

Rochester Institute of Technology

RIT Digital Institutional Repository

Theses

9-2023

Robust Weakly Supervised Learning for Real-World Anomaly Detection

Hitesh Sapkota
hxs1943@rit.edu

Follow this and additional works at: <https://repository.rit.edu/theses>

Recommended Citation

Sapkota, Hitesh, "Robust Weakly Supervised Learning for Real-World Anomaly Detection" (2023). Thesis. Rochester Institute of Technology. Accessed from

This Dissertation is brought to you for free and open access by the RIT Libraries. For more information, please contact repository@rit.edu.

Robust Weakly Supervised Learning for Real-World Anomaly Detection

by

Hitesh Sapkota

A dissertation submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in Computing and Information Sciences

B. Thomas Golisano College of Computing and
Information Sciences

Rochester Institute of Technology
Rochester, New York
09/2023

Robust Weakly Supervised Learning for Real-World Anomaly Detection

by
Hitesh Sapkota

Committee Approval:

We, the undersigned committee members, certify that we have advised and/or supervised the candidate on the work described in this dissertation. We further certify that we have reviewed the dissertation manuscript and approve it in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Computing and Information Sciences.

Qi Yu	Date
Dissertation Advisor	

Zhiqiang Tao	Date
Dissertation Committee Member	

Rui Li	Date
Dissertation Committee Member	

Dongfang Liu	Date
Dissertation Committee Member	

Dan Phillips	Date
Dissertation Defense Chairperson	

Certified by:

Pengcheng Shi	Date
Ph.D. Program Director, Computing and Information Sciences	

Robust Weakly Supervised Learning for Real-World Anomaly Detection

by

Hitesh Sapkota

Submitted to the

B. Thomas Golisano College of Computing and Information Sciences Ph.D. Program in
Computing and Information Sciences

in partial fulfillment of the requirements for the

Doctor of Philosophy Degree

at the Rochester Institute of Technology

Abstract

Anomaly detection has attracted increasing attentions from diverse domains, including medicine, public safety, and military operations. Despite its wide applicability, anomaly detection is inherently challenging as abnormal activities are usually rare and unbounded in nature. This makes it difficult and expensive to identify all potential anomalous events and label them during the training phase so that a detection model can provide robust prediction on unseen event data. Unsupervised and semi-supervised learning models have been explored, which achieve a decent detection performance with no or much less annotations. However, these models can be highly sensitive to outliers (*i.e.*, normal samples that look different from other normal ones) or multimodal scenarios (*i.e.*, existence of multiple types of anomalies), leading to much worse detection performance under these situations. The imbalanced class distribution of the normal data samples poses a further challenge as the model may be confused between the normal samples from a minority class and true anomalies.

To systematically address the key challenges outlined above, this dissertation contributes the first Robust Weakly Supervised Learning (RWSL) framework that provides fundamental support for real-world anomaly detection using only weak and/or sparse learning signals. The proposed RWSL framework offers a principled learning paradigm to deal with the rare and unbounded nature of real-world anomalies, which allows a statistical learning model to be robustly trained using only high-level supervised signals while generalizing well in few-shot settings. The framework is comprised of three interconnected components. The first research component integrates Robust Distributionally Optimization (DRO) with Bayesian learning, leading to a novel Bayesian DRO model that achieves robust detection performance using weak learning signals coupled with outliers and multimodal anomalies. The Bayesian DRO model is further augmented with non-parametric sub-modular optimization and active instance sampling to improve both the reliability and accuracy of the detection performance. The second research component leverages the evidential theory and

its fine-grained uncertainty formulation to tackle anomaly detection coupled with imbalanced class distribution of normal data samples. An adaptive Distributionally Robust Evidential Optimization (DREO) training process is developed to boost the anomaly detection performance by accurately differentiating minority class samples and true anomalies using evidential uncertainty. Evidential learning is further integrated with a transformer architecture, leading to an Evidential Meta Transformer (MET) for reliable anomaly detection in the few-shot setting. Finally, the third research component aims to ensure an unbiased (*i.e.*, fair) and better-calibrated model with improved anomaly detection performance by avoiding the overconfidence predictions stemming from the memorization effect seen in deep neural networks. To achieve this, the Distributionally Robust Ensemble (DRE) is proposed that learns multiple diverse and complementary sparse sub-networks through the utilization of DRO properties. By facilitating these sparse sub-networks to capture different data distributions across varying levels of complexity, they naturally complement each other resulting in improved model calibration with enhanced anomaly detection capability.

Acknowledgments

First of all, I would like to express my heartfelt gratitude to my advisor Professor Qi Yu for his exceptional guidance, unwavering support and invaluable suggestions throughout my doctoral journey. It would be unattainable to achieve any accomplishment in my PhD career without his help. Also, I feel lucky to have Professor Qi Yu as my advisor who prioritized creating a comfortable and friendly working environment, fostering constructive collaborations that have contributed to the ongoing progression throughout my PhD journey. I am grateful to my advisor for genuinely believing in my capabilities that has significantly boosted my confidence in my life. I am thankful to my committee members Professor Rui Li, Professor Dongfang Liu, and Professor Zhiqiang Tao, for their insightful comments, constructive criticisms, and valuable suggestions that have helped to enrich the quality of this thesis.

Secondly, I would like to extend my sincere appreciation to Professor Feng Chen from UT Dallas, Professor Yiming Ying from University at Albany SUNY, and Professor Pradeep Murukannaiah from TU Delft. I would also like to thank PhD program director Professor Phengcheng Shi, and faculty members Charles Gruener, MinHong Fu for continuous guidance, insights, and administrative support.

Thirdly, I am grateful to the members of the RIT-MINING lab, including Wei Shi Shi, for their constructive feedback along with collaboration that have become pivotal to broaden my perspectives and deepened my understanding in the field.

Finally, I would like to thank my wife Kusum Parajuli, my parents Cheena Sapkota, Laxman Sapkota, and my Sisters Bima Sapkota, and Moona Sapkota for their unconditional support, guidance and encouragement during my challenging times. Also, I am grateful to all relatives, and my friends for their constant support.

This page is dedicated to my wife Kusum Parajuli and my parents Laxman Sapkota, and Cheena Sapkota.

Contents

1	Introduction	1
1.1	Problem Statement and Research Challenges	2
1.1.1	Anomaly Detection under Weakly Supervised Learning	3
1.1.2	Anomaly Detection under Few Shot Learning	4
1.1.3	Anomaly Detection under Imbalanced Class Distribution	5
1.1.4	Anomaly Detection under Undercalibrated Model	5
1.2	Robust Weakly Supervised Learning Framework	5
1.3	Summary of Contributions	7
2	Literature Review	10
2.1	Anomaly Detection	10
2.2	Distributionally Robust Optimization (DRO)	11
2.3	Openset Detection	12
3	Robust Multiple Instance Learning for Anomaly Detection	14
3.1	Distributionally Robust Optimization for Deep Kernel Multiple Instance Learning .	15
3.1.1	Related Work	17

3.1.2	DRO Deep Kernel Multiple Instance Learning (DRO-DKMIL)	17
3.1.3	Experiments	24
3.1.4	Datasets and Experimental Settings	24
3.1.5	Impact of Key Model Parameters	30
3.1.6	Conclusion	32
3.2	Bayesian Nonparametric Submodular Video Partition for Robust Anomaly Detection	32
3.2.1	Related Work	35
3.2.2	Methodology	36
3.2.3	Experiments	43
3.2.4	Datasets and Experimental Settings	43
3.2.5	Performance Comparison	45
3.2.6	Detecting Multimodal and Outlier Segments	46
3.2.7	Qualitative Analysis	48
3.2.8	Conclusion	49
4	Multiple Instance Active Learning for Anomaly Detection	50
4.1	Related Work	52
4.2	Methodology	53
4.2.1	Variance Regularization	53
4.2.2	Distributionally Robust Bag Likelihood	55
4.2.3	P-F Active Sampling	57
4.3	Experiments	60

4.3.1	Experimental Setup	62
4.3.2	Performance Comparison	63
4.3.3	Ablation Study	66
4.3.4	Qualitative analysis	66
4.4	Conclusion	67
5	Anomaly Detection under Class Imbalanced Setting	68
5.1	Related Work	70
5.2	Methodology	70
5.3	Experiments	76
5.4	Conclusion	81
6	Anomaly Detection under Few-Shot Learning Settings	83
6.1	Related Work	86
6.2	Methodology	87
6.2.1	Preliminaries	87
6.2.2	Transformer based FSOSR	88
6.2.3	Meta Evidential Transformer (MET)	90
6.3	Experiments	94
6.3.1	Results and Discussion	95
6.3.2	Ablation Study	95
6.4	Conclusion	97
7	Anomaly Detection under Sparse Network Training	98

7.1	Related Work	101
7.2	Methodology	102
7.2.1	Preliminaries	102
7.2.2	Distributionally Robust Ensemble (DRE)	103
7.2.3	Theoretical Analysis	105
7.3	Experiments	107
7.3.1	Experimental Settings	108
7.3.2	Performance Comparison	108
7.3.3	Additional Results, Ablation Study, and Qualitative Analysis	112
7.4	Conclusion	112
8	Conclusion and Future Works	113
8.1	Conclusion	113
8.2	Future Works	114
9	List of Publications	117
9.1	Published	117
9.2	Submitted (Preprints)	118
	Appendices	134
A		135
A.1	Distributionally Robust Optimization for Deep Kernel Multiple Instance Learning .	135
A.2	Bayesian Nonparametric Submodular Video Partition for Robust Anomaly Detection	141

B	144
C	157
C.1 Summary of Notations	157
C.2 Additional Related Work	157
C.3 Theoretical Proof	159
C.3.1 Proof of Lemma 6.1	160
C.3.2 Proof of Theorem 6.2	160
C.4 Experimental Details and Additional Results	162
C.4.1 Dataset Distribution	162
C.4.2 Experimentation Details	163
C.4.3 Closed-Set Performance	164
C.4.4 ROC Curves	164
C.4.5 Ablation Study	164
C.4.6 Qualitative Analysis	167
D	169
D.1 Summary of Notations	169
D.2 Robust Loss Optimization in DRO	169
D.2.1 Robust Loss Optimization	169
D.2.2 Hyperparameter settings	171
D.3 Theoretical Proof	172
D.3.1 Proof of Lemma 7.1	172

D.3.2	Proof of Theorem 7.2	174
D.4	Experimental Details and Additional Results	175
D.4.1	Detailed Dataset Description	175
D.4.2	Hardware Details for Experimentation	176
D.4.3	Single-view and Multi-view Examples	176
D.4.4	Additional Result on Cifar10 and Cifar100	177
D.4.5	Additional Baseline Results on TinyImageNet	177
D.4.6	Performance from Ensemble Members	178
D.4.7	Comparison with Common Calibration Techniques	178
D.4.8	Ablation Study	179
D.4.9	Parameter Size and Inference Speed	181
D.4.10	Diversity on Sparse Sub-networks	181
D.4.11	Qualitative Analysis	182
D.5	Broader Impact, Limitations, and Future Work	182
D.5.1	Broader Impact	184
D.5.2	Limitations and Future Works	184

List of Figures

1.1	Weakly supervised and FSL Setting	2
1.2	Examples of outlier (a-b) and multimodal frames (c-e) from the Avenue dataset . . .	3
1.3	Robust Weakly Supervised Learning Framework	6
3.1	ROC Performance on Three Video Datasets (a)-(c); Multimodal (d) and Outlier Prediction (e)	27
3.2	Abnormal Frames Identified by DRO-DKMIL but not DK-MMIL	29
3.3	Abnormal Frame Prediction	29
3.4	Uncertainty of Different Frames	30
3.5	AUC Performance vs. η and Comparison with Average top- k	30
3.6	Highly fluctuating detection performance w.r.t. k	34
3.7	Output of the top- k based approach in a video from Avenue dataset (missing some of the abnormal segments in top- k).	36
3.8	Example frames from different scenes in an explosion video from UCF-Crime: (a-b) scene 1, (c) scene 2, (d-e) scene 3	37
3.9	Performance comparison with top- k ranking models	44
3.10	ROC curves on three video datasets (a)-(c), multimodal (d) and outlier (e)	46

3.11	Frames from UCF-Crime Stealing019; (a) Correct BN-SVP, Avg Topk, (b) Correct BN-SVP, Incorrect Avg Topk	48
4.1	(a) Example of a challenging bag; (b) MI-AL performance on instance-level predictions; (c)-(e) Prediction scores of instances in the bag in different MI-AL steps . . .	51
4.2	Example of challenging bags from different topics in 20NewsGroup	58
4.3	MI-AL performance	63
4.4	Effectiveness of P-F active sampling	63
4.5	Impact of model parameter λ	65
4.6	Impact of model parameter β	65
4.7	Impact of hyperparameter k	65
4.8	(a-b) Poorly explored bags in Pascal VOC; (c) Description of these bags and their mAP scores; (d) Additional true positive bags successfully explored by P-F sampling	66
5.1	Examples of Scheduler Functions	73
5.2	OSD performance comparison from imbalanced Cifar10 dataset.	80
5.3	(a) Top row: minority class; bottom-row: majority classes; (b) sample ranking. . . .	81
6.1	OSR performance (AUROC) of a difficult task consisting of similar closed- and open-set images with examples shown in (a). (b) SnaTCHer (72.84%) and (c) MET (83.34%) that uses Class Mamalute (83) serves as the opponent class for better separation between Ferrets (closed) and Golden Retriever (open).	84
6.2	(a) MET training pipeline and opposing class selection (c) for compact closed-set representation learning (c).	88
6.3	(a) A model trained using evidential loss fails to identify the Bengal Cat as open-set because it shares feature similarity with Maine Coon and is very distinct from the Dog class. (b) EVR and evidential cross-attention help to recognize the open-set sample. (c) Overall inference process that integrates EVR and evidential cross-attention. . .	92

6.4	OSR performance comparison on MiniImageNet.	97
7.1	Calibration performance by expected calibration error (ECE) on Cifar100 dataset with ResNet101 architecture with density $\mathcal{K} = 15\%$. EP refers to the Edge Popup algorithm [110].	99
7.2	Robust ensemble where η defines the size of an uncertainty set with $\eta_1 \leq \eta_2 \leq \eta_3$. . .	104
7.3	Open-set detection performance on different confidence thresholds.	111
C.1	ROC curves on both 5-way-1-shot and 5-way-5-shot tasks	165
C.2	OSR performance with respect to hyperparameter ϵ : (a-b) MiniImageNet, (c-d) TieredImageNet.	165
C.3	OSR performance with respect to hyperparameter λ : (a-b) MiniImageNet, (c-d) TieredImageNet.	166
C.4	Examples of difficult images with the corresponding ranking	168
D.1	Examples of single-view and multi-view samples.	176
D.2	(a-b) Impact of λ on ECE using ResNet101 architecture on Cifar100 dataset. . . .	180
D.3	Confidence scores of incorrectly classified samples in CIFAR100 with ResNet101 . .	183
D.4	Confidence scores of correctly classified samples in CIFAR100 with ResNet101 . . .	183

List of Tables

3.1	Symbols with Descriptions	18
3.2	Video Level Distribution on Different Datasets	24
3.3	Comparison of AUC Scores	26
3.4	AUC on Multimodal and Outlier Detection	28
3.5	Symbols with Descriptions	38
3.6	Comparison with Other Models	47
3.7	AUC Scores on Multimodal and Outlier Detection	48
4.1	Number of positive and negative bags on different datasets	62
4.2	MIL Performance in Passive Setting	64
5.1	Symbols with Descriptions	71
5.2	OSD (MAP) performance on all datasets	79
5.3	Closed set performance (MAP) on all datasets	79
6.2	Ablation study results on MiniImageNet.	95
6.1	OSD (AUROC) performance on different datasets.	96

7.1	Accuracy and ECE performance with 9% density for Cifar10 and Cifar100.	109
7.2	Accuracy and ECE on TinyImageNet.	109
7.3	Accuracy and ECE performance on out-of-distribution datasets.	110
C.1	Notations with Descriptions	158
C.2	Train/Evaluation/Test partition on different datasets.	163
C.3	Closed set performance (ACC) on different datasets.	164
C.4	MiniImageNet performance with: (a) Different backbones, (b) Original data split. . .	166
C.5	OSD (AUROC) performance on additional datasets.	167
D.1	Symbols with Descriptions.	170
D.2	Accuracy and ECE performance with 15% density for Cifar10 and Cifar100 Dataset. .	177
D.3	Additional baseline results on TinyImageNet using ResNet50 with $\mathcal{K} = 15\%$	178
D.4	Different subnetworks performance on Cifar100 Dataset.	178
D.5	Different calibration techniques on the top of EP Algorithm with $\mathcal{K} = 9\%$	179
D.6	ACC and ECE with different: (a) backbones and (b) number of subnetworks.	180
D.7	Parameter size and inference speed.	181
D.8	Accuracy, ECE, and prediction disagreement performance with a $\mathcal{K} = 15\%$ density. .	182

Chapter 1

Introduction

There has been tremendous progress in the field of artificial intelligence after the arrival of deep neural networks (DNNs). To achieve a high prediction accuracy, the training of a DNN usually requires an extensive amount of labeled data samples. Despite the outstanding prediction performance in many computer vision and natural language process tasks, anomaly detection in real-world settings still poses fundamental challenges for the current machine learning models. First, abnormal activities are usually rare and unbounded in nature. This makes it difficult and expensive to identify all potential events and label them during the training phase to learn a detection model that can perform robustly on previously unseen event data. Second, the existence of outliers (*i.e.*, normal samples that appear to be different from other normal ones) and multimodal scenarios (*i.e.*, co-occurrence of multiple types of anomalies) further complicates the detection of truly abnormal events. Third, the imbalanced class distribution of normal data samples (*i.e.*, number of samples for from certain known classes is much less than the rest) may make the model easily confused between minority class samples and true anomalies. Finally, the spurious correlations that are introduced with the data collection process coupled with overparameterized network resulting from the memorization effect may result in a biased and un-calibrated model that tends to misidentify the less representative samples as anomalies. Furthermore, such an un-calibrated model may wrongly identify an anomalous sample as the known sample with high confidence.

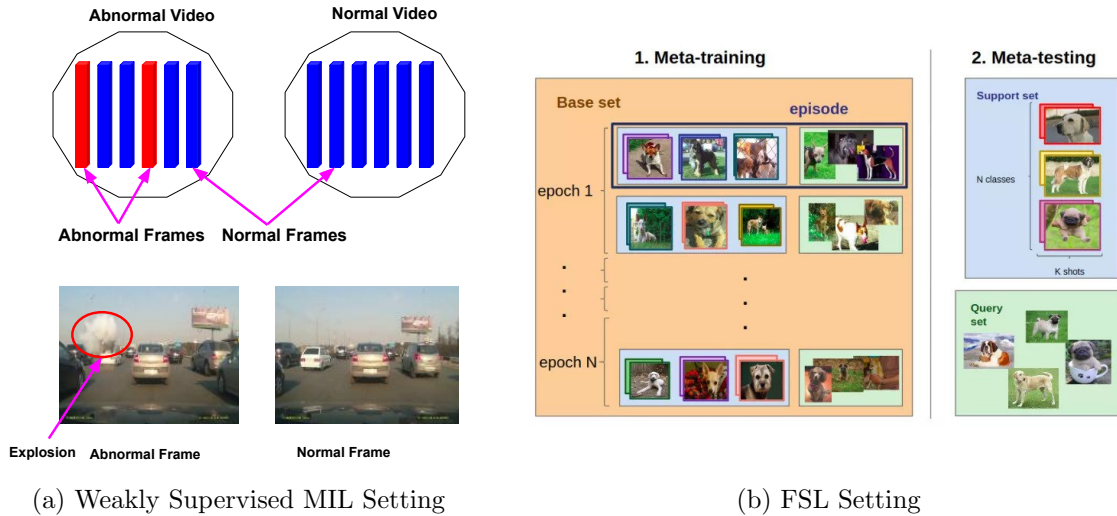


Figure 1.1: Weakly supervised and FSL Setting

1.1 Problem Statement and Research Challenges

Our study mainly focuses on robust learning under weak supervision with only weak and/or sparse learning signals. There are four major tasks in anomaly detection in real-world settings. The first task is related to learning from weak supervision, referred to as weakly supervised learning where only high level learning signals are present. Multiple Instance Learning (MIL) is one of the widely used weakly supervised learning techniques, where we have a collection of positive and negative bags (*e.g.*, videos) as shown in Figure 1.1 (a). A bag (video) is considered to be positive (abnormal) if at least one instance (frame) is positive (abnormal) otherwise negative (normal). During the training process, only the bag (video) level annotation is present whereas, that of the instance (frame) level annotation is missing. Based on the bag (video) level information, we want to train the model that can perform well in the anomaly detection task indicating whether a given frame is abnormal or normal. However, the lack of instance labels, the rare and unbounded nature of abnormal events, coupled with the existence of outliers and multimodal scenarios, make weakly supervised anomaly detection highly challenging.

The second task concerns the few-shot learning (FSL) setting, where strong supervised signals are available but the total annotated data samples are extremely limited. Meta-learning offers a promising vehicle to tackle FSL. In meta-learning, a dataset is divided into meta-training and meta-testing phase as shown in Figure 1.1 (b). During the meta-training phase, a task is constructed by randomly picking N -classes with each class having K -samples, and the corresponding problem is



Figure 1.2: Examples of outlier (a-b) and multimodal frames (c-e) from the Avenue dataset

referred to as N -way K -shot problem. The model is trained using multiple tasks, and finally, it is tested on the meta-testing dataset. During the testing phase, the model has to classify the given query set example into one of the N -classes based on the limited samples present in the support set. It should be noted that the number of samples per class *i.e.*, K is very small and therefore, the model has to make a prediction based on the limited samples. The anomaly detection under the FSL is a very challenging problem as the model should be able to discriminate the normal class samples and anomalous samples based on the very weak signal.

The third task is concerned with learning under the class imbalanced setting where a number of samples for a specific class (classes) is very small compared to the rest. The anomaly detection under such a setting is difficult as the model be confused between normal samples from minority class and true anomalies. The final task is related to learning under the spurious correlation setting coupled with the overparamaterized network resulting from the memorization effect, where the biased and/or uncalibrated model resulting from such data may be confused between the in-distribution samples and anomalous ones. Also, in this case, because of the poor calibration, the model may incorrectly identify the anomalous samples as the known samples with high confidence which would reduce the trust of the people representing minority groups toward the model.

Next, we present each sub-task in detail and point out the limitations of the existing techniques that motivate us to formulate the proposed Robust Weakly Supervised Learning (RWSL) framework.

1.1.1 Anomaly Detection under Weakly Supervised Learning

To tackle anomaly detection under weakly supervised learning, the MIL paradigm has been used that models each video as a bag and its segments (or frames) as instances within the bag [129]. One effective MIL learning objective considered in the past is to maximize the gap between two instances having the respective highest anomaly scores from a pair of positive and negative bags, called maximum-based multiple instance learning (MMIL). However, those techniques are highly sensitive to the outliers and multimodal scenarios where some of the representative examples are demonstrated in Figure 1.2. The images shown in (a-b) are outliers that look like abnormal frames

but are indeed normal frames. For example, the frame in (a) is a normal frame from burglary activity but looks similar to the arson-related abnormal frames. Similarly, the frame in (b) is a normal frame that looks similar to fighting-related frames. If there exist such type of frames, the MMIL techniques may miss actual abnormal frames during the training process and instead may try to maximize the gap between the outlier and the instance with the highest anomaly score from the normal bag yielding lower anomaly detection performance. Figures (c-e) are different types of abnormal frames from the same video and the situation is called multimodality. As MMIL considers a single frame from an abnormal video, it may try to make the anomaly score of only one type of abnormal frame while ignoring the rest during the training process and thereby consequently leading low anomaly detection performance.

Top- k ranking loss has been adopted in an attempt to address the issues outlined above. It maximizes the gap between the mean score of the top- k predicted instances from a positive bag and that of a negative one [116, 137]. However, there are inherent limitations to using a top- k loss. First, it tends to be extremely sensitive to the selected k value shown in Figure 3.6. Since there is no frame (or segment) labels available during model training, setting an optimal k through cross-validation is infeasible or highly costly. Second, given the diverse videos, the number of abnormal instances may vary significantly from one video to another implying we should have a different k for each video. Hence, applying the same k to all videos as in the existing approaches fails to capture the nature of the data. The third issue is that all (or most of) the selected k segments may come from the same sub-sequence of the video. Using a consecutive set of visually similar segments is less effective for model training, making it more likely to suffer from outlier and multimodal scenarios.

1.1.2 Anomaly Detection under Few Shot Learning

Anomaly detection under the FSL setting has become a challenging problem because of the limited samples per normal classes as it can easily mislead the model to incorrectly identify anomalous samples as the normal data samples and vice versa. There have been few attempts to address few-shot open-set recognition (FSOSR) [56, 80]. However, those techniques have failed to learn a compact representation of the normal classes specifically in the case where abnormal activities share some similarities with the normal classes. Therefore it is important to consider the subtle differences leading to discriminate the abnormal events from the normal samples.

1.1.3 Anomaly Detection under Imbalanced Class Distribution

Anomaly detection has become a challenging case in the limited data sample setting specifically in the imbalanced class distribution and FSL setting. Considering the imbalanced class distribution setting, DRO can be used which has proven to be effective to learn a robust representation under such setting [108, 164]. Further, imbalanced class distribution in the closed-set is handled through oversampling to achieve a more balanced class distribution [18]. Although the existing techniques are shown to be effective in a closed set classification setting, neither of them is adequate to anomaly detection in the imbalanced class setting. A fundamental challenge lies in the interplay between normal data samples from the minority class and the difficult samples from the majority classes. As a result, simply oversampling the minority class may neglect these difficult samples. Similarly, applying DRO with a flexible uncertainty set may put too much emphasis on these difficult samples and ignore the minority class as well as some representative samples from the majority classes, which affects proper model training.

1.1.4 Anomaly Detection under Undercalibrated Model

The modern deep neural network tends to exhibit an overfitting phenomenon resulting from the memorization effect. Furthermore, the presence of a spurious correlation between shortcuts and associated classes makes the situation worse. As such, the trained model becomes heavily biased and highly uncalibrated. Therefore, if the anomaly event shares some similarity with training known samples, the model may confidently detect the anomalous sample as the normal sample. Also, because of the poor calibration and biases, the known samples (especially the minority ones) may be misclassified as the anomalous sample. There have been existing efforts to make the model sparse which may potentially improve the calibration and reduce the biases of existing models. However, the primary focus of those sparse techniques is to match the accuracy as that of dense networks without explicitly paying attention to calibration and biases. Despite having some improvements, those techniques do not yield optimal performance in terms of better calibration and improvement toward biases.

1.2 Robust Weakly Supervised Learning Framework

Our solutions to the aforementioned challenges are well organized under a novel robust weak supervision learning framework. DRO and related robust techniques are powerful tools to help design

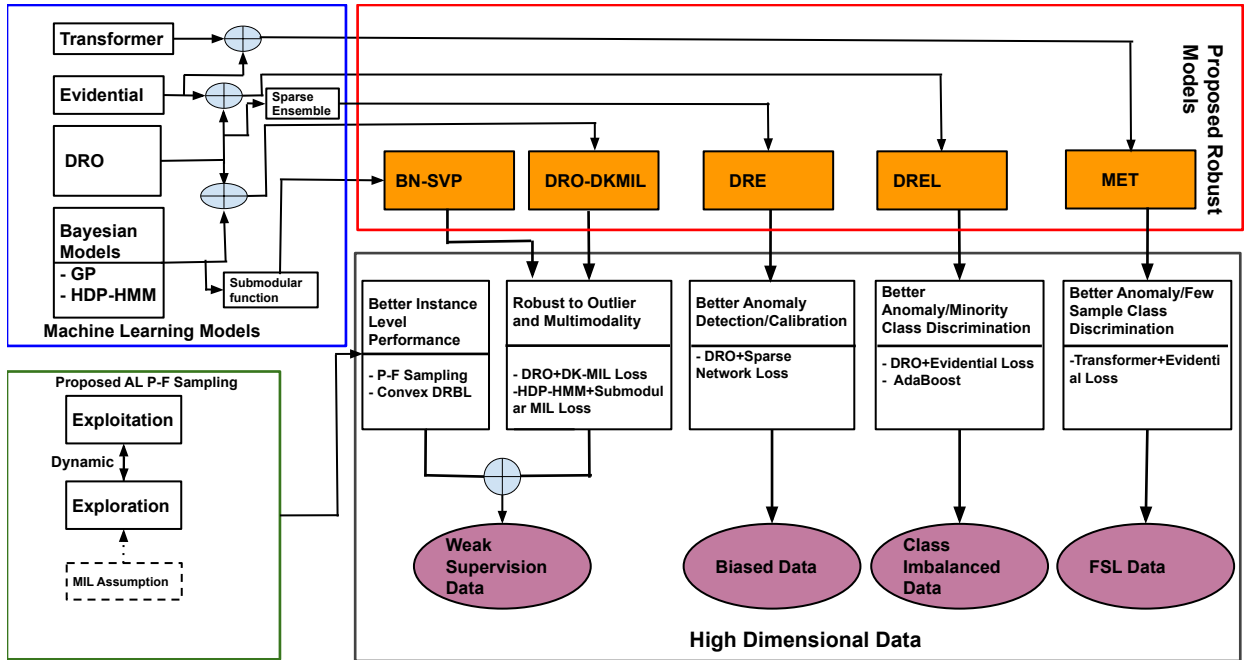


Figure 1.3: Robust Weakly Supervised Learning Framework

learning paradigms that are robust under outlier, multimodal, spurious correlation, and class/group imbalanced settings. Figure 1.3 shows the overall diagram depicting our framework. Integration of the DRO with Bayesian models (such as GP) helps to address the multiple issues present in the anomaly detection task under a weakly supervised learning setting. Specifically, under a weak supervision setting like MIL, such technique helps to make the approach robust to outlier and multimodal scenario demonstrated in Figure 1.2. Further augmentation with non-parametric models like HDP-HMM coupled with submodular function and active learning strategy further augments the performance by ensuring diverse types of abnormal patterns during the training process. Also, the proposed technique automatically adjusts the k -value depending on the nature of the video yielding better performance comparison. The novel active learning techniques like P-F sampling help to identify and label the rare abnormal patterns and thereby improving both reliability as well as accuracy of the anomaly detection performance.

A non-trivial combination of evidential learning with the adaptive DRO provides an effective way to perform anomaly detection under the imbalanced class distribution setting by providing high uncertainty for the abnormal samples. It should be noted that neither evidential learning nor a simple straightforward combination of DRO and evidential network is sufficient to tackle the anomaly detection task.

Considering the complexity of the anomaly detection task under the limited data setting *i.e.*, few-shot learning, we devise a novel evidential transformer network, called Meta Evidential Transformer (MET) by combining evidential theory and transformer network. This novel MET based architecture learns a compact representation for each normal class by leveraging the properties of transformer network.

To tackle anomaly detection under an over-parametrized network and spurious correlation setting, we propose Distributionally Robust Ensemble (DRE) that uses DRO along with the sparse network. In this technique, we aim to learn multiple diverse and complementary sparse sub-networks (tickets) with the guidance of uncertainty sets through DRO, which encourages tickets to gradually capture different data distributions from easy to hard and naturally complement each other. Specifically, through the DRO framework, the proposed technique avoids learning from the spurious features and/or noises while ensuring diversity among the learned sparse sub-networks. Therefore, the resulting ensemble model becomes more trustworthy with better calibration and unbiased. It is worth noting that the final ensemble model may be able to better recognize the anomalous samples instead of being confidently wrong. Also, by avoiding learning from spurious correlation, the model will be able to better recognize minority group samples as known samples.

1.3 Summary of Contributions

Considering anomaly detection under weak-level supervision, we propose three approaches that essentially solve the multiple challenges that exist in the previous techniques. First two approaches are discussed in Chapter 3 and the last one is discussed in Chapter 4. In chapter 3, first we discuss the DRO-based deep kernel multiple instance learning technique. This technique assigns the non-zero probability to each instance in a bag through a Gaussian Process (GP) latent mixture model. This way the proposed technique becomes more robust to outlier and multimodal scenarios as it gives chances to all frames to participate in the optimization with different probabilities. Further, the DRO constraint defined over the mixture model relaxes the limitations of specifying a fixed k value which is very challenging in the existing top- k approaches. Further, through the integration of non-parametric GP with the powerful DNN, our approach becomes a very effective approach to handle high dimensional data resulting in state-of-the-art performance.

The second approach (in chapter 3) is defined to address the limitations that exist in the top- k approach along with our DRO-based deep kernel multiple instance learning model. In addition to setting fluctuating k (in top- k), there is another limitation resulting from the frame redundancy. It

is quite possible that all (or most of) the selected k segments may come from the same sub-sequence of the video because of temporal consistency and feature similarity. However, using a consecutive set of visually similar segments is less effective for model training, making it more likely to suffer from outlier and multimodal scenarios. To solve this problem we propose novel Bayesian non-parametric construction of a submodular set function, which is integrated with multiple instance learning to deliver robust video anomaly detection performance under practical settings. The hierarchical dirichlet process (HDP) prior on the state transition probability of the hidden markov model (HMM) helps to avoid the challenging problem of setting k value. In fact, HDP-HMM helps the model automatically decide the k -value for each video depending on the nature of the video. Also, novel submodular function helps to ensure the diverse frames to be selected during the training process which increases the chance of including many abnormal events in the training. This technique addresses both limitations of setting proper k value as well as maintaining the diversity which results in a better performance.

The third approach (in chapter 4) is related to multiple instance active learning. It is found that only using the weakly supervised approach such as MIL may not be sufficient to increase the instance level prediction in challenging cases. The underlying reason for the less accurate instance-level prediction is due to the lack of instance labels. For positive instances that are relatively rare across bags, detecting them by only relying on bag labels is inherently challenging as the weakly supervised signal (*i.e.*, bag label) cannot be propagated to the instance level without sufficient statistical evidence. By leveraging the key MIL assumption, the novel P-F sampling function can explore the most challenging bags and effectively detect their positive instances for annotation, which significantly improves the instance-level prediction. Further, we theoretically show that the proposed distributionally robust bag likelihood (DRBL) helps to detect potentially positive instances to support the proposed P-F sampling.

We consider anomaly detection under the imbalanced class distribution of normal data samples in Chapter 5. Specifically, to tackle this problem, we develop an adaptive Distributionally Robust Evidential Optimization (DREO) training process that provides the principled way to quantify sample uncertainty through evidential learning while optimally balancing the model training over all classes in the closed set through adaptive DRO framework. The proposed DREO accurately differentiates the minority class samples and true anomalies through evidential uncertainty. To avoid the model from primarily focusing on the most difficult samples by following the standard DRO, the adaptive learning strategy gradually increases the size of the uncertainty set, which allows the model to learn from easy to hard samples. Further, our experimentation along with the theoretical analysis justifies the robustness of the proposed technique in anomaly detection.

We devise a Meta Evidential Learning (MET) framework to deal with anomaly detection under the few-shot learning setting in Chapter 6. The proposed MET framework uses an evidential open-set loss to learn more compact closed-set class representations by effectively leveraging similar closed-set classes. MET further integrates an evidence-to-variance ratio to detect fundamentally challenging tasks and uses an evidence-guided cross-attention mechanism to better separate the difficult open-set samples. As such, during the testing process, the model will be able to better identify the challenging anomalous samples that may exhibit similarity with the known classes. Also, because of the compact representation capability coupled with EVR, the model will be able to better identify the known class samples despite having very few support set samples.

Finally, in Chapter 7 we aim to devise a model that is better calibrated, trustworthy, and unbiased for better anomaly detection. To accomplish this, we propose a novel Distributionally Robust Optimization (DRO) framework to achieve an ensemble of lottery tickets toward calibrated network sparsification. Specifically, the proposed DRO ensemble aims to learn multiple diverse and complementary sparse sub-networks (tickets) with the guidance of uncertainty sets, which encourage tickets to gradually capture different data distributions from easy to hard and naturally complement each other. As such, the DRE framework avoids learning from the spurious correlation and/or noises and thereby avoiding the overfitting phenomenon. During testing, the resulting ensemble model becomes much more trustworthy, better calibrated, and unbiased. Therefore, the model correctly identifies the unknown sample without being confidently wrong. Furthermore, as it avoids learning from the spurious correlation, the model correctly identifies minority group samples as the known samples.

Chapter 2

Literature Review

In this chapter, we discuss the existing work that is common to multiple chapters. Specifically, we will discuss related work on (1) Anomaly Detection, (2) Distributionally Robust Optimization, and (3) Openset Detection. The first topic addresses mainly unsupervised, and multiple instance learning-based anomaly detection techniques. In this topic, we review existing techniques to solve anomaly detection and figure out shortcomings. Next, we explain the DRO related work and connect it with our technique regarding how it can address shortcomings present in the existing multiple instance learning techniques for anomaly detection. Finally, we perform a literature review on the openset detection techniques in a general setting.

2.1 Anomaly Detection

Encoding and sparse reconstruction-based approaches have been employed for anomaly detection, assuming that abnormal events are rare and deviate from normal patterns. They aim to capture the normal patterns using models, such as Gaussian processes (GPs) [76] and HMMs [66], to identify anomalies as outliers based on the reconstruction loss. Sparse representation-based approaches construct a dictionary for normal events and identify the events with the high reconstruction error as anomalies [87]. Recent approaches consider both abnormal and normal events in the training process. For video anomaly detection, since only video-level labels are assumed to be available during model training [48], MIL offers a natural solution by modeling each video as a bag and the associated segments (frames) as instances of the bag. Sultani et al. proposed an MIL based approach that enables to maximize the gap between highest prediction scores from a positive and

negative bags, respectively [129]. However, this maximum score based MIL model (*i.e.*, MMIL) is insufficient to handle outlier and multimodal scenarios.

top- k ranking loss based MIL models have been developed to address the limitations of the MMIL model [137]. These models produce state-of-the-art detection performance given that a suitable k value can be assigned in advance. However, the detection performance of such models is highly sensitive to the chosen k value. The main issue with the top- k ranking loss is how to set a suitable k , which can be quite challenging in practice. More importantly, since k takes discrete values, the prediction performance may fluctuate significantly when k changes. To address this fundamental challenge, we first, propose to integrate a DRO constraint into the GP mixture framework, which can essentially function as a soft version of the top- k constraint, thus removing the need to specify a fixed k value while ensuring a more stable (and robust) prediction [116].

Although the DRO based proposed approach avoids the issue of setting k value while stabilizing the performance, we may still need to manually set the size of the uncertainty set is controlled by the radius (*i.e.*, η) of the uncertainty ball. Furthermore, both the DRO based approach as well as top- k variants may put more focus on a set of consecutive segments with the highest prediction scores and ignore some other potentially positive segments resulting into degradation in the performance. To overcome those issues, we next propose a novel submodular set function in a non-parametric way by inferring the diversity from data automatically. By jointly optimizing the submodular function and the MIL loss, it automatically chooses a diverse set of segments and lets the model better differentiate these (potentially positive) segments from those of a negative bag to ensure good detection performance.

2.2 Distributionally Robust Optimization (DRO)

Distributionally robust optimization is based on principled statistical learning theory, where the worst case weighted loss is optimized by searching the weights in a given uncertainty set [29,98,164]. DRO offers a systematic way to handle the imbalanced class distribution and has been commonly used in supervised learning setting [108,164] as well as in multiple instance learning [116]. DRO has been employed in supervised learning to assign different weights to different losses so as to maximize the overall weighted loss over an uncertainty set for the distributional variable [98,107]. Depending on how the uncertainty set is defined, the DRO-based loss reduces to different types of widely known loss functions. For example, by restricting the distribution of the distributional variable within a certain ball with a center given by the uniform distribution, DRO-based loss

becomes variance regularized loss [99]. Similarly, by making the distributional variable take any value between 0 to 1, the corresponding DRO-based loss becomes a maximal loss and the top- k loss when further restricting the distributional variable value between 0 and $\frac{1}{k}$ [32]. In a similar way, [77] proposes a technique called Tilted Empirical Risk Minimization (TERM) by redefining the ERM with the introduction of hyperparameter t . Depending on the tunable parameter t value, different variants of loss (maximum, minimum, and average) are recovered and thereby provide a unified way to perform effective training in the presence of outlier and class imbalance scenarios. In the DRO-based anomaly detection task work, we integrate DRO to constrain the parameter that governs the probability of each frame being a positive instance in the proposed GP mixture model. In this work, we are the first one to introduces DRO into MIL, leading to a DRO based GP mixture model that provides robust MIL predictions.

Considering the Openset Detection Setting (OSD), while DRO may help to improve the close set performance, it is not sufficient to address the OSD problem with imbalanced data. This is because DRO with a flexible uncertainty set may put too much emphasis on the difficult samples and ignores the ones from the minority class as well as representative samples from majority classes. Therefore, we propose an adaptive learning strategy to learn from easy samples in the early training phase and gradually shift the focus to the difficult samples. Furthermore, the class-ratio biased loss ensures proper learning from the limited samples in the minority class.

2.3 Openset Detection

Various SVM based techniques [54, 118, 119] have proposed for OSD. For instance, Scheirer et al. [119] proposed an SVM based model, which performs detection using a Weibull-calibrated SVM (W-SVM) by leveraging Extreme Value Theory (EVT). Reconstruction based approaches have been proposed [157], where a threshold defined over the reconstruction error is used to decide whether the sample is from a known or an unknown class. Other traditional models, such as nearest neighbor [59], quasi-linear function [15], have also been explored as well. Deep learning models have been increasingly applied for open set detection [7, 130, 154]. As an example, OpenMAX replaces the softmax function and probability of the softmax is redistributed to produce the probability of a sample being unknown [7]. Sun et al. [130] proposed VAE based open set recognition, where the probability of a sample belonging to each of the known classes is used as a proxy to detect whether the sample is known or unknown. In their case, each known class distribution is modeled as a Gaussian using the training data.

Recently, systematic approaches have been presented to break the closed set limitation by explicitly modeling the uncertainty mass belonging to the unknown distribution. One of the representative work inline with this is the evidential deep learning (EDL) model [121]. EDL treats the predicted multi-class probability as a multinomial opinion by leveraging the subjective logic principle. Similar to this work, Malinin et al. [92] propose Prior Networks (PNs) that explicitly consider the distributional uncertainty to quantify the distribution mismatch. Despite having a natural way to quantify the uncertainty, both of these methods require OOD data samples for model training, which is less practical. Charpentier et al. [17] propose the posterior networks that leverage the normalizing flows for density estimation in the latent space in order to predict the posterior distribution by only using in-distribution samples. Despite the significant progresses in OSD, limited attention has been drawn to the scenario, where the close set involves highly imbalanced classes, which is common in practical settings, such as anomaly detection, medical diagnosis, and so on. In such cases, existing standard ERM based approaches may not learn properly from the minority class due to lack of positive samples resulting in the mis-identification of a minority-class sample as an unknown class samples. Few recent works try to tackle this fundamental challenge based on the assumption that visual similarity exists between head and tail classes in the close set [85]. A model is designed to leverage this similarity to make it more robust for recognizing minority class samples. However, such an assumption may not universally hold, which limits the applicability of the model in general settings.

Chapter 3

Robust Multiple Instance Learning for Anomaly Detection

In this chapter, we propose two MIL approaches for anomaly detection. The goal of both techniques is to address the limitations present in the existing MIL-based techniques. Our first approach tries to address the challenging problem of selecting a suitable k -value in the top- k based approaches. Our proposed approach uses the DRO constraint in the Gaussian process (GP) mixture framework which can essentially serve as a soft version of the top- k constraint and therefore, removing the need to specify a fixed k value while ensuring a more stable (and robust) prediction performance. The GP in our technique offers a natural way to capture the interaction among the instances. Also, the Bayesian nature of the GP outputs the predictive uncertainty in a principled way which could be used for the decision-making process whenever the model is uncertain. Further, considering the limitation of GP which is restricted by a kernel with a fixed basis function, we propose to integrate the DNN with the GP while enabling end-to-end training. The powerful DNN deals with the high dimensional data and provides the adaptive basis function to GP that enables to capture the of complex patterns and interactions. Through multiple real-world benchmark datasets, we empirically show that our powerful Bayesian approach has a better as well as robust performance compared to the competitive top- k variants.

Although the DRO-based proposed approach avoids the issue of setting k value while stabilizing the performance, we may still need to manually set the size of the uncertainty set controlled by the radius (*i.e.*, η) of the uncertainty ball. Furthermore, both the DRO-based approach as well as top- k variants may put more focus on a set of consecutive segments with the highest prediction scores

and ignore some other potentially positive segments resulting in degradation in the performance. To overcome those issues, we next propose a novel submodular set function in a non-parametric way by inferring the diversity from data automatically. In this proposed work, we design a special submodular set functions that enables the discovery of a representative set of a frame from a video and thereby avoiding only choosing visually similar consecutive video frames. Next, our novel HDP prior to the state transition distribution of a HMM helps to automatically identify the optimal number of frames required to consider from each abnormal video. This makes the number of frames to be considered dynamic without needing to set any k that may have otherwise resulted in the suboptimal performance.

3.1 Distributionally Robust Optimization for Deep Kernel Multiple Instance Learning

In MIL, there is a collection of *positive* and *negative* bags where each bag consists of several instances. A bag is considered to be positive if at least one of the instances is positive and negative if none of the instances are positive [27]. Among many other useful applications, such as text classification and protein identification, MIL offers a particularly powerful tool to some important computer vision tasks, such as video anomaly detection, where the models have to solely rely on video level labels due to the lack of expensive frame-level labels [129, 160].

Various approaches have been developed to tackle the MIL problem by treating it as a missing-label problem [146, 163]. Those classical MIL techniques focus on the most positive instance (often referred to as the *witness*), instead of simultaneously considering multiple instances from a positive bag. In particular, the most positive instance is the one mainly responsible for determining the label of a bag [62]. For instance, in SVM based techniques [3], they maximize the margin of the instance with the most positive confidence w.r.t. the current model. Different from other works, a graph-based approach is developed to capture the interactions between instances within a bag and thereby using the information of multiple instances [162].

For many MIL tasks such as video anomaly detection, it is important to capture the interactions among frames in a video to correctly identify the abnormal frames given the temporal and spatial relationship naturally embedded in the data [62]. Further, due to the lack of instance-level labels, the model prediction may be much more uncertain and the uncertainty information is essential for many critical domains (e.g., security surveillance) [47]. Gaussian processes (GP) offer a natural way to capture the interactions among the instances through its covariance function. The Bayesian na-

ture of GP outputs the predictive uncertainty in a principled way. In addition, as a non-parametric model, GP allows the modeling power to scale well with the increase in the dataset. By leveraging these modeling advantages, a number of GP based MIL models have been developed, where maximum score from positive and negative bags are considered in the for model training [47, 62].

However, there are two key limitations with using a maximum score. First, the presence of noisy outliers may significantly impact the overall performance. This is because the defined objective function solely focuses on an individual instance with the highest score from positive and negative bags. Second, if a multimodal situation (e.g., multiple types of abnormal events in a single video) presents, maximum score based approaches may only detect one type of positive instances due to its inability to consider multiple instances from a single bag in the training process.

To address these limitations, we propose a general GP mixture framework that assigns a non-zero probability to each instance in a bag through a latent mixture model. By adding a top- k constraint, it is equivalent to choosing the top- k most positive instances in a bag, making it more robust to outliers and multimodal scenarios. Most importantly, we further integrate a Distributionally Robust Optimization (DRO) constraint that relaxes the limitation of specifying a fixed k value. By combining DRO with a Bayesian non-parametric GP, *this is the first work that develops a Bayesian DRO model for MIL*. To ensure the prediction power over high-dimensional data that are common in MIL problems, we augment the GP kernel with fixed basis functions by using a deep neural network to perform deep kernel (DK) learning [149]. As a result, it learns adaptive basis functions so that the covariance structure of high-dimensional input data can be accurately captured. Finally, different components of the proposed DRO-DKMIL model are jointly optimized through stochastic variational inference (SVI) that leverages local kernel interpolation and structure exploiting algebra [147] to conduct end-to-end model training, ensuring good efficiency and scalability. In summary, our key contribution is fourfold:

- a general GP mixture framework for MIL that gives flexibility for each instance to take non-zero membership probability in each bag,
- a novel Bayesian DRO MIL model that ensures the participation of multiple instances from each bag in model training, making the prediction robust to outlier and multimodal scenarios,
- the first approximate inference algorithm to train the new Bayesian DRO model in MIL setting,
- state-of-the-art prediction performance that outperforms all existing competitive MIL models.

Experiments are conducted on three challenging real-world video anomaly detection datasets with varied scales: UCF-Crime [129], ShanghaiTech [82], and Avenue [87]. Results show that DRO-DKMIL achieves best performance in all cases.

3.1.1 Related Work

Work related to Multiple Instance Learning (MIL) and Distributionally Robust Optimization (DRO) are described in Section 2. In this section, we will be describing the work related to deep kernel learning.

Deep Kernel Learning (DKL). DKL provides a powerful learning paradigm by combining the non-parametric flexibility of kernel methods (e.g., GP) and representation learning ability of deep neural networks. State-of-the-art performance has been demonstrated over multiple supervised learning tasks [147, 149]. One of the key challenges comes from the computation bottleneck of GP which can work only for a few thousand data points [150]. Such issue has been alleviated through structure exploiting techniques [21, 26, 139], local kernel interpolation [61], and other advances in this field. Building upon these efforts, we develop a stochastic variational inference (SVI) algorithm to conduct end-to-end model training to ensure good efficiency and scalability under the MIL setting.

3.1.2 DRO Deep Kernel Multiple Instance Learning (DRO-DKMIL)

We consider that each bag has a fixed number of instances. For a positive bag, there is at least one positive instance whereas for a negative bag all instances are normal. Let $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ be a set training instances. Each $\mathbf{x}_n \in \mathcal{R}^D$ is a D -dimensional feature vector associated with bag $b \in [1, B]$ with corresponding label t_b indicating its bag type, where $+1$ denotes positive and -1 otherwise. Further, consider $\mathbf{y} = \{y_1, \dots, y_B\}$ be the set of predicted labels. Table 3.1 shows the notations used in this section.

DRO-GP MIL

For most MIL problems, it is essential to capture the interactions among instances in the same bag (e.g. frames in a video) as the spatially and/or temporally close instances usually belong to the same event, which should be assigned the same labels. Further, capturing the uncertainty associated with each instance is crucial in MIL tasks such as anomaly detection from surveillance videos. Gaussian Processes (GP) naturally capture the interactions among instances through its

Table 3.1: Symbols with Descriptions

Notation	Description
\mathbf{X}	Set of training bag instances
B	Total number of bags in a training set
\mathbf{y}	Set of predicted probabilities of B bags
t_b	Binary label for bag b
\mathbf{f}_b	Set of functional values of instances in a bag b
\mathbf{z}_b	Indicator variable drawn from a multinomial distribution
n	Total number of instances in each bag
Q	DNN final layer (L) feature representation of each bag instance
\mathbf{w}	DNN parameters
u^j	Inducing variables for j^{th} GP
$\boldsymbol{\mu}_j$	Posterior distribution mean for u^j
\mathbf{S}_j	Posterior distribution co-variance matrix for u^j
A	Mixing parameter to combine J-GPS functional values
\mathbf{U}	Set of inducing variables for J GPS
\mathbf{Z}	Set of multinomial variables for bags B
\mathbf{F}	Set of functional values for bags B
M	Sparse interpolation matrix
\mathbf{r}_b	Posterior distribution parameter for \mathbf{z}_b
η	Hyper-parameter used to define the ball radius in DRO framework
T	Total number of likelihood samples used
P	Mini-batch of bag size
\mathbf{L}	Lower diagonal matrix with real and positive entries
$\boldsymbol{\pi}_b$	Prior distribution parameter for \mathbf{z}_b
N	Total number of segments in a training set
D	Feature dimension of each bag instance
\mathcal{N}	Gaussian distribution
\mathbb{E}_q	Expectation with respect to the distribution q
$\mathbf{K}_{A,B}$	Kernel matrix computed between A and B
R	Total number of inducing inputs considered in each GP
$L(q)$	Marginal likelihood lower bound with variational distribution q
\otimes	Kronecker decomposition operator

covariance function and its non-parametric flexibility allows the modeling power to scale well with the increase of data. Being a Bayesian model, GP also directly outputs the predictive distribution that quantifies the prediction uncertainty in a principled way.

We propose a GP based mixture framework to address the limitation of existing models as discussed earlier. By integrating GP with a latent mixture model, the proposed framework assigns a non-zero membership probability for each instance present in a bag resulting in robustness to the outlier and multimodal scenarios.

We start by defining the bag level likelihood:

$$p(y_b | \mathbf{f}_b, \mathbf{z}_b) = \prod_{i=1}^n \left\{ \frac{1}{1 + \exp(-t_b f_{bi})} \right\}^{z_{bi}} \quad (3.1)$$

$$p(\mathbf{z}_b | \boldsymbol{\pi}_b) = \prod_{i=1}^n \pi_{bi}^{z_{bi}}, \pi_{bi} \geq 0, \sum_i \pi_{bi} = 1 \quad (3.2)$$

where \mathbf{z}_b is an indicator variable drawn from a multinomial distribution parameterized by $\boldsymbol{\pi}_b, \forall b \in [1, B]$. For a negative bag with $t_b = -1$, the model is expected to output a small *score* f_{bi} (which can be negative) to maximize the bag level likelihood. In contrast, f_{bi} will be high for a positive bag with $t_b = 1$. Since $\pi_{bi} \geq 0$, each instance has a chance to be predicted as positive.

We denote $\mathcal{P}_{\boldsymbol{\pi}_b, n}$ as an uncertainty set, defining the constraints over the mixing coefficient $\boldsymbol{\pi}_b$. Without adding any additional constraints other than being non-negative and summing to one, we have $\mathcal{P}_{\boldsymbol{\pi}_b, n}^{max} := \{\boldsymbol{\pi}_b \in R^n : \boldsymbol{\pi}_b^T \mathbf{1} = 1, 0 \leq \boldsymbol{\pi}_b\}$. It turns out performing multiple instance learning under the GP mixture framework with constraints $\mathcal{P}_{\boldsymbol{\pi}_b, n}^{max}$ is equivalent to a maximum score based model.

Lemma 3.1. *With \mathcal{P}^{max} as constraints, MIL under the GP mixture framework only considers the most positive instance (equivalent to maximum score MIL).*

Proof. Marginalizing over \mathbf{z}_b leads to the marginal likelihood of the bag-level label:

$$p(y_b | \mathbf{f}_b, \boldsymbol{\pi}_b) = \sum_{i=1}^n \pi_{bi} \frac{1}{1 + \exp(-t_b f_{bi})} \quad (3.3)$$

Denote $p(f_{bi}) = \frac{1}{1 + \exp(-t_b f_{bi})}$ and maximizing (3.3) over $\boldsymbol{\pi}_b$ leads to

$$\pi_{bi} = \begin{cases} 1, & \text{if } p(f_{bi}) = \max_{i \in b} p(f_{bi}) \\ 0, & \text{otherwise} \end{cases} \quad (3.4)$$

Thus, the bag level likelihood is given by:

$$p(y_b|\mathbf{f}_b) = \max_{i \in b} p(f_{bi}) \quad (3.5)$$

which only relies on the most positive instance [62]. \square

To more effectively involve multiple instances, we can instead consider a top- k constraint, given by

$$\mathcal{P}_{\pi_b, n}^{\text{top-}k} := \{\pi_b \in R^n : \pi_b^T \mathbf{1} = 1, 0 \leq \pi_{bi} \leq \frac{1}{k}\} \quad (3.6)$$

where k indicates the number of instances being potentially positive.

Lemma 3.2. *With $\mathcal{P}^{\text{top-}k}$ as constraints, MIL under the GP mixture framework considers the top- k most positive instances (equivalent to average top- k MIL).*

Proof. Maximizing (3.3) under $\mathcal{P}_{\pi_b, n}^{\text{top-}k}$ constraints gives

$$\pi_{bi} = \begin{cases} \frac{1}{k}, & \text{if } p(f_{bi}) \geq p(f_{b[k]}) \\ 0, & \text{otherwise} \end{cases} \quad (3.7)$$

where $p(f_{b[k]})$ indicates the instance with the k^{th} highest value. Thus, the bag level likelihood is

$$p(y_b|\mathbf{f}_b) = \frac{1}{k} \sum_{i=1}^k \frac{1}{1 + \exp(-t_b f_{b[i]})} \quad (3.8)$$

which is collectively determined by the top- k instances with largest scores. \square

Lemma 3.2 shows that by leveraging the top- k constraint, the GP mixture framework involves the top- k most positive instances that can effectively overcome outlier and multimodal situations. However, a remaining issue is how to set a suitable k , which can be quite challenging in practice. More importantly, since k takes discrete values, the prediction performance may fluctuate significantly when k changes. To address this fundamental challenge, we propose to integrate a DRO constraint into the GP mixture framework, which can essentially function as a soft version of the top- k constraint, thus removing the need to specify a fixed k value while ensuring a more stable (and robust) prediction. More specifically, the DRO constraint restricts π_b within a certain ball with a center given by the uniform distribution [98]:

$$\mathcal{P}_{\pi_b, n}^{\text{DRO}} := \{\pi_b \in R^n : \pi_b^T \mathbf{1} = 1, \pi_b \geq 0, D_f(\pi_b || \frac{\mathbf{1}}{n}) \leq \eta\} \quad (3.9)$$

where η controls the radius of a ball and D_f is the f-divergence. A large η gives more flexibility on π_b , which allows it to deviate significantly from the uniform distribution so that one single instance

may play a dominant role in the bag level likelihood (equivalent to maximum score MIL when $\eta \rightarrow \infty$); a small η leads to near equal probability for each instance (equivalent to averaging overall all instances in a bag when $\eta \rightarrow 0$).

Deep Kernel MIL

While a GP has the non-parametric flexibility along with its Bayesian nature to capture model uncertainty, it is restricted by the kernels with fixed basis functions that are less effective when applied to high dimensional data. To address this issue, one viable solution is to integrate a deep neural network (DNN), which uses adaptive basis functions to learn the rich representations from high dimensional input data.

In terms of network architecture, the proposed deep kernel multiple instance learning consists of three main components: (1) deep neural network, (2) additive Gaussian Processes, and (3) mixing model. For each instance ($\mathbf{x}_i \in \mathcal{R}^D$) present in a bag b , we perform non-linear transformation using a mapping function $\mathbf{h}(\mathbf{x}, \mathbf{w})$ parameterised by neural network weights \mathbf{w} to generate Q -dimensional features at the final layer L , i.e., $h_i^{L1}, \dots, h_i^{LQ}$. Next, we use J Gaussian Processes with corresponding base kernels k_1, \dots, k_J applied to the subset of those extracted features constituting an additive GP model [150]. As the base kernels act on low dimensional inputs, local kernel interpolation (for scalability) become more natural. The resulting GP functional values from J-GPS (f_i^1, \dots, f_i^J) are linearly mixed by a training matrix $A \in R^{J \times 1}$ to produce a single functional value f_i . Finally collecting the functional values for all instances present in a bag b , we arrive at the bag-level likelihood in (3.1).

For the j^{th} Gaussian process in the additive GP layer, let $\mathbf{f}^j = \{f_i^j\}_{i=1}^N$ be the latent functions on the input data features for all the instances in a bag. By introducing a set of latent inducing variables \mathbf{u}^j indexed by m inducing inputs [109] (denoted as R), we have

$$p(\mathbf{f}^j | \mathbf{u}^j) = \mathcal{N}(\mathbf{f}^j | \mathbf{K}_{X,R}^j \mathbf{K}_{R,R}^{j-1} \mathbf{u}^j, \hat{\mathbf{K}}^j), \quad \hat{\mathbf{K}} = \mathbf{K}_{X,X} - \mathbf{K}_{X,R} \mathbf{K}_{R,R}^{-1} \mathbf{K}_{R,X} \quad (3.10)$$

where $X \in R^{N \times Q}$ is the feature representation learned from N training instances through DNN. Performing the local interpolation approximation (similar to [150]) $\mathbf{K}_{X,X} \approx \mathbf{M} \mathbf{K}_{R,R} \mathbf{M}^T$, $\hat{\mathbf{K}}^j$ becomes zero, yielding $\mathbf{f}^j = \mathbf{K}_{X,R} \mathbf{K}_{R,R}^{-1} \mathbf{u} = \mathbf{M} \mathbf{u}$, where \mathbf{M} is $N \times m$ matrix of interpolation weights that can be extremely sparse with the relationship $\mathbf{K}_{X,R} \approx \mathbf{M} \mathbf{K}_{R,R}$. This means with the help of local interpolation along with inducing points, we can obtain a deterministic relationship between \mathbf{f} and \mathbf{u} governed by the sparse matrix \mathbf{M} .

Denote $\mathbf{U} = \{\mathbf{u}^j\}_{j=1}^J$ as the collection of inducing variables for J additive GPs along with the pos-

terior distribution as $q(\mathbf{U}) = \prod_{j=1}^J \mathcal{N}(\mathbf{u}^j | \boldsymbol{\mu}_j, \mathbf{S}_j)$. Further, let $q(\mathbf{z}_b | \mathbf{r}_b) = \prod_{i=1}^n r_{bi}^{z_{bi}}$ be the posterior distribution for a multinomial variable corresponding to a bag b parameterized by \mathbf{r}_b . To update: (1) variational parameters ($\{\boldsymbol{\mu}_j, \mathbf{S}_j\}_{j=1}^J, \{r_{bi}\}_{i=1}^n; \forall b \in [1, B]$) (2) GP kernel hyper-parameters, (3) $\{\pi_{bi}\}_{i=1}^n; \forall b \in [1, B]$, (4) mixing coefficients A , and (5) neural network parameters \mathbf{w} , we optimize a lower bound of the marginal likelihood using an efficient stochastic variational procedure.

Stochastic Variational Inference

Exact inference and parameter learning with a non-Gaussian bag level likelihood is intractable. We develop the first stochastic variational inference method that combines a fast sampling scheme to work on a mini-batch setting to ensure efficient and scalable end-to-end training of the new DRO-DK-MIL model.

We start by defining the log marginal bag-level likelihood and applying Jensen’s inequality

$$\log p(\mathbf{y}) = \log \int p(\mathbf{y}, \mathbf{Z}, \mathbf{F}, \mathbf{U}) d\mathbf{Z} d\mathbf{F} d\mathbf{U} \geq \mathbb{E}_q[\log p(\mathbf{y}, \mathbf{Z}, \mathbf{F}, \mathbf{U})] - \mathbb{E}_q[\log q(\mathbf{Z}, \mathbf{F}, \mathbf{U})] \quad (3.11)$$

We formally define the lower bound as

$$\begin{aligned} L(q) &\triangleq \mathbb{E}_q[\log p(\mathbf{y}, \mathbf{Z}, \mathbf{F}, \mathbf{U})] - \mathbb{E}_q[\log q(\mathbf{Z}, \mathbf{F}, \mathbf{U})] \\ &= \mathbb{E}_q[\log p(\mathbf{y} | \mathbf{Z}, \mathbf{F})] - KL(q(\mathbf{Z}) || p(\mathbf{Z})) - KL[q(\mathbf{U}) || p(\mathbf{U})] \end{aligned} \quad (3.12)$$

where $KL(P || Q)$ is the KL divergence between two distributions P and Q .

As likelihood function presented in (3.12) factorizes over each bag, i.e., $p(\mathbf{y} | \mathbf{Z}, \mathbf{F}) = \prod_{b=1}^B p(y_b | \mathbf{f}_b, \mathbf{z}_b)$, we can optimize the lower bound in a minibatch setting. The variational parameters corresponding to $q(\mathbf{U})$, kernel hyper-parameter parameters, mixing coefficients A , and neural network parameters are updated using SGD through the noisy approximation of the gradient of the lower bound on mini-batches, as detailed below.

Update $q(\mathbf{Z})$. To update $q(\mathbf{Z})$, we further simplify (3.12) by absorbing terms that do not depend on \mathbf{Z} to a constant term,

$$L(q(\mathbf{Z})) \triangleq \mathbb{E}_{q(\mathbf{U})q(\mathbf{Z})}[\log p(\mathbf{y} | \mathbf{F}, \mathbf{Z})] - \mathbb{E}_{q(\mathbf{Z})}[\log(q(\mathbf{Z}))] + \mathbb{E}_{q(\mathbf{Z})}[\log(p(\mathbf{Z}))] + \text{const} \quad (3.13)$$

By taking derivative with respect to r_{bi} , we have

$$r_{bi} = \pi_{bi} \exp(\mathbb{E}_{q(\mathbf{U})}[\log(p(t_b f_{bi}))]), \quad \forall i \in [1, n] \quad (3.14)$$

As long as $\pi_{bi} \geq 0$, we have $r_{bi} \geq 0$. To satisfy the second constraint $\sum_{i=1}^n r_{bi} = 1$, we normalize it as

$$r_{bi} = r_{bi} / \sum_{j=1}^n r_{bj} \quad (3.15)$$

Update $\boldsymbol{\pi}$. To update $\boldsymbol{\pi}_b$, we focus on $\mathbb{E}_{q(\mathbf{Z})}[\log(p(\mathbf{Z}))]$, which is the only term as a function of $\boldsymbol{\pi}_b$ and proceed as

$$\max_{\boldsymbol{\pi}_b} \mathbb{E}_{q(\mathbf{z})} \sum_{i=1}^n z_{bi} \log(\pi_{bi}) = \max_{\boldsymbol{\pi}_b} \sum_{i=1}^n r_{bi} \log(\pi_{bi}) \quad (3.16)$$

where $\mathbb{E}_{q(\mathbf{z})}[z_{bi}] = r_{bi}, \forall i \in [1, n]$. It should be noted that maximization of the above objective function is performed under the DRO constraints in (3.9).

Update $q(\mathbf{U})$. Due to the non-Gaussian bag-level likelihood function in (3.12), expectation cannot be evaluated analytically. Therefore, we use a sampling method, which is proven to be highly efficient with structured reparametrization, local kernel interpolation, and structure exploiting algebra [147, 150]. Using the local kernel interpolation, the latent function \mathbf{f} is expressed as a deterministic local interpolation of the inducing variables \mathbf{u} and therefore, allowing us to make the difficult posterior approximation over \mathbf{f} easier. As such, we can perform direct reparameterization over $q(\mathbf{U})$ and compute \mathbf{f} directly through interpolation $\mathbf{f}^t = \mathbf{M}\mathbf{u}^t$ (for notation simplicity, we have omitted the index j corresponding to j^{th} GP). Using Cholesky decomposition for the covariance matrix of $q(\mathbf{U})$: $\mathbf{S} = \mathbf{L}^T\mathbf{L}$, we have the sampling procedure:

$$\mathbf{u}^t = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{\epsilon}^t, \quad \boldsymbol{\epsilon}^t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (3.17)$$

where each step of the above standard sampler has a complexity of $O(m^2)$ with m inducing points.

As the above sampling procedure requires a matrix-vector product, it may become expensive with many inducing points which are required for large datasets with a high dimensional input [150]. To further scale up the sampling procedure, we can take the advantage of both Toeplitz and circulant structure along with the Kronecker decomposition on $\mathbf{L} = \bigotimes_{d=1}^D \mathbf{L}_d$ with D being the input dimension of the base kernel.

As both KL divergence terms have a closed form, only the bag-level likelihood function requires sampling for the expectation computation. With T samples of \mathbf{u} and mini-batch of bag size P , we can estimate the marginal likelihood lower bound as:

$$L(q) \approx \frac{1}{TP} \sum_{t=1}^T \sum_{b=1}^P \sum_{i=1}^n z_{bi}^t \log \left(\frac{1}{1 + \exp(-t_b f_{bn}^t)} \right) - KL[q(\mathbf{U})||p(\mathbf{U})] - KL[q(\mathbf{Z})||p(\mathbf{Z})] \quad (3.18)$$

Table 3.2: Video Level Distribution on Different Datasets

Split	ShanghaiTech		UCF-Crime		UCF-Crime Multimodal		Avenue	
	<i>Normal</i>	<i>Abnormal</i>	<i>Normal</i>	<i>Abnormal</i>	<i>Normal</i>	<i>Abnormal</i>	<i>Normal</i>	<i>Abnormal</i>
Train	175	63	810	800	150	150	13	17
Test	155	44	150	140	30	30	3	4

where we can efficiently compute the $KL(q(\mathbf{U})||p(\mathbf{U}))$ term and its gradient with the Kronecker method (details are provided in the Appendix).

Update other parameters. Update of other parameters, including the mixing coefficients A , kernel hyperparameters, variational parameters $\{\boldsymbol{\mu}, \{\mathbf{L}_d\}_{d=1}^D\}$, and neural network parameters, can be achieved through gradient decent as detailed in the Appendix.

3.1.3 Experiments

We conduct extensive experiments to evaluate the proposed DRO-DKMIL model. We first introduce three real-world video datasets for anomaly detection. Anomaly detection is regarded as one of the most challenging computer vision tasks under the MIL setting. Next, we demonstrate the overall performance of DRO-DKMIL and compare it with existing state-of-the-art video anomaly detection models. Further, we assess the effectiveness of our proposed model in multimodal and outlier scenarios. We also provide a qualitative analysis to justify the superior performance of our model. Finally, we investigate the impact of the key parameters to the model performance. The GitHub repository that includes the source code and detailed documentation can be accessed via this link.

3.1.4 Datasets and Experimental Settings

Datasets. Our experiments involve three anomaly detection video datasets of different scales: ShanghaiTech [90], Avenue [87], and UCF-Crime [129]. On those videos, the assumption is that in the training set, frame level annotation is missing and only video level information (indicating whether the video is of abnormal type or normal type) is available.

- **ShanghaiTech** consists of 437 videos with 330 normal and 107 abnormal videos. In the original setting, all training videos are normal. To fit into our scenario, we follow the data split in [159] to assign normal and abnormal videos in both training and testing sets.

- **Avenue** consists of 16 training and 21 testing videos. We perform 80:20 split separately in the abnormal and normal video sets to generate training and testing instances.
- **UCF-Crime** consists of 13 different anomalies with a total of 1900 videos: 1610 for training and 290 for testing. In this dataset, frame labels are available only for the testing videos.

Table 3.2 shows how the videos are partitioned into the training and testing sets in each dataset.

Evaluation metric and model training. For evaluation, we report the frame-level receiver operating characteristics (ROC) curve along with the corresponding AUC score, which captures the robustness of the prediction performance at varying thresholds. For the Avenue and ShanghaiTech datasets, we extract the visual features from FC7 layer of a pre-trained C3D network [138]. To extract the features, we first re-size each video frame to 240×340 pixels and fix the frame rate to 30fps. Next, we use a pre-trained C3D model to compute the C3D features for every 16-frame video clip. This may yield a different number of clips (each clip having 2048 dimensional feature vector) depending on the number of frames in each video. Thus, we fit any number of clips to the 32 segments by taking an average of clip features in a specific segment.

In terms of the DNN architecture, we follow the 2-dimensional neural network followed by the GP base kernels. The first FC layer has 32 units followed by 16 units. We adopt a 60% dropout regularization between FC layers along with the ReLU activation. For the UCF-Crime dataset, we extract features using I3D network [13]. We uniformly sample 1512 frames and pass an 8-frame video clip into the network. This yields 189 segment clips each with 1024 dimensional feature vector. For this dataset, we use a 5-layer LSTM network, where each layer has 189 hidden units followed by a batch normalization layer and FC layer of 16 nodes. Finally, base GP kernels are applied to the DNN output features. In the uncertainty set of parameter π , we define the f-divergence as a Kullback-Leibler (KL)-divergence. For hyper-parameter η , we conduct a grid search in a range from 10^{-9} to 1.0 and find the one with best validation AUC score as the optimal η value. The details about η value selection and its impact are provided in the Appendix. For DNN training, we use SGD with a learning rate of 0.001 and l_2 regularization with parameter $\lambda = 0.001$ whereas, for variational parameters, mixing coefficient (A), and hyper-parameters, we use a learning rate of 0.1.

Performance Comparison

In our comparison study, we include baselines that are used in the video anomaly detection tasks. We also compare with the maximum score based GP model [62] but augment it with deep kernel learning to properly handle high-dimension data (referred as DK-MMIL). We further implement the variational inference algorithm developed in [47] and refer to this model as VGPMIL. We also

Table 3.3: Comparison of AUC Scores

Approach	AUC (%)
UCF-CRIME	
Hasan et al. [46] (C3D)	50.60
Lu et al. [87] (C3D)	65.51
Lu et al. [87] (I3D)	61.98
Sultani et al. [129] (C3D)	75.41
Ilse et al. [52] (I3D)	76.52
Zhong et al. [159] (GCN (C3D))	81.08
Zhong et al. [159] (TSN^{RGB})	82.12
Zhong et al. [159] ($TSN^{OpticalFlow}$)	78.08
Haußmann et al. [47] VGPMIL (I3D)	79.56
DK-MMIL (I3D)	82.32
DK-TKMIL (I3D)	82.66
DRO-DKMIL (I3D)	85.93
SHANGHAI TECH	
Lu et al. [87] (C3D)	72.90
Zhong et al. [159] (GCN (C3D))	76.44
Zhong et al. [159] (TSN^{RGB})	84.44
Zhong et al. [159] ($TSN^{OpticalFlow}$)	84.13
Ilse et al. [52] (C3D)	85.78
Haußmann et al. [47] VGPMIL (C3D)	87.78
DK-MMIL (C3D)	92.00
DK-TKMIL (C3D)	92.30
DRO-DKMIL (C3D)	94.39
AVENUE	
Lu et al. [87] (C3D)	62.14
Ilse et al. [52] (C3D)	72.39
Haußmann et al. [47] VGPMIL (C3D)	72.84
DK-MMIL (C3D)	73.93
DK-TKMIL (C3D)	75.12
DRO-DKMIL (C3D)	78.66

compare with the average top- k constraint as introduced in Lemma 2 (refer to as DK-TKMIL) with a pairwise hinge loss (details are provided in the Appendix). In addition, for each dataset, we also include other competitive models that have been applied to that dataset.

UCF-Crime. Table 3.3 shows the AUC scores of all competitive techniques. As can be seen, DRO-DKMIL has superior performance compared to other existing techniques. The corresponding ROC performance is shown in Figure 3.1 (a). As shown, DRO-DKMIL has higher TPR for all FPR

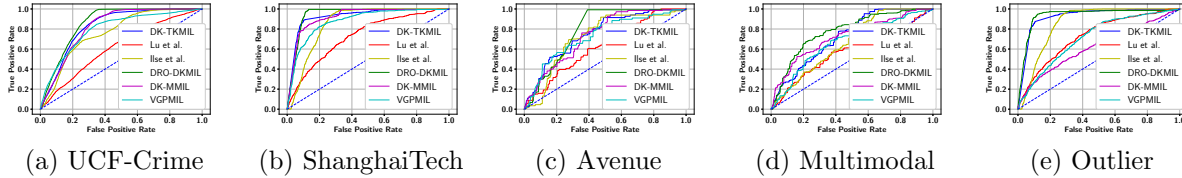


Figure 3.1: ROC Performance on Three Video Datasets (a)-(c); Multimodal (d) and Outlier Prediction (e)

below 0.5, which demonstrates the robustness of the approach.

ShanghaiTech. Besides the common baselines, we also compare our method with the recent GCN based model using three feature extractors ($C3D$, TSN^{RGB} , $TSN^{OptimalFlow}$) [159]. The result is reported in Table 3.3. The corresponding ROC curves are shown in Figure 3.1 (b). The result shows that DRO-DKMIL significantly outperforms other competitive methods.

Avenue. Table 3.3 summarizes the AUC scores on Avenue of the proposed approach along with other techniques. The result confirms that DRO-DKMIL outperforms all existing techniques. The corresponding ROC performance is shown in Figure 3.1 (c). Similarly, the proposed approach achieves higher recall compared to other approaches.

Multimodal and Outlier Detection

In this section, we assess the effectiveness of the proposed DRO-DKMIL in outlier and multimodal settings. For this, we create a multimodal scenario by extending the UCF-Crime dataset. For the outlier scenario, we deliberately impose some outliers in the ShanghaiTech dataset and evaluate the performance.

Multimodal Detection. The original UCF-Crime dataset does not explicitly consider the multimodal scenario. Although it is natural to have multimodal scenario in the real-world videos (as evidenced by the superior performance of the proposed model), it is hard to identify the actual videos for this specific evaluation. In case of UCF-Crime, we have abnormal videos categorized into different activity types. Therefore, we create a multimodal scenario by combining multiple abnormal videos from different anomaly types. To create a multimodal scenario, we randomly select three activity types. Then, we form a positive (abnormal) bag by concatenating three abnormal videos, one video per activity type. To construct a normal bag, we randomly pick three normal videos and concatenate them. In the process, the training bags are constructed using training videos only and testing bags are constructed using testing videos only. The corresponding video statistics is shown in the Table 3.2. Each bag is a concatenation of three videos yielding total 50 abnormal and

50 normal bags in the training set. The testing set consists of 10 normal and 10 normal videos. Table 3.4 reports the AUC scores and the ROC plot is shown in Figure 3.1 (d). We can observe that the ROC curve from DRO-DKMIL clearly outperforms all baselines. This means, compared to the baselines, our approach is more robust to the multimodal scenario at various thresholds.

Table 3.4: AUC on Multimodal and Outlier Detection

Approach	AUC (%)	
	<i>Multimodal</i>	<i>Outlier</i>
Lu et al. [87] (C3D)	58.67	72.90
Ilse et al. [52] (C3D)	66.85	85.65
Haußmann et al. [47] (C3D)	67.16	71.31
DK-MMIL (C3D)	72.44	62.89
DK-TKMIL (C3D)	72.75	92.61
DRO-DKMIL (C3D)	77.89	93.49

Outlier Detection. To test the robustness of the proposed approach with outliers, we extend the ShanghaiTech dataset by explicitly including outliers. Specifically, we randomly select 120 segments from abnormal videos and replace their features with points drawn from a standard multivariate Gaussian distribution. As shown in Table 3.4, DK-MMIL suffers heavily by the outliers compared to the proposed DRO-DKMIL. This is because, it is likely to have an outlier predicted as the maximum prediction score from an abnormal video. As a result, the overall training process may be heavily influenced by outliers. However, as our approach gives chance to other actual abnormal segments as well in the training process, it makes the model robust to the outliers. It is also noted that DK-TKMIL performs very well with outliers, which benefits from the top- k constraints. However, setting a proper k value is highly challenging in practice and the prediction performance fluctuates significantly with the change of k (see Appendix for details).

Qualitative Analysis

To get deeper insight regarding the effectiveness of our approach, we analyze videos where the proposed DRO-DKMIL and maximum score-based DK-MMIL provide different predictions. Figure 3.2 shows abnormal frames from two videos in which DRO-DKMIL correctly predicts as abnormal whereas DK-MMIL fails. The resulting prediction scores for all abnormal frames for the videos ARREST001 and EXPLOSION010 are shown in Figure 3.3. The prediction threshold (shown as a horizontal line) in each approach is determined such that FPR is maintained at 0.3. As shown in the video ARREST001, DK-MMIL fails to detect the abnormal frames near the transition phase. Since transitioning frames may be far from the abnormal frame with the maximum prediction score,

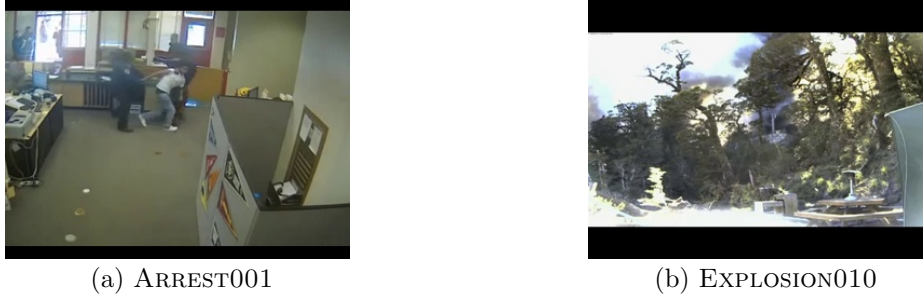


Figure 3.2: Abnormal Frames Identified by DRO-DKMIL but not DK-MMIL

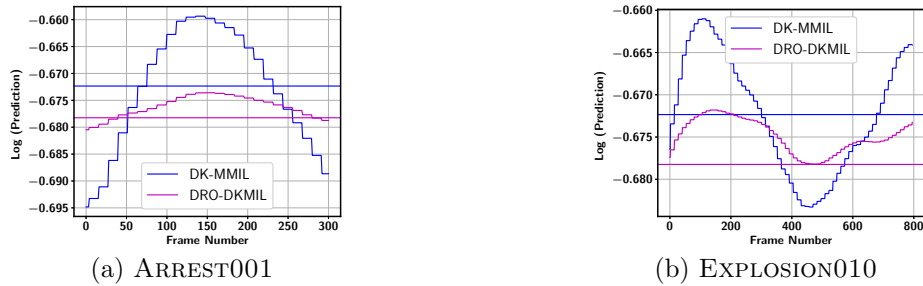


Figure 3.3: Abnormal Frame Prediction

DK-MMIL does not consider those types of abnormal frames during model training. However, for the proposed DRO-DKMIL, it is more likely to involve these transitioning abnormal frames in the training process. Thus, it can correctly identify similar frames during testing.

In the video EXPLOSION010, DK-MMIL fails to correctly identify the abnormal frames that are in the middle. This may be because more extreme abnormal frames of the explosion type may only participate in the training process. As a result, the maximum score based approach may not consider the frame as shown in Figure 3.2(b). However, the proposed approach may be more likely to involve this type of abnormal frames as it allows the participation of multiple abnormal frames from each abnormal video.

Uncertainty Analysis

Being a Bayesian model, the proposed DRO-DKMIL is able to accurately capture the prediction uncertainty, which provides important complementary information for video anomaly detection. The uncertainty score can guide a human decision maker to not only focus on the predicted positive frames to examine the abnormality but also pay attention to the highly uncertain areas in the videos that may also include important information to support decision-making. To show this, we use the Avenue dataset and report the standard deviation (SD) output by DRO-DKMIL for each testing

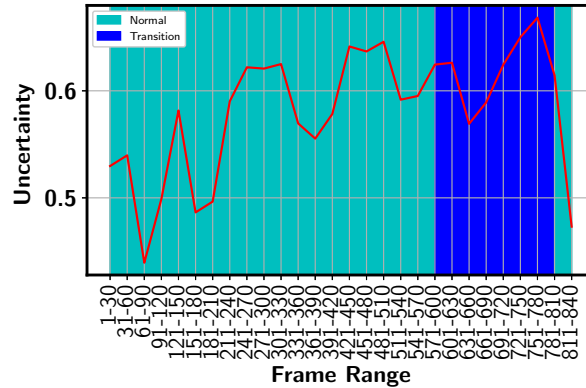


Figure 3.4: Uncertainty of Different Frames

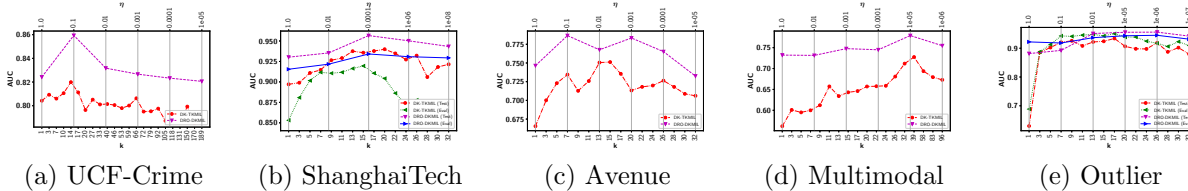


Figure 3.5: AUC Performance vs. η and Comparison with Average top- k

frame. We maintain $FPR = 0.3$ and identify all correct and incorrect instances. By setting a threshold as 0.67, we identify 70 incorrect and 42 correct frames with a SD above the threshold. This means a larger uncertainty score (i.e., sd) indicates a higher chance of prediction errors, which is a desirable property.

Figure 3.4 shows the uncertainty associated with different frames in Avenue video TEST-10. In the video, the first 569 are normal frames, where the model has a relatively high confidence. After that until frame 817, the transition occurs nine times between abnormal and normal frames. Therefore, we observe much higher uncertainty. As the transition is rapid, the consecutive frames may look very similar to each other, which may confuse the model, leading to a (correctly) predicted high uncertainty score for those frames.

3.1.5 Impact of Key Model Parameters

Impact of η . We analyze the impact of the hyperparameter η in the AUC score for all datasets (UCF-Crime, ShanghaiTech, Avenue, UCF Crime Multimodal, and ShanghaiTech outlier). For the ShanghaiTech and ShanghaiTech outlier, we use 20% of the original testing set to construct a validation set and use the rest to report the model performance. The propose of constructing the validation set is to determine the optimal η value. To get robustness in the performance,

we randomly choose the validation set 20 times producing 20 pairs of the validation-testing split. Figures 3.5 (b) and (e) show the validation and testing AUC change for the randomly selected test-validation pair from ShanghaiTech and ShanghaiTech Outlier datasets, respectively. For a lower η value, the model allows the participation of most of the frames. As a result, the model tries to make a prediction score of most of the frames from an abnormal video to be higher than from a normal video, resulting in the misclassification of many normal frames from abnormal videos. Therefore, we observe lower performance for a lower η value. As we increase η , the model limits the participation of the frames from both abnormal and normal videos. This increases the chances of including only abnormal frames while leaving out normal frames from abnormal videos in the optimization process. As a result, the model learns to have a higher score for the abnormal frames compared to all the normal frames, resulting in improvement in the performance. However, a very high η value allows the participation of a very limited number of abnormal as well as normal frames during the training process. As a result, the model may be highly influenced by outliers and multimodal scenarios. Therefore, we can see the degradation in the performance for a high η value.

For the UCF-Crime, Avenue, and Multimodal datasets, we directly report the performance in the testing dataset, instead of using a separate validation set. For UCF-Crime, the use of a separate validation set may not be effective because of the limited testing videos of a given type. Therefore, similar to [129], we evaluate the testing performance with respect to η and report the one with the best performance as the best η value. For the Avenue dataset, there are very limited testing videos, so determining the η value using a separate validation set may not be feasible. As can be seen in the Figures 3.5 (a), (c), and (d), the trend is similar to what we have observed in the ShanghaiTech dataset for the same reason explained above.

Comparison with Average top- k . As proved in Lemma 2, by using a top- k constraint, the proposed framework is equivalent to perform average top- k MIL. As the top- k most positive frames are simultaneously considered by the training process, it can potentially handle the outlier and multimodal scenarios as well. In this set of experiments, we further compare the proposed DRO-DKMIL with the average top- k model (the deep kernel version is referred to as DK-TKMIL). Figure 3.5 compares the AUC scores between DRO-DKMIL and DK-TKMIL while varying η and k . We have three key observations: (i) With a properly chosen k , DK-TKMIL achieves a decent prediction performance, especially when dealing with outliers, as shown in 3.5(e). (ii) DRO-DKMIL achieves even better prediction performance. In all datasets, the test AUC curve of DRO-DKMIL stays on top of DK-TKMIL for almost all different η and k values. (iii) In almost all datasets, the AUC score of DK-TKMIL changes more significantly when compared with the AUC score change

of DRO-DKMIL with η . In addition, η varies in a much wider range (i.e., 10^{-8} to 1) than the k values. This clearly confirms the advantage of DRO-DKMIL over an average top- k model as setting a proper k may be highly challenging. In addition, due to the discrete nature of k , the prediction performance may fluctuate significantly when k changes. The DRO based constraint essentially offers a soft version of the top- k constraint, which effectively addresses the limitation of an average top- k model.

3.1.6 Conclusion

We present a general GP mixture framework for multiple instance learning under noisy and multimodal settings. The proposed framework can flexibly incorporate multiple instances into the bag-level likelihood so that the model can most effectively learn from these potentially positive instances to make more robust predictions with the presence of outliers and different event types in the same bag. A key modeling ingredient is a DRO constraint applied to the mixture model parameters that acts as a soft top- k constraint to identify the subset of most positive instances in a bag. We further augment the GP kernel by using a deep neural network that uses adaptive basis functions to learn the rich representations from high dimensional input data. A stochastic variational inference method combines a fast sampling scheme to work on a mini-batch setting that ensures efficient and scalable end-to-end model training. Experiments on three challenging real-world video anomaly detection datasets clearly demonstrate the effectiveness of the proposed model.

3.2 Bayesian Nonparametric Submodular Video Partition for Robust Anomaly Detection

Anomaly detection from videos poses fundamental challenges as abnormal activities are usually rare, complicated, and unbounded in nature [83]. Furthermore, segment or frame labels are typically unavailable due to high labeling cost and therefore, the detection models have to rely on the weak video level labels [129]. There are two main streams of work to handle the challenging anomaly detection task. The first stream treats anomaly detection as an unsupervised learning problem [136]. It assumes that an event is considered to be abnormal if it deviates significantly from a predefined set of normal events included in the training data [20, 142, 151]. However, a model trained on limited normal data is likely to capture only specific characteristics present in the training dataset, and therefore, testing normal events deviating significantly from the training normal events will

lead to a high false alarm rate [159]. The second stream of research has attempted to address the limitation by formulating the problem as multiple instance learning (MIL) that models each video as a bag and its segments (or frames) as instances within the bag [129]. The goal is to learn a model that can effectively make frame-level anomaly predictions relying on the video-level labels during the training process. One effective MIL learning objective is to maximize the gap between two instances having the respective highest anomaly scores from a pair of positive and negative bags. The maximum score based MIL (referred as MMIL) model outperformed the unsupervised approaches and achieved promising performance in real-world long surveillance videos [129].

However, there are two key limitations with the MMIL model. First, the presence of noisy outliers (different from other normal events) in both abnormal and normal videos may significantly impact the overall model performance. This is because the objective function solely focuses on the individual segments from both positive and negative bags, making the training process sensitive to noises. Figure 1.2 (a-b) shows the example normal frames that are significantly different from other normal ones in real-world surveillance videos. The first frame is from the burglary video that looks similar to an abnormal frame from a video with an arson event. The second frame is from the shooting video that looks similar to a fighting frame. Hence, they may serve as outliers in the corresponding videos.

Second, if multiple types of abnormal events (referred to as multi-modal anomaly) present in a single abnormal video, MMIL may only detect one type of anomaly while missing other important ones due to the limitation of the objective function. Figure 1.2 (c)-(e) demonstrate three frames with different anomaly types from an example video in the Avenue dataset [87]. In Figure 1.2 (c), the person is running, which is regarded as a strange action in that context [87]. In (b), it shows a person waiting in a place holding some object in the hand, and (c) involves a person walking in the wrong direction. Therefore, the single video has multiple anomaly frames leading to a multimodal scenario.

top- k ranking loss has been adopted in attempt to address the issues as outlined above. It maximizes the gap between the mean score of the top- k predicted instances from a positive bag and that of a negative one [116, 137]. However, there are inherent limitations by using a top- k loss. First, it tends to be extremely sensitive to the selected k value. Figure 3.6 shows the highly fluctuating detection performance from two real-world surveillance video datasets. Since there is no frame (or segment) labels available during model training, setting an optimal k through cross-validation is infeasible or highly costly. Furthermore, given the diverse videos, the number of abnormal instances may vary significantly from one video to another implying we should have a different k for each



Figure 3.6: Highly fluctuating detection performance w.r.t. k

video. Hence, applying the same k to all videos as in the existing approaches fails to capture the nature of the data. Another serious but more subtle issue is that all (or most of) the selected k segments may come from the same sub-sequence of the video. Using a consecutive set of visually similar segments is less effective for model training, making it more likely to suffer from outlier and multimodal scenarios. As a result, top- k approaches will fall short in providing a reliable detection performance in most practical settings as evidenced by our experiments.

To address the fundamental limitations of existing solutions, we propose novel Bayesian non-parametric construction of a submodular set function, which is integrated with multiple instance learning to deliver robust video anomaly detection performance under practical settings. Instead of choosing a set of instances with the highest prediction scores that are likely from a consecutive sub-sequence, maximizing a specially designed submodular function can involve a more diverse set of instances and expose the model to all potentially abnormal segments for more effective model training. Furthermore, the submodular set function is constructed in a non-parametric way, which induces a pairwise similarity among different segments in a video based on the diverse nature of the data. More specifically, an infinite Hidden Markov Model with a Hierarchical Dirichlet prior (HDP-HMM) [135] augmented with an enhanced self-transition is employed to partition a video through dynamic non-parametric clustering of its segments. To more effectively accommodate the dynamic and noisy nature of real-world surveillance videos, the emission process of the HMM is also governed by a non-parametric mixture model to allow segments within the same hidden state to have visual and spatial variations. This unique design is instrumental to discover temporally consistent and semantically coherent hidden states that can be naturally interpreted as scenes. Pairwise similarity among different segments are defined according to the state-component structure, which leads to the construction of a submodular set function. We then develop a novel submodularity diversified MIL loss function to ensure robust anomaly detection from real-world surveillance videos with outlier and multimodal scenarios. Our key contributions include:

- Formulation of a novel **submodularity diversified MIL loss** that simultaneously extracts a diverse set of potentially positive instances while maximizing the gap between the mean score of

these instances from a positive bag and a negative one, respectively.

- **Bayesian non-parametric construction of the submodular set function** that infers the diversity from the video data to induce a pairwise similarity among different segments in a video and provide an upper bound on the size of the diverse set.
- **A greedy algorithm** that leverages the state-component hierarchical structure resulting from the non-parametric construction for submodular set function optimization and efficient model training.
- **Theoretical results** to ensure strong performance guarantee of the greedy algorithm.

The proposed approach achieves the state-of-the-art robust anomaly detection performance on real-world surveillance videos with noisy and multimodal scenarios.

3.2.1 Related Work

Encoding and sparse reconstruction-based approaches have been employed for anomaly detection, assuming that abnormal events are rare and deviate from normal patterns. They aim to capture the normal patterns using models, such as Gaussian processes (GPs) [76] and HMMs [66], to identify anomalies as outliers based on the reconstruction loss. Sparse representation-based approaches construct a dictionary for normal events and identify the events with the high reconstruction error as anomalies [87]. Recent approaches consider both abnormal and normal events in the training process. For video anomaly detection, since only video-level labels are assumed to be available during model training [48], MIL offers a natural solution by modeling each video as a bag and the associated segments (frames) as instances of the bag. Sultani et al. proposed an MIL based approach that enables to maximize the gap between highest prediction scores from a positive and negative bags, respectively [129]. However, this maximum score based MIL model (*i.e.*, MMIL) is insufficient to handle outlier and multimodal scenarios as discussed earlier.

top- k ranking loss based MIL models have been developed to address the limitations of the MMIL model [116,137]. These models produce state-of-the-art detection performance given that a suitable k value can be assigned in advance. However, as demonstrated earlier, the detection performance of such models is highly sensitive to the chosen k value. Meanwhile, given the diverse nature of videos, applying the same k value to all the videos is sub-optimal. More importantly, since instance level labels are not available during training time, choosing a single k value through cross-validation is infeasible or incurs a high annotation cost. Distributionally Robust Optimization (DRO) has been used to convert the top- k set into an uncertainty set that allows the model to focus on instances

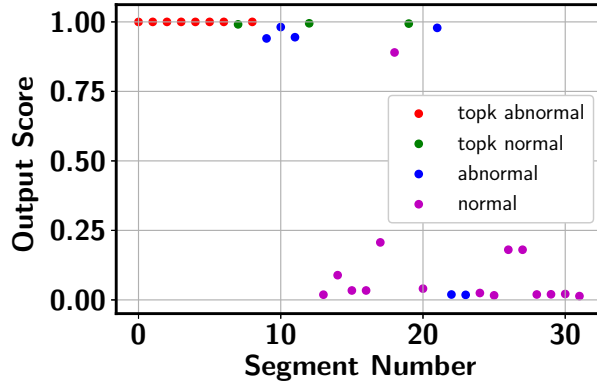


Figure 3.7: Output of the top- k based approach in a video from Avenue dataset (missing some of the abnormal segments in top- k).

in proportion to their prediction scores [116]. This is equivalent to assigning soft membership to involve instances into the MIL loss function. However, the size of the uncertainty set is controlled by the radius (*i.e.*, η) of the uncertainty ball, which needs to be manually set. Furthermore, the model may put more focus on a set of consecutive segments with the highest prediction scores and ignore some other potentially positive segments.

The proposed approach constructs a novel submodular set function in a non-parametric way by inferring the diversity from data automatically. By jointly optimizing the submodular function and the MIL loss, it automatically chooses a diverse set of segments and lets the model better differentiate these (potentially positive) segments from those of a negative bag to ensure good detection performance.

3.2.2 Methodology

Following the standard MIL assumption, we consider, for a positive bag, there is at least one abnormal segment whereas, for a negative bag all segments are of normal types.

Preliminaries

Let \mathbf{x}_i^+ be the i^{th} segment in the positive bag \mathcal{B}_{pos} and \mathbf{x}_j^- indicates the j^{th} segment in the negative bag \mathcal{B}_{neg} . Also consider n as the total number of instances per bag. The maximum score based MIL (MMIL) model tries to maximize the gap between the maximum prediction score from positive

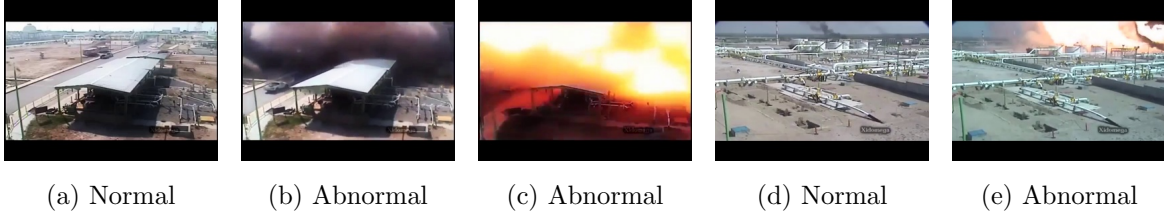


Figure 3.8: Example frames from different scenes in an explosion video from UCF-Crime: (a-b) scene 1, (c) scene 2, (d-e) scene 3

bag and that from the negative bag [129]:

$$L(\mathcal{B}_{pos}, \mathcal{B}_{neg}) = \left[1 - \max_{i \in \mathcal{B}_{pos}} (f(\mathbf{x}_i^+)) + \max_{j \in \mathcal{B}_{neg}} (f(\mathbf{x}_j^-)) \right]_+ \quad (3.19)$$

where $f(\mathbf{x}_i^+)$ (or $f(\mathbf{x}_j^-)$) is the prediction score of \mathbf{x}_i^+ (or \mathbf{x}_j^-) and $[a]_+ = \max\{0, a\}$. As mentioned earlier, MMIL is less effective to handle outlier and multimodal scenarios. The top- k ranking loss partially addresses the limitation of MMIL by maximizing the gap between an average of k highest segment predictions from the positive bag and maximum segment prediction score from a negative bag:

$$L(\mathcal{B}_{pos}, \mathcal{B}_{neg}) = \left[1 - \frac{1}{k} \sum_{i=1}^k f(\mathbf{x}_{[i]}^+) + \max_{j \in \mathcal{B}_{neg}} f(\mathbf{x}_j^-) \right]_+ \quad (3.20)$$

where the positive bag segment predictions are sorted in a non-decreasing order, *i.e.*, $f(\mathbf{x}_{[1]}^+) \geq \dots \geq f(\mathbf{x}_{[k]}^+)$. Table 3.5 demonstrates the important symbols used in this section.

There are two major issues associated with the top- k ranking loss. First, choosing an optimal k value is a key challenge as the number of abnormal instances may vary significantly from one video to another implying different k value for each video. Second, all the selected top- k instances may come from same sub-sequence of the video. Including all those visually similar instances does not contribute much in the model training process. Furthermore, concentrating only on a specific sub-sequence may make the approach less effective to handle multimodal and outlier scenario. Figure 3.7 presents the output of the top- k based model in the Avenue dataset. It can be seen that the top- k based approach picks the consecutive video segments while missing quite a few other abnormal frames.

Table 3.5: Symbols with Descriptions

Notation	Description
\mathcal{B}_{pos}	Positive bag (video)
\mathcal{B}_{neg}	Negative bag (video)
n	Number of segments in each bag
\mathbf{x}_i^+	Segment in a positive bag
$\mathbf{x}_{[i]}^+$	i^{th} largest prediction segment in a positive bag
\mathbf{x}_j^-	Segment in a negative bag
M	Feature dimension of each video segment
\mathbf{w}	Network parameters
\mathbf{b}	Network bias
k	Number of segments considered in the top- k formulation
η	Learning rate
\mathcal{C}^+	Set of instances from positive bag involve in model training
\mathcal{G}_0	Base distribution in DP
γ	Concentration parameter for the distribution \mathcal{G}_0
β_k	Weight associated with the k^{th} atom
ϕ_k	Atom k drawn from the distribution H
\mathcal{G}_j	Transition probability distribution of j^{th} state
$\hat{\pi}_{jl}$	Stick breaking weight associated with l^{th} atom in j^{th} group
α	Concentration parameter for $\hat{\pi}_j$
ϕ_{jl}	l^{th} atom corresponding j^{th} group
β_k	Stick breaking weight corresponding to atom ϕ_k
γ	Concentration parameter for β_k
ρ	Parameter defining the self transitioning
z_i	Scene assignment for the i^{th} segment in a video
s_i	Mixture component assignment for the i^{th} segment in a video
\mathcal{N}	Multivariate Gaussian distribution
$\boldsymbol{\mu}_{k,t}$	Mean of the k^{th} state, t^{th} mixture component
$\boldsymbol{\Sigma}_{k,t}$	Covariance of k^{th} state, t^{th} mixture component
$S_{i,j}$	Pairwise similarity between i^{th} and j^{th} segments
$F(\mathcal{C})$	Submodular set function
f_s^*	Maximum output score among segments assigned to the same cluster
i_s^*	Index of the representative segment
$\widehat{\mathcal{C}}^+$	Representative set constructed using the greedy algorithm
ϵ	Threshold to exclude segments with low prediction score from the representative set
κ	Upper bound of number of representative segments

Bayesian Non-parametric Submodular Set Function Construction

The proposed Bayesian non-parametric submodular video partition (BN-SVP) approach offers a novel integrated solution to address the above two fundamental challenges simultaneously. In particular, since submodular set functions provide a natural measure for diversity, we design a special submodular set function that enables discovery of a representative set of segments from a video. This avoids only choosing visually similar consecutive video segments like in the top- k approach, which enhances the model’s exposure to potential abnormal instances during model training. As a result, the model’s capability to handle multimodal and outlier scenarios can be effectively improved.

However, maximizing a submodular set function still requires to specify the size of the set. As mentioned above, choosing a set with an optimal size in video anomaly detection is highly challenging. To this end, we propose a novel Bayesian non-parametric construction of the submodular set function. The non-parametric construction leverages both visual features of the video segments and their temporal closeness to derive a similarity measure that allows us to define a submodular set function $F(\mathcal{C}^+)$, where \mathcal{C}^+ represents a subset of segments in a video. The size of \mathcal{C}^+ is automatically determined through Bayesian non-parametric analysis of the video. Intuitively, most videos, especially those with anomalies, usually consist of multiple scenes, where each scene is comprised of a consecutive set of visually similar segments. Figure 3.8 shows the example frames from three different scenes in a video that records an explosion event. Ideally, if a video could be partitioned based on these scenes, we can choose representative (and potentially positive) segments from each scene. Such information can significantly facilitate the optimization of the submodular set function. However, both the number and the types of the scenes are unavailable during model training.

The proposed BN-SVP addresses the above issue through non-parametric partition of a video. It builds upon and extends an HDP-HMM model that places a Hierarchical Dirichlet Process (HDP) prior on the state transition distribution of a Hidden Markov Model (HMM) model [135]. By using an HMM to model a video (as a sequence of segments), each discovered hidden state can be naturally interpreted as a scene in the video. The HDP prior allows us to determine the optimal number of states (*i.e.*, scenes) according to the nature of the data. However, real-world videos may be highly noisy and directly using an HDP-HMM model may extract too many scenes with less significant visual characteristics (*e.g.*, spatial changes of objects or addition/removal of a small number of objects). To address this issue, we follow the sticky HDP-HMM to encourage a stronger self-transition of a state [36]. This will result in temporal persistence of states to produce longer and semantically coherent scenes. To further accommodate spatial changes or variations

in certain objects, we allow the emission distribution to follow another non-parametric DP that automatically determines the number of mixture components (*i.e.*, sub-scenes) within the same scene. For example, scene 1 in Figure 3.8 is comprised of two sub-scenes: the first with a clear sky and the second with smoke in the sky.

More specifically, consider a collection of hidden states (*i.e.*, scenes in a video), the transition probability of state j to other states is governed by a DP:

$$\mathcal{G}_j = \sum_{l=1}^{\infty} \hat{\pi}_{jl} \delta_{\hat{\phi}_{jl}}, \hat{\pi}_j \sim \text{GEM}(\alpha) \quad (3.21)$$

where $\text{GEM}(\alpha)$ is formed through a stick breaking construction process with parameter α [135], $\hat{\phi}_{jl}$ is drawn from a base distribution \mathcal{G}_0 , which follows another DP

$$\mathcal{G}_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}, \beta \sim \text{GEM}(\gamma), \phi_k \sim H \quad (3.22)$$

Because of the discrete nature of \mathcal{G}_0 , there can be multiple $\hat{\phi}_{jl}$'s taking an identical value of ϕ_k . Considering the unique set of atoms ϕ_k , we can rewrite \mathcal{G}_j as

$$\mathcal{G}_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}, \pi_j \sim \text{DP}(\alpha, \beta), \phi_k \sim H \quad (3.23)$$

Given the highly dynamic and noisy nature of many real-world surveillance videos, directly applying the standard HDP-HMM model to partition a video may result in many redundant scenes and rapidly switches among them. This is problematic in our setting in which it is critical to infer semantically coherent scenes along with a slower transition among them. As a result, it is essential to ensure temporal persistence of the discovered scenes [36]. This can be achieved through enhanced self transitions. In particular, the transition probability of the j 's state is augmented by

$$\pi_j \sim \text{DP} \left(\alpha + \rho, \frac{\alpha \beta + \rho \delta_j}{\alpha + \rho} \right) \quad (3.24)$$

This has the effect of increasing the expected probability of staying in the same state.

$$\mathbb{E}[\pi_{jk} | \beta] = \begin{cases} \frac{\alpha \beta_j + \rho}{\alpha + \rho} & \text{if } k = j \\ \frac{\alpha \beta_k}{\alpha + \rho}, & \text{otherwise} \end{cases} \quad (3.25)$$

To allow certain levels of variations within the same scene and accommodate the highly dynamic nature of a video sequence, we propose to model the emission process using a mixture distribution

governed by another non-parametric DP. This design offers three unique advantages. First, it further ensures the temporal persistence of a scene as for a segment with less significant visual differences, it can stay in the same scene by switching to a different mixture component instead of transitioning to another (redundant) scene. Second, it offers a fine-grained partition of the video sequence, which is instrumental to separate abnormal segments (*e.g.*, frames (b) and (e) in Figure 3.8) from normal ones (*e.g.*, frames (a) and (d) in Figure 3.8) that share a common background. Last, the number of mixture components is automatically determined by the DP (*e.g.*, scenes 1 & 3 have two mixture components while scene 2 only has one).

For the k -th scene, there is a unique stick-breaking distribution $\psi_k \sim \text{GEM}(\tau)$ that defines the weights of the mixture components within the scene. Then, given the scene and mixture component assignment ($z_i = k, s_i = t$) of a segment \mathbf{x}_i^+ in a video, it is drawn from a specific multivariate Gaussian: $\mathcal{N}(\boldsymbol{\mu}_{k,t}, \Sigma_{k,t})$.

Posterior inference of the augmented HDP-HMM model with a DP mixture for emission can be achieved through direct assignment [135] or blocked sampling with an increasing mixing rate [32]. Hyper-parameters can also be inferred by placing a vague prior on them and conduct Gibbs sampling.

The scene and component assignments of BN-SVP induces a pairwise similarity among segments in a video:

$$\begin{cases} S_{i,j} = (\mathbf{x}_i^+)^{\top} \Sigma_{z_i, s_i}^{-1} \mathbf{x}_j^+ & \text{if } s_i == s_j \wedge z_i == z_j \\ S_{i,j} = 0 & \text{otherwise} \end{cases} \quad (3.26)$$

It is worth to note that the similarity between two segments \mathbf{x}_i^+ and \mathbf{x}_j^+ is evaluated using the learned feature representations (through a DNN) instead of the raw features. The induced similarity allows us to define a submodular set function summarized by the follow proposition.

Proposition 3.1. *Let κ denote the number of unique mixture components across all the discovered states in a bag \mathcal{B}_{pos} and $\mathcal{C} \subset \mathcal{B}_{pos}$ is a subset of \mathcal{B}_{pos} with size κ . Given the BN-SVP induced pairwise similarity defined in (3.26), the following function is a submodular set function:*

$$F(\mathcal{C}) = \sum_{i \in \mathcal{B}_{pos}} \max_{j \in \mathcal{C}} S_{i,j} \quad (3.27)$$

Based on the definition of $S_{i,j}$ as shown above, it is straightforward to show that $F(\mathcal{C})$ is a special instance of the location facility function [79], which is submodular. Since each mixture component captures a unique sub-scene, maximization of $F(\mathcal{C})$ can extract a diverse set of segments that best represent the all the scenes (and sub-scenes) in the entire video.

By further integrating the margin loss given in (3.20), we achieve a submodularity diversified MIL loss:

$$\min_{\mathbf{w}, \mathcal{C}^+ \in \mathcal{B}_{pos}, |\mathcal{C}^+| \leq \kappa} L(\mathcal{C}^+) - \lambda F(\mathcal{C}^+) \quad (3.28)$$

where the margin loss is defined over instances in a set \mathcal{C}^+ with size no larger than κ :

$$L(\mathcal{C}^+) = \left[1 - \frac{1}{|\mathcal{C}^+|} \sum_{i \in \mathcal{C}^+} f(\mathbf{x}_i^+) + \max_{j \in \mathcal{B}_{neg}} f(\mathbf{x}_j^-) \right]_+ \quad (3.29)$$

In essence, \mathcal{C}^+ includes the set of instances in a positive bag that participate in the model training. The constraint $|\mathcal{C}^+| \leq \kappa$ has the effect of excluding some representative segments from the margin loss as these segments are less likely to be abnormal (*e.g.*, with a very low predicted score). Including these segments will increase the margin loss and coefficient λ controls the balance between the margin loss and the diversity among the chosen segments.

Greedy Submodular Function Optimization

We propose a greedy algorithm for optimizing the submodular function in (3.27) to ensure efficient model training. The proposed algorithm leverages the special structure of the state and mixture component space resulted from the HDP-HMM partition of the video segments. The performance guarantee of the greedy algorithm is ensured by our theoretical result presented at the end of this section.

Recall that we use s_i to denote the mixture component of segment \mathbf{x}_i^+ . Let f_s^* denote the maximum score among all the segments assigned to the same mixture component and i_s^* be the index of the corresponding representative segment:

$$i_s^* = \arg \max_{\forall i: s_i = s} f(\mathbf{x}_i^+), \quad f_s^* = f(\mathbf{x}_{i_s^*}^+) \quad (3.30)$$

We construct a representative set $\widehat{\mathcal{C}}^+$ as follows. Let $\widehat{\mathcal{C}}^+ = \Phi$ and for each mixture component s , we set

$$\begin{cases} \widehat{\mathcal{C}}^+ \leftarrow \widehat{\mathcal{C}}^+ \cup \{i_s^*\}, & \text{if } f_s^* \geq \epsilon \\ \widehat{\mathcal{C}}^+ \leftarrow \widehat{\mathcal{C}}^+, & \text{otherwise} \end{cases} \quad (3.31)$$

where ϵ is a threshold to exclude segments with a low prediction score, which plays an equivalent role as constraint $|\mathcal{C}^+| \leq \kappa$ in (3.28). In our experiments, we use ϵ equal to the output of the segment staying in the 35th percentile among video specific segments so as to avoid skipping any

potential abnormal segments. Once a representative set $\widehat{\mathcal{C}}^+$ is constructed, model training can proceed by solving the following MIL loss:

$$\min_{\mathbf{w}} \left[1 - \frac{1}{|\widehat{\mathcal{C}}^+|} \sum_{i \in \widehat{\mathcal{C}}^+} f(\mathbf{x}_i^+) + \max_{j \in \mathcal{B}_{neg}} f(\mathbf{x}_j^-) \right]_+ \quad (3.32)$$

Given the state and mixture component assignment of each segment in a video, the representative set can be quickly constructed by sorting segments within each component according to their predicted scores and choosing the representative segment from each component by comparing its score with the threshold ϵ . Next, we provide a strong theoretical guarantee that the greedy algorithm can ensure the inclusion of a diverse set of segments for model training.

Theorem 3.3. *The representative set based MIL loss given in (3.32) is equivalent to the submodularity diversified MIL loss given in (3.28). Furthermore, using the proposed greedy algorithm to locate the κ representative segments essentially provides a κ -constrained greedy approximation to the maximization of the submodular set function $F(\mathcal{C})$. As a result, the obtained solution is guaranteed to be no worse $(1 - e^{-1})$ of the optimal solution.*

The detailed proof is provided in the Appendix.

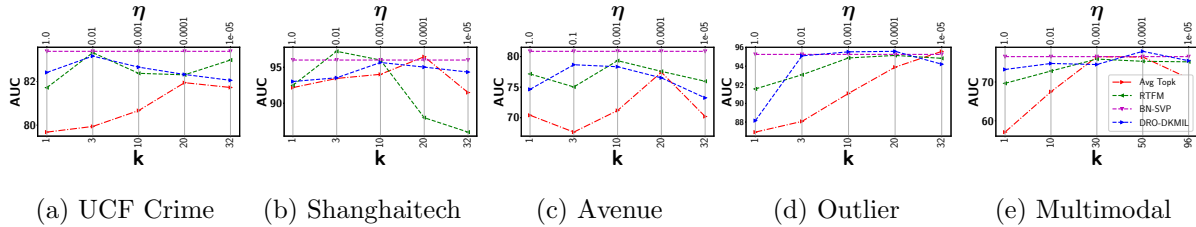
3.2.3 Experiments

We conduct extensive experiments to evaluate the effectiveness of the proposed BN-SVP approach. Through these experiments, we aim to demonstrate: (i) outstanding anomaly detection performance by comparing with competitive top- k , MIL, and other video anomaly detection models, (ii) robustness to outlier and multimodal scenarios, and (iii) deeper insights on the better detection performance through a qualitative study.

3.2.4 Datasets and Experimental Settings

Our experimentation includes three video datasets of different scales: ShanghaiTech [90], Avenue [87], and UCF-Crime [129]. Table 3.2 in the Appendix shows how the videos are partitioned into the training/testing sets in each dataset.

- **ShanghaiTech** consists of 437 videos with 330 normal and 107 abnormal videos. In the original setting, all training videos are normal. To fit into our setting, we follow the data split in [159] to assign normal and abnormal videos in both training and testing sets.

Figure 3.9: Performance comparison with top- k ranking models

- **Avenue** consists of 16 training and 21 testing videos. We perform 80:20 split separately in the abnormal and normal video sets to generate training and testing instances.
- **UCF-Crime** consists of 13 different anomalies with a total of 1900 videos, where 1610 are training videos and 290 are testing videos. In this dataset, frame labels are available only for the testing videos.

To show the robustness of the proposed approach in the multimodal and outlier scenarios, we also generate the Multimodal and Outlier datasets. Specifically, we create a multimodal scenario by extending the UCF-Crime dataset. For the outlier scenario, we deliberately impose some outliers in the ShanghaiTech dataset. More details of these two datasets are provided in Section 3.2.6. For evaluation, we report the frame-level receiver operating characteristics (ROC) curve along with the corresponding area under curve (AUC). The AUC score indicates the robustness of the performance at various thresholds.

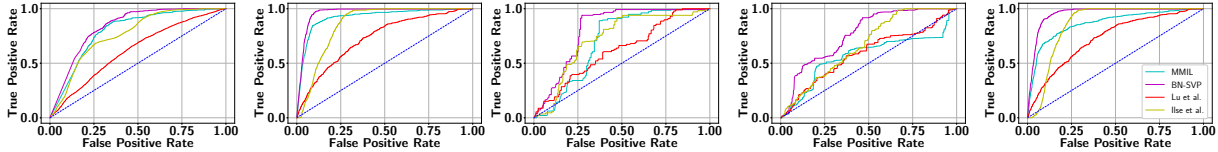
For Avenue and ShanghaiTech datasets, we extract visual features from the FC7 layer of a pre-trained C3D network [138]. We re-size each video frame to 240×340 pixels and fix the frame rate to 30 fps. We compute the C3D features for every 16-frame video clip. This may yield a different number of clips (each clip having a 2048 dimensional feature vector) depending on the number of frames in each video. Thus, we fit any number of clips to the 32 segments by taking an average of clip features in a specific segment. In case of UCF-Crime, we extract the features using an I3D network [13] by using the pretrained network as described in [144]. For all datasets, we use parallel GCN networks to capture the feature similarity and temporal consistency. The outputs of the parallel branches are combined and passed through a 5-layer LSTM network where each layer has 32 hidden units followed by batch normalization. Finally, an FC layer with sigmoid activation is applied to bring the prediction score to $(0, 1)$. For model training, we use SGD with a learning rate of 0.001 and l_2 regularization with parameter $\lambda = 0.001$. Detailed information about the network architecture is provided in the Appendix.

3.2.5 Performance Comparison

Comparison with top- k Models We first compare the detection performance with two most recent top- k based models, including Robust Temporal Feature Magnitude learning (RTFM) [137] and the DRO based deep kernel MIL (DRO-DKMIL) [116]. We also compare with a standard average top- k model (Avg Topk) as the baseline. Avg Topk uses the rank loss in (3.20) with the same network architecture as BN-SVP. For RTFM, we get the result by re-running the original implementation for different k values. Similarly, for DRO-DKMIL, we run the original implementation for different η values that control the size of the uncertainty set. The proposed BN-SVP removes the dependency on these highly sensitive parameters through non-parametric modeling. Detailed comparison results are shown in Figure 3.9.

We have several important observations. First, all the top- k models are very sensitive to the selection of the k value (or η that defines a soft version of the top- k set). Both RTFM and DRO-DKMIL outperform the standard Avg Topk. DRO-DKMIL achieves relatively more stable performance across all datasets. This may attribute to its conversion of discrete optimization (*i.e.*, choose a specific k) to a continuous optimization problem (*i.e.*, choosing η). However, for certain dataset (*e.g.*, Avenue), its performance still varies more than 8%. Second, while for some rare cases that RTFM or DRO-DKMIL achieves the best performance for a specific k or η , they under-perform BN-SVP in most cases. This is mainly due to that these models tend to choose a consecutive set of segments, which limits the model’s exposure of other potentially positive segments. This issue has been effectively addressed by BN-SVP, which extracts a diverse set of potentially positive segments through submodular optimization.

Comparison with Other Models We also make comparison with other existing techniques that do not depend on the k value. Specifically, our comparison study includes the maximum score based MIL model (MMIL) by Sultani et al. [129], attention based deep MIL model proposed by Ilse et al. [52], a dictionary based approach proposed by Lu et al. [87], and an MIL model for soft bags (MILS) proposed by Li & Vasconcelos [78] as common baselines for all datasets. Sultani et al. [129] used the loss function in (3.19) along with the temporal similarity and consistency as a regularizer. Ilse et al. used a permutation invariant aggregation function to detect the positive instances in the bag, where the function operators are learned using the attention network [52]. Li & Vasconcelos used a large-margin based latent support vector machine model with the goal to correctly classify positive and negative bags [78]. In case of the approach presented by Zhong et al. [159], we directly report the performance from the original paper for the UCF-Crime and ShanghaiTech datasets. This approach involves multiple rounds of alternative optimization between



(a) UCF-Crime (b) ShanghaiTech (c) Avenue (d) Multimodal (e) Outlier

Figure 3.10: ROC curves on three video datasets (a)-(c), multimodal (d) and outlier (e)

classification and cleaning and may produce unstable performance [137]. Considering its difficulty in the training and replication process, we do not include it in other datasets.

Table 3.6 reports the AUC scores of BN-SVP along with the results from the comparison models as described above. It can be seen that BN-SVP clearly outperforms other models in all datasets and a large margin (*i.e.*, 6-8%) is achieved on both ShanghaiTech and Avenue datasets. The corresponding ROC curves are shown in Figure 3.10, which demonstrates a consistent trend. For example, on UCF-Crime, BN-SVP has a more than 10% better True Positive Rate (TPR) compared to MMIL at a False Positive Rate (FPR) of 0.2. Also at varying FPR, BN-SVP consistently outperforms the other competitive baselines, which justifies its outstanding detection capability.

3.2.6 Detecting Multimodal and Outlier Segments

Multimodal Detection The original UCF-Crime dataset does not explicitly consider a multimodal scenario. Even though the real-world surveillance videos may indeed contain those cases (which is evidenced by the superior performance of the BN-SVP model), it is hard to identify actual videos for this specific information. In UCF-Crime dataset, different types of anomalies are present. This allows us to explicitly create multimodal scenarios by combining multiple abnormal videos from different activity types. To this end, we randomly select three activity types and form an abnormal bag by concatenating three abnormal videos, one video per activity type. The training bags are constructed using the training dataset whereas testing bags are constructed using the testing dataset. In total, we construct 50 abnormal and 50 normal training bags. In the testing set, there are 10 normal and 10 abnormal videos. Table 3.7 shows the AUC scores and corresponding ROC curve is shown in the Figure 3.10 (d). BN-SVP achieves a more superior performance compared to other baselines. Furthermore, BN-SVP stays consistently on the top in the ROC curve justifying the effectiveness of the approach toward the multimodal scenario. As an example, at $FPR = 0.1$, BN-SVP is at least 20% better than other approaches on TPR.

Table 3.6: Comparison with Other Models

Approach	AUC (%)
UCF-CRIME	
Hasan et al. [46] (C3D)	50.60
Lu et al. [87] (C3D)	65.51
Lu et al. [87] (I3D)	61.98
MMIL [129] (C3D)	75.41
Li & Vasconcelos [78] (I3D)	77.95
Ilse et al. [52] (I3D)	76.52
Zhong et al. [159] (GCN (C3D))	81.08
Zhong et al. [159] (TSN ^{RGB})	82.12
Zhong et al. [159] (TSN ^{OpticalFlow})	78.08
MMIL [129] (I3D)	79.68
BN-SVP (I3D)	83.39
SHANGHAI TECH	
Lu et al. [87] (C3D)	72.90
Li & Vasconcelos [78] (C3D)	90.40
Zhong et al. [159] (GCN (C3D))	76.44
Zhong et al. [159] (GCN(TSN ^{RGB}))	84.44
Zhong et al. [159] (GCN(TSN ^{Optical Flow}))	84.13
Ilse et al. [52] (C3D)	85.78
MMIL [129] (C3D)	92.18
BN-SVP (C3D)	96.00
AVENUE	
Binary SVM (C3D)	69.11
Lu et al. [87] (C3D)	62.14
Li & Vasconcelos [78] (C3D)	72.23
Ilse et al. [52] (C3D)	72.39
MMIL [129]	70.40
BN-SVP (C3D)	80.87

Outlier Detection To assess the robustness on outlier detection, we extend the ShanghaiTech dataset with outliers. Specifically, we randomly select 120 segments from abnormal videos and replace their features with points drawn from a standard multivariate Gaussian distribution. As shown in Table 3.7, MMIL suffers heavily by the outliers compared to the proposed BN-SVP. This is because, it is likely to have an outlier prediction as the maximum prediction score from the abnormal video. As a result, the overall optimization process may be heavily influenced by outliers.

Table 3.7: AUC Scores on Multimodal and Outlier Detection

Approach	AUC (%)	
	<i>Multimodal</i>	<i>Outlier</i>
Lu et al. [87] (C3D)	58.67	72.90
Li & Vasconcelos [78] (C3D)	70.96	90.95
Ilse et al. [52] (C3D)	66.85	85.65
MMIL [129]	57.08	86.47
BN-SVP	76.53	95.27



(a) Frame 1



(b) Frame 2

Figure 3.11: Frames from UCF-Crime Stealing019; (a) Correct BN-SVP, Avg Topk, (b) Correct BN-SVP, Incorrect Avg Topk

3.2.7 Qualitative Analysis

To show the effectiveness of extracting a diverse set of segments for model training, we present illustrative sample frames in a stealing video from UCF-Crime, where BN-SVP correctly identifies all abnormal frames and a top- k approach (*e.g.*, Avg Topk) misses some of them. In Figure 3.11, both frames are of abnormal types and but they occur in two distinct time intervals within the video. The first frame is more obvious for a stealing event. Consequently, both the proposed BN-SVP and Avg Topk are able to correctly identify it. In contrast, the second frame is less obvious for a stealing activity. Nevertheless, it is still chosen by BN-SVP due to its diverse coverage of potentially abnormal frames during the training process. On the other hand, Avg Topk only focuses on frames with high prediction scores that are usually co-located in the same time interval. This will narrow the scope of the model being exposed to other abnormal frames. Therefore, Avg Topk is not able to correctly predict the second frame and falsely classify it as normal. More qualitative analysis that demonstrates the robustness of the proposed approach on multimodal and outlier scenarios is provided in the Appendix along with an ablation study for the prediction score threshold ϵ defined in (3.31).

3.2.8 Conclusion

In summary, we propose a novel Bayesian non-parametric submodularity diversified MIL model for robust video anomaly detection in practical settings that involve outlier and multimodal scenarios. By integrating submodular optimization with the minimization of an MIL loss, the proposed approach identifies a diverse set of segments to ensure comprehensive coverage of all potential positive segments for effective model training. The Bayesian non-parametric construction of the submodular set function automatically determines the upper bound on the size of the diverse set, which serves as a key constraint for minimizing the submodularity diversified MIL loss function. The resulting state-component structure also leads to a greedy submodular optimization algorithm to support efficient model training. Effectiveness of the proposed approach is demonstrated through the state-of-the-art robust anomaly detection performance on real-world surveillance videos with noisy and multimodal scenarios.

Chapter 4

Multiple Instance Active Learning for Anomaly Detection

As a widely used weakly supervised learning scheme, modern MIL models achieve competitive performance at the bag level. However, instance-level anomaly detection prediction, which is essential for many important applications, remains largely unsatisfactory.

To achieve a high bag level prediction, most existing MIL models primarily focus on the most positive instance from a positive bag that is mainly responsible for determining the bag label [4, 47, 63, 129]. However, they suffer from two major limitations, which lead to poor instance-level predictions. First, solely focusing on the most positive instance is sensitive to outliers, which are negative instances that look very different from other negative ones [12]. As a result, these instances may be wrongly assigned a high score indicating they are positive. Second, there may be multiple types (*i.e.*, multimodal) of positive instances in a single bag (*e.g.*, different types of anomalies in a surveillance video or different types of skin lesions in a dermatology image). Thus, focusing on a single most positive instance will miss other positive ones. Both cases will result in a low instance-level prediction performance. A possible solution to improve the detection of positive instances is to consider the top- k most positive instances. However, the number of positive instances may vary significantly across different bags and applying the same k to all bags may be inappropriate. Furthermore, finding an optimal k for each bag is highly challenging as it takes a discrete value.

The underlying reason for the less accurate instance-level prediction is due to the lack of instance labels. For positive instances that are relatively rare across bags, detecting them by only relying

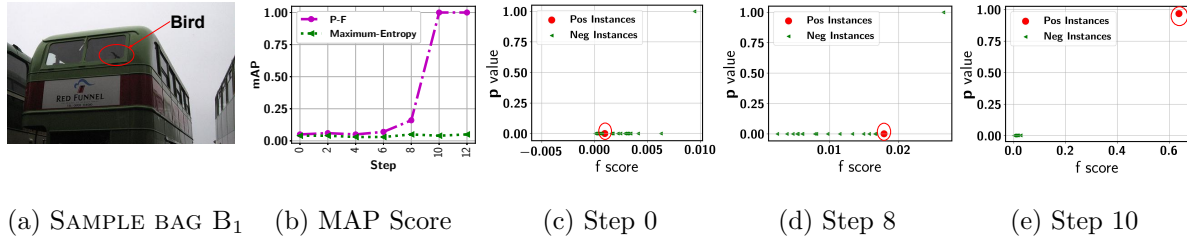


Figure 4.1: (a) Example of a challenging bag; (b) MI-AL performance on instance-level predictions; (c)-(e) Prediction scores of instances in the bag in different MI-AL steps

on bag labels is inherently challenging as the weakly supervised signal (*i.e.*, bag label) cannot be propagated to the instance level without sufficient statistical evidence. One promising direction to tackle this challenge is to augment MIL with active learning (AL). Multiple instance AL (or MI-AL) aims to select a small number of informative instances to improve the instance level prediction in MIL. In most MIL problems, the data is highly imbalanced at the instance level, where the positive ones are much more sparse. Since the positive instances usually carry more important information, a primary goal of MI-AL is to effectively sample the positive instances from a candidate pool dominated by the negative ones. If a true positive instance can be sampled and labeled, it can help to identify other similar positive instances in the same and different bags, which will significantly improve the instance-level predictions.

However, existing MIL models may easily miss some rare positive instances [129]. They may also focus on the wrongly identified negative instances due to their sensitivity to outliers or incapability of handling multimodal bags. Thus, the true positive instances may be assigned a low prediction score, indicating that they are predicted as negative with a high confidence. As a result, commonly used uncertainty based sampling will miss these important instances. Figure 4.1 (a) shows a challenging bag, which is an image that contains the shadow of a bird (as the positive class). The positive instances are patches that cover (part of) the bird shadow. Figure 4.1 (b) shows that combining uncertainty sampling with a maximum score based MIL model (the green curve) is not able to sample effectively so that instance-level prediction remains very low over the AL process. Figure 4.1 (c) further confirms that the initial prediction score (F-score) of the positive instance is close to 0, making it hard to be sampled.

We propose a novel MI-AL model for effective instance sampling to significantly boost the instance-level prediction in MIL. We design an unique variance regularized MIL loss that encourages a high variance of the prediction scores to address bags with a highly imbalanced instance distribution and/or those with outliers and multimodal scenarios. Since the variance regularizer is non-convex, we propose to optimize a distributionally robust bag likelihood (DRBL), which provides a good

convex approximation of the variance based loss with a strong theoretical guarantee. The DRBL automatically adjusts the impact of the bag-level variance, making it more effective to detect potentially positive instances to support active sampling. It can also be naturally integrated with a deep architecture to support deep MIL model training using mini-batches of positive-negative bag pairs. Finally, a novel P-F sampling function is developed that combines a probability vector (*i.e.*, \mathbf{p}) and predicted instance scores (*i.e.*, \mathbf{f}), obtained by optimizing the DRBL. By leveraging the key MIL assumption, the sampling function can explore the most challenging bags and effectively detect their positive instances for annotation, which significantly improves the instance-level prediction. Novel batch-mode sampling is developed to work seamlessly with the deep MIL, leading to a powerful active deep MIL (ADMIL) model to support sampling of high-dimensional data used in most MIL applications. Figure 4.1 (b) shows the proposed model (purple curve) that significantly improves instance predictions. Figures 4.1 (c)-(e) show P-F sampling dynamically updates the probability \mathbf{p} and score \mathbf{f} values to effectively sample the positive instance from the highly challenging bag in a few steps.

Our main contribution includes: (i) an unique variance regularized MIL loss and its convex surrogate that address inherent MIL challenges to best support active sampling, (ii) a novel P-F sampling function to effectively explore most challenging bags with rare positive instances, (iii) mini-batch training and batch-mode active sampling to support ADMIL in broader MIL applications, and (iv) state-of-the-art instance prediction performance in MIL while maintaining low instance annotations.

4.1 Related Work

The related work in MIL is described in Section 2. In this section, we will be describing about the Active Learning (AL).

Active Learning (AL). Uncertainty and margin based measures are commonly leveraged in existing AL models to achieve efficient data sampling [124]. Distributionally robust optimization has also been adopted in multi-class AL to address sampling bias and imbalanced data distribution [165]. Deep learning (DL) models are good candidates for AL because of their high-dimensional data processing and automatic feature extraction capability. Existing models mainly target at improving uncertainty quantification of the network for reliable sampling [39,60,74,143]. Batch-mode sampling is commonly used in active DL to avoid frequent model re-training. It focuses on constructing representative batches to avoid redundant information given by similar instances [5,64,120]. AL in the MIL setting has been rarely investigated. One exception is the MI logistic model and its three

uncertainty measures to simultaneously consider both instance and bag level uncertainty [125]. However, uncertainty sampling is ineffective to explore challenging bags, where all instances are confidently predicted as negative. In addition, the original model is a simple linear model, which does not provide sufficient capacity for high-dimensional data. There is no systematic way to support batch-mode sampling, either. A reinforcement learning based AL technique is developed in [14], where segments are chosen to be labeled in each AL step. However, segmentation level annotations are required to compute the reward during the training process, which violates the assumption of MIL. Another AL framework is developed for MIL tasks in [156]. However, sampling is conducted at the bag level (*i.e.*, choosing bags instead of instances). Thus, it is essentially a multi-label AL model, aiming to improve the bag-level predictions with fewer annotated bags. This is fundamentally different from the design goal of ADMIL.

4.2 Methodology

Let $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ denote a set of instances associated with each bag \mathcal{B} , where each $\mathbf{x}_i \in \mathbb{R}^D$ is a feature vector. Let $t_{\mathcal{B}} \in \{+1, -1\}$ indicates the bag type. Following the standard MIL assumption discussed earlier, active sampling will focus on instances in the positive bags as all instances in a negative bag are negative. We also allow the number of instances to vary from one bag to another.

4.2.1 Variance Regularization

Let \mathbf{x}_i^+ (or \mathbf{x}_j^-) be the i^{th} (or j^{th}) instance in a positive bag \mathcal{B}_{pos} (or a negative bag \mathcal{B}_{neg}). Following the MIL assumption, a commonly used loss function for training deep MIL models is to make the maximum prediction score of instances from a positive bag to be higher than a negative bag [129]. We define as

$$\mathcal{L}^{\text{MS}} = \left\{ 1 - \max_{i \in \mathcal{B}_{pos}} [f(\mathbf{x}_i^+; \mathbf{w})] + \max_{j \in \mathcal{B}_{neg}} [f(\mathbf{x}_j^-; \mathbf{w})] \right\}_+ \quad (4.1)$$

where $f(\mathbf{x}; \mathbf{w}) \in [0, 1]$ is the prediction score of instance \mathbf{x} provided by a deep neural network parameterized by \mathbf{w} and $[a]_+ = \max\{0, a\}$. We will omit \mathbf{w} from $f(\mathbf{x}; \mathbf{w})$ to keep the notation uncluttered. The above objective function aims to maximize the gap between the maximum prediction score of instances from a positive bag and maximum score from a negative bag. Model training can be performed by sampling pairs of positive and negative bags $(\mathcal{B}_{pos}, \mathcal{B}_{neg})$, using their bag-level labels to evaluate the loss, and performing back-propagation. The maximum score based MIL (referred to as MS-MIL) models are designed primarily for bag label prediction as it aims to

identify a single most positive instance from a positive bag and maximizes its prediction score. In this way, it fully leverages the MIL assumption (*i.e.*, at least one positive instance in \mathcal{B}_{pos}) and the weakly supervised signal (*i.e.*, bag-level label).

As discussed earlier, MS-MIL and its top- k extensions suffer from key limitations that impact their instance-level prediction performance. Meanwhile, they provide inadequate support to sample the most informative instances to enhance the instance predictions. Inspired by the recent advances in learning theory to automatically balance bias and variance in risk minimization [30], we propose a novel variance regularized MIL loss function to capture the inherent characteristics of MIL, aiming to collectively address highly imbalanced instance distribution, existence of outliers, and multimodal scenarios. As a result, minimizing the new MIL loss can effectively improve the prediction scores of the positive instances, making them easier to be sampled for annotation by the proposed sampling function. In particular, the variance regularized loss introduces *two novel changes* to (4.1), which are formalized below:

$$\mathcal{L}^{\text{VAR}} = \left\{ 1 - \left[\frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i^+) + C \sqrt{\frac{\text{Var}_n[f(X^+)]}{n}} \right] + \max_{j \in \mathcal{B}_{neg}} [f(\mathbf{x}_j^-)] \right\}_+ \quad (4.2)$$

where $\forall i \in [1, n], \mathbf{x}_i^+ \in \mathcal{B}_{pos}$, n is the size of \mathcal{B}_{pos} , Var_n is the empirical variance of $f(X^+)$ with X^+ being a random variable representing an instance from a positive bag, and parameter C balances the mean score and the variance.

The first key change is to use the mean score to replace the maximum score in (4.1), which avoids the model to only focus on the most positive instance in a bag to make it robust to outliers and multimodal scenarios. Since positive bags are guaranteed to include positive instances and instances in a negative bag are all negative, it is desirable that the mean score for a positive bag should be high. Maximizing the mean score in a positive bag using a complex model (*e.g.*, a DNN) could effectively reduce the training loss (by reducing the bias) in estimating the bag-level labels. However, using the mean score alone is problematic as most instances in a positive bag are usually negative in a typical MIL setting. As a result, such a low bias model will lead to a very high false positive rate, which negatively impacts the overall instance-level prediction. The proposed loss function addresses this issue through the novel variance term, which effectively handles the highly imbalanced instance distribution. With only a small number of instances being truly positive, the empirical variance Var_n for the bag should be high due to the large deviation of a small number of high scores from the majority of low scores. It is worth to note that the variance term in (4.2) plays a distinct role than risk minimization in standard supervised learning, where it is minimized to control the estimation error. In contrast, the variance in (4.2) is encouraged to be large to allow a small set of instances in a bag to be positive, aiming to precisely capture the imbalanced

distribution. To our best knowledge, this is the first bias-variance formulation in the MIL setting.

Conducting MI-AL using variance regularization still faces two challenges. First, its effectiveness hinges on an optimal balance between the mean score and the empirical variance, which is controlled by the hyperparameter C . Similar to the standard supervised learning, there lacks a systematic way of setting such a hyperparameter to achieve an optimal trade-off. Second, the variance term is non-convex with multiple local minima [30], which makes model training much more difficult and time-consuming. Thus, it is not suitable for real-time interactions to support active sampling.

4.2.2 Distributionally Robust Bag Likelihood

To address the challenges as outlined above, we propose to formulate a distributionally robust bag level likelihood (DRBL) as a convex surrogate of the variance regularized loss in (4.2). By extending the distributionally robust optimization framework developed for risk minimization in supervised learning [30, 99], we theoretically prove the equivalence between DRBL and variance regularization with high probability. Being convex, DRBL is easier to optimize that facilitates MIL model training to support fast active sampling. Furthermore, by setting a proper uncertainty set as introduced next, we show that the parameter C is directly obtained when optimizing the DRBL, where the instance distribution in the bag is constrained by the uncertainty set. As a result, it achieves automatic trade-off between the mean prediction score and the variance.

We first introduce a probability vector $\mathbf{p} = (p_1, \dots, p_n)^\top$, where $\sum_i p_i = 1, p_i \geq 0, \forall i \in \{1, \dots, n\}$ and let p_i denote the probability that instance $\mathbf{x}_i^+ \in \mathcal{B}_{pos}$ can represent the bag. We further introduce a binary indicator vector $\mathbf{z} = (z_1, \dots, z_n)^\top$, where $p(z_i = 1) = p_i$. Let Y be a binary random variable that denotes the bag label. Conditioning on all the instances in the bag, the (conditional) bag likelihood for bag \mathcal{B}_{pos} is given by $p(Y = 1|\mathbf{z}, \mathbf{f}) = \prod_i f(\mathbf{x}_i^+)^{z_i}$, where $\mathbf{f} = (f(\mathbf{x}_1^+), \dots, f(\mathbf{x}_n^+))^\top$. By integrating out the indicator variables, we have the marginal bag likelihood as $p(Y = 1|\mathbf{p}, \mathbf{f}) = \sum_i p_i f(\mathbf{x}_i^+)$. Instead of letting a single most positive instance to determine the bag label, where $p(y = 1|\mathbf{p}, \mathbf{f}) = f(\mathbf{x}_k^+)$ with $k = \arg \max_i f(\mathbf{x}_i^+)$, which is equivalent to MS-MIL, or assigning equal probability to each instance (*i.e.*, $p_i = 1/n$), which is equivalent to the mean score, we introduce an uncertainty set \mathcal{P}_n that allows \mathbf{p} to deviate from a uniform distribution to some extent:

$$\mathcal{P}_n := \left\{ \mathbf{p} \in \mathbb{R}^n, \mathbf{p}^\top \mathbb{1} = 1, 0 \leq \mathbf{p}, D_f \left(\mathbf{p} \parallel \frac{\mathbb{1}}{n} \right) \leq \frac{\lambda}{n} \right\} \quad (4.3)$$

where $D_f(\mathbf{p}|\mathbf{q})$ is the f -divergence between two distributions \mathbf{p} and \mathbf{q} , $\mathbb{1}$ is a n -dimensional unit vector, and λ controls the extent that \mathbf{p} can deviate from a uniform vector, which essentially corresponds to the imbalanced instance distribution in the bag. Note that \mathcal{P}_n only specifies a

neighborhood that \mathbf{p} may deviate from a uniform distribution. Since \mathcal{P}_n is a convex set, an optimal \mathbf{p} can be easily computed for each specific bag by optimizing the robust bag likelihood according to its specific imbalanced instance distribution. This is fundamentally more advantageous than a top- k approach, where k is discrete and hard to optimize. Next, we show that the optimal robust bag likelihood is equivalent to the variance regularized mean prediction score with high probability, which allows us to define a new MIL loss based on DRBL.

Theorem 4.1. *Let X^+ be a random variable representing an instance from a positive bag, $f(X^+) \in [0, 1]$ is the score assigned to an instance, $\sigma^2 = \text{Var}[f(X^+)]$ and $\text{Var}_n[f(X^+)]$ denote the population and sample variance of $f(X^+)$, respectively, and D_f takes the form of χ^2 -divergence. For a fixed λ and with $n \geq \max(2, \frac{\lambda}{\sigma^2} \max(8\sigma, 44))$,*

$$\max_{\mathbf{p} \in \mathcal{P}_n} \sum_{i=1}^n p_i f(\mathbf{x}_i^+) = \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i^+) + \sqrt{\frac{\lambda \text{Var}_n[f(X^+)]}{n}} \quad (4.4)$$

with probability at least $1 - \exp\left(-\frac{7n\sigma^2}{20}\right)$, where \mathcal{P}_n is an uncertainty set defined by (4.3).

It is worth to note that given the highly imbalanced positive instances in a typical MIL setting, the true variance σ^2 should be high. For a bag with a decent size, it guarantees the equivalence in (4.4) with high probability. Furthermore, maximizing the robust bag likelihood given on the l.h.s. of (4.4) assigns $C = \sqrt{\lambda}$, which automatically adjusts the impact of variance based on the uncertainty set. Theorem 4.2 below further generalizes this result to the KL-divergence.

Theorem 4.2. *Let X^+ be a random variable representing an instance from a positive bag, $f(X^+) \in [0, 1]$ is the score assigned to an instance, $\sigma^2 = \text{Var}[f(X^+)]$ and $\text{Var}_n[f(X^+)]$ denote the population and sample variance of $f(X^+)$, respectively, and D_f takes the form of KL-divergence. We have*

$$\max_{\mathbf{p} \in \mathcal{P}_n} \sum_{i=1}^n p_i f(\mathbf{x}_i^+) = \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i^+) + \sqrt{\frac{2\lambda \text{Var}_n[f(X^+)]}{n}} + \epsilon\left(\frac{\lambda}{n}\right) \quad (4.5)$$

where $\epsilon\left(\frac{\lambda}{n}\right) = \frac{\lambda}{3n} \frac{\kappa_3(f(X^+))}{\text{Var}_n[f(X^+)]} + \mathcal{O}\left(\left(\frac{\lambda}{n}\right)^{3/2}\right)$ with $\kappa_3 = \mathbb{E}_0[(f(X^+) - \mathbb{E}_0[f(X^+)])^3]$ and \mathbb{E}_0 denotes the expectation taken over \mathbf{p}_0 .

Remark: Given a bag with a decent size $n \gg 1$ and since λ is usually set to $\lambda \ll 1$ (0.01 is used in our experiments), we have $\epsilon\left(\frac{\lambda}{n}\right) \rightarrow 0$. When the empirical variance $\text{Var}_n[f(X^+)]$ is sufficiently large (which is true for MIL), the r.h.s. of (4.5) is dominated by the first two terms, which implies

$$\max_{\mathbf{p} \in \mathcal{P}_n} \sum_{i=1}^n p_i f(\mathbf{x}_i^+) \approx \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i^+) + \sqrt{\frac{2\lambda \text{Var}_n[f(X^+)]}{n}} \quad (4.6)$$

Detailed proofs are given in Appendix. Leveraging the above theoretical results, we formulate a DRBL-based MIL loss as

$$\mathcal{L}^{\text{DRBL}} = \left\{ 1 - \max_{\mathbf{p} \in \mathcal{P}_n} \left[\sum_{i=1}^n p_i f(\mathbf{x}_i^+) \right] + \max_{j \in \mathcal{B}_{neg}} \left[f(\mathbf{x}_j^-) \right] \right\}_+ \quad (4.7)$$

The DRBL loss offers a very intuitive interpretation on the newly introduced probability vector \mathbf{p} . Since it can deviate from the uniform distribution as specified by the uncertainty set \mathcal{P}_n , each entry p_i essentially corresponds to the contribution (or weight) of \mathbf{x}_i^+ to the bag likelihood (being positive). As a result, to maximize the robust bag likelihood, instances with a higher prediction score should receive a higher weight. Meanwhile, constrained by \mathcal{P}_n , multiple instances will contribute to the bag likelihood with a sizable weight as \mathbf{p} cannot deviate too much from being uniform. Hence, their prediction scores will simultaneously be brought up by the model. This makes DRBL robust to the outlier and multimodal cases as it increases the chance for the true positive instances or multiple types of true positive instances to be assigned a high prediction score. This provides fundamental support to the proposed P-F active sampling function that combines the probability vector \mathbf{p} and the prediction score \mathbf{f} in a novel way to choose the most informative instances in a bag for annotation.

4.2.3 P-F Active Sampling

Since we have the prediction score $f(\mathbf{x}_i^+) \in [0, 1]$, it can be naturally interpreted as the probability of instance \mathbf{x}_i^+ being positive. A straightforward way to perform uncertainty based instance sampling is to compute the f -score based entropy of the instances, referred to F-Entropy:

$$\mathbf{x}_* = \arg \max_{i \in \mathcal{B}_{pos}} H[f(\mathbf{x}_i^+)], \quad (4.8)$$

where $H[f] = -[f \log f + (1 - f) \log(1 - f)]$. Since the sampled instance has the largest prediction uncertainty (according to F-Entropy), labeling such an instance can effectively improve the model's instance-level performance. Active sampling using (4.8) is straightforward, which involves evaluating $H[f(\mathbf{x}^+)]$ for all the instances from positive training bags (note that all the instances in a negative bag are negative). Since we consider a deep learning model to better accommodate high-dimensional data, sampling one instance at a time requires frequent model training, which is computationally expensive. Instead, we sample a small batch of instances in each step based on their predicted F-Entropy. It is worth to note that, due to the highly imbalanced instance distribution, the majority of the prediction scores, including many positive instances, may be very low. The goal is to assign a relatively higher score to the potentially positive instances so that

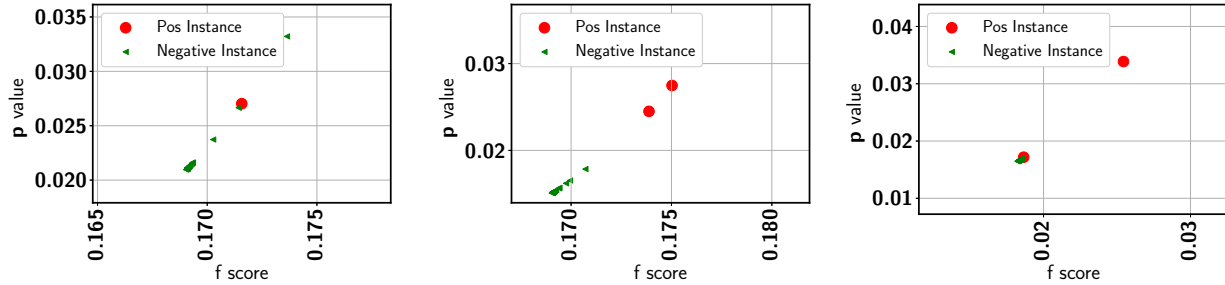


Figure 4.2: Example of challenging bags from different topics in 20NewsGroup

their entropy is not too low, indicating a confident negative prediction, which will be missed by the sampling function.

As discussed earlier, using the robust bag likelihood as the MIL loss can directly benefit instance sampling by increasing the chance to assign a higher prediction score to a positive instance so that it is more likely to be sampled. However, F-Entropy sampling still suffers from two major limitations. First, for some very difficult bags, such as the sample image shown in Figure 4.1 (a), identifying the positive instances (*e.g.*, the patch in the image containing the shadow of a bird) can be highly challenging. As a result, they may be assigned a very low f score. In fact, as shown in Figure 4.1 (c), all the instances in this bag receive a very low score with the highest less than 0.01, leading to a very low entropy. Some additional examples of challenging bags from the 20NewsGroup dataset are shown in Figure 4.2, where all the instances are predicted with a very low score. Hence, all these instances are predicted as negative with low uncertainty, making them less likely to be chosen by entropy based sampling. Second, since batch-mode sampling is adopted to reduce the training cost of a deep network, it is essential to diversify the selected instances in the same batch to minimize the annotation cost. However, choosing data instances solely based on their predicted entropy may lead to the annotation of similar instances, which is not cost-effective.

The proposed P-F active sampling overcomes the above two limitations simultaneously through effective bag exploration by combining the probability vector \mathbf{p} and the prediction score \mathbf{f} through a minmax function according to their distinct roles in a bag. The key design rationale of P-F sampling is rooted in the standard MIL assumption that ensures at least one positive instance in each positive bag to guide effective bag exploration. Both \mathbf{p} 's and \mathbf{f} 's along with the bag structure are dynamically updated during bag exploration to increase the chance of sampling the positive instances in an under-explored bag. A hybrid loss function further utilizes labels of sampled negative instances in the same bag to boost the prediction scores of the positive instances. More specifically, let B be the total number of positive training bags, P-F sampling will choose the following data

instance:

$$\mathbf{x}_*^{PF} = \arg \min_{b \in \{1, \dots, B\}} f(\mathbf{x}_{b_*}^+), \quad \text{and } b_* = \arg \max \mathbf{p}_b \quad (4.9)$$

where \mathbf{p}_b is the probability vector of bag b . For each bag, the sampling function first identifies the instance $\mathbf{x}_{b_*}^+$ with the largest p value in each bag. Such an instance can be regarded as the most representative instance in the bag as it makes the largest contribution (according to \mathbf{p}_b) to the bag likelihood. According to the prediction score of $\mathbf{x}_{b_*}^+$, we can categorize bags into three groups: (1) easy bags, where $f(\mathbf{x}_{b_*}^+)$ takes a large value, indicating that the model makes confidently correct predictions, (2) confusing bags, where $f(\mathbf{x}_{b_*}^+)$ is reasonably large but uncertain, indicating the model is still confusing about its prediction, and (3) difficult bags, where $f(\mathbf{x}_{b_*}^+)$ is very low, indicating the model makes confidently wrong predictions. It is desirable to sample from both confusing and difficult bags as the model already makes accurate instance predictions for easy bags. Sampling instances from the confusing bags can be achieved through the proposed F-Entropy as the model makes uncertain predictions, which leads to a high entropy. Finally, sampling from the difficult bags is fundamentally more challenging due to low prediction scores for the entire bag. However, the MIL assumption provides a general direction for bag-level exploration of positive instances as there must be at least one positive instance in each positive bag. The P-F sampling function in (4.9) chooses the representative instance from the bag with the lowest prediction score. Such an instance is guaranteed to be sampled from an under-explored (*i.e.*, difficult) bag as it has the lowest prediction score despite being predicted as the most positive instance in the bag.

Extension to the batch-mode sampling is conducted in two directions, within a bag and across bags, for more effective exploration while ensuring diversity of the sampled instances. First, instead of only sampling the most positive instance from the identified under-explored bag, we propose to sample $k > 1$ instances as the positive instances may be ranked lower than multiple negative instances in the bag according to the current prediction scores (see Figure 4.1 (c) for an example). This helps to more effectively explore very difficult bags. To ensure diversity among the sampled instances, we keep k small but sample across multiple bags simultaneously. Only bags with a max prediction score $f(\mathbf{x}_{b_*}^+)$ less than a threshold (0.3 is used in our experiments) will be explored as these represent the difficult bags as discussed above. For bags with a larger $f(\mathbf{x}_{b_*}^+)$, they are either easy bags or confusing bags that can be effectively sampled using F-Entropy. Our overall P-F sampling function integrates bag exploration and F-Entropy and gives priority to the former to perform diversity-aware bag exploration first. As more bags are successfully explored along with MI-AL, less instances will be sampled by exploration and the focus will be naturally shifted to F-Entropy to perform model fine-tuning. The detailed sampling process is summarized by Algorithm 1.

Similar to AL in standard supervised learning, the sampled annotated instances should be used

to improve the model prediction performance. However, the MIL loss primarily focuses on the bag-level labels due to the lack of instance labels. To this end, we propose a hybrid loss function that integrates the bag and instance labels. Let $\mathbf{X}^l = \{\mathbf{x}_1^l, \mathbf{x}_2^l, \dots, \mathbf{x}_m^l\}$ be the m labeled instances queried by the proposed active learning function and $\mathbf{t}^l = \{t_1^l, t_2^l, \dots, t_m^l\}$ with $t_i^l \in \{0, 1\}$ be the corresponding instance labels. We formulate a supervised binary cross-entropy (BCE) loss as

$$L^{\text{BCE}} = -\frac{1}{m} \sum_{i=1}^m \left[t_i^l \log(f(\mathbf{x}_i^l)) + (1 - t_i^l) \log(1 - f(\mathbf{x}_i^l)) \right] \quad (4.10)$$

It is clear that the sampled positive instances provide important supervised signals so that the model will predict a high score for similar positive instances, which will directly benefit instance-level prediction. In contrast, the sampled negative instances, especially those chosen from the under-explored bags, contribute less to improve the prediction performance as their original prediction scores are already low. However, they play a subtle but essential role to achieve more effective bag-level exploration. First, if a sampled instance is labeled as negative, it will be removed from the bag, which does not violate the MIL assumption. Meanwhile, since we have $\sum_i p_i = 1$, the p values will be redistributed and the chance for each remaining instance to be sampled is therefore increased. Furthermore, the BCE loss will further bring down the prediction scores of negative instances that are similar to the sampled one. This may help to improve the score of the positive instance so that it can have a higher chance to be sampled in the future. Finally, the hybrid loss that combines the MIL loss and the supervised loss is used to retrain the model after a new batch of instances are queried:

$$\mathcal{L}^{\text{Hybrid}} = \mathcal{L}^{\text{DRBL}}(\mathcal{B}_{\text{pos}}, \mathcal{B}_{\text{neg}}) + \beta \mathcal{L}^{\text{BCE}}(\mathbf{X}^l, \mathbf{t}^l) \quad (4.11)$$

where β is used to trade-off bag- and instance-level losses.

4.3 Experiments

We conduct extensive experimentation over multiple real-world MIL datasets to justify the effectiveness of the proposed ADMIL model. The purpose of our experiments is to demonstrate: (i) the state-of-the-art instance prediction performance by comparing with existing competitive baselines, (ii) effectiveness of the proposed P-F active sampling function through comparison with other sampling mechanisms, (iii) impact of key model parameters through a detailed ablation study, and (iv) qualitative evaluation through concrete examples to provide deeper and intuitive insights on the working rationale of the proposed model.

Algorithm 1: P-F Active Sampling**Input:** $\mathbf{p}_{B_{pos}}, \mathcal{Q}_{prev}, Th_{PF}, Th_H, BSize, k$ **Output:** \mathcal{Q} **Data:** B positive training bags // Feature vector for each bag

```

1 Initialization:  $\mathcal{U}_B = \{\}, \text{count} = 0, \mathcal{Q}_{P-F} = \{\}, \mathcal{Q}_F = \{\}$ 
2 for  $b \in [B]$  do
3    $\mathbf{p}_b \leftarrow \mathbf{p}_{B_{pos}}[b], b_* \leftarrow \arg \max \mathbf{p}_b \setminus \mathcal{Q}_{prev}[b]$ 
4   if  $f(\mathbf{x}_{b_*}^+) \leq Th_{PF}$  then
5      $\mathcal{U}_B \leftarrow b_*$ 
   /* Adding instances from unexplored bags */
6  $\mathcal{U}_B = \arg \text{sortAsc}_{b_* \in \mathcal{U}_B} f(\mathbf{x}_{b_*}^+)$ 
7 for  $b_* \in \mathcal{U}_B$  do
8   if  $b_* \in \mathcal{Q}_{prev}$  then
9     if  $\text{positive ins} \in \mathcal{Q}_{prev}[b]$  then
10       $\text{continue}$ 
11   else
12      $\mathcal{X}^{PF} = \arg \text{sortDesc}_{b_*} (f(\mathbf{x}_{b_*}^+) \setminus \mathcal{Q}_{prev}[b_*])[:k]$ 
13     for  $\mathbf{x}_i \in \mathcal{X}^{PF}$  do
14       if  $\text{count} \geq BSize$  then
15          $\text{break}$ 
16        $\mathcal{Q}_{P-F}[b_*] \leftarrow \mathbf{x}_i$ 
17        $\text{count} \leftarrow \text{count} + 1$ 
18  $\mathcal{Q}_{prev} = \mathcal{Q}_{prev} \cup \mathcal{Q}_{P-F}$ 
   /* Adding instances with highest F-Entropy;
       $H[f(\mathbf{x}_i^+)] = - [f(\mathbf{x}_i^+) \log f(\mathbf{x}_i^+) + (1 - f(\mathbf{x}_i^+)) \log(1 - f(\mathbf{x}_i^+))]$  */
19  $\mathcal{C}_{idx} = \arg \text{sortDesc}_i (H[f(\mathbf{x}_i^+)] \geq Th_H)$ 
20 for  $i \in \mathcal{C}_{idx}$  do
21   if  $\text{count} \geq BSize$  then
22      $\text{break}$ 
23   if  $\mathbf{x}_i^+ \in \mathcal{Q}_{prev}[b_i]$  then
24      $\text{break}$ 
25    $\mathcal{Q}_F[b_i] \leftarrow \mathbf{x}_i^+$ 
26    $\text{count} \leftarrow \text{count} + 1$ 
27  $\mathcal{Q} = \mathcal{Q}_{prev} \cup \mathcal{Q}_F$ 

```


Table 4.1: Number of positive and negative bags on different datasets

Split	20NewsGroup		Cifar10		Cifar100		Pascal VOC	
	<i>Positive</i>	<i>Negative</i>	<i>Positive</i>	<i>Negative</i>	<i>Positive</i>	<i>Negative</i>	<i>Positive</i>	<i>Negative</i>
Train	30	30	500	500	500	500	124	124
Test	20	20	100	100	100	100	84	84

4.3.1 Experimental Setup

Datasets. Our experiments involve four datasets covering both textual and image data: 20New-Group [162], Cifar10 [67], Cifar100 [67], and Pascal VOC [31]. The detailed description of each dataset is given below and bag level statistics is summarized in the Table 4.1

- **20NewsGroup:** In this dataset, an instance refers to a post from a particular topic. For each topic, a bag is considered as positive if it contains at least one instance from that topic and negative otherwise. This dataset is particularly challenging because of the severe imbalance where there are very few ($\approx 3\%$) positive instances in each positive bag. While number of instances per bag may vary, on average there are around 40 instances per bag.
- **Cifar10:** In the original dataset, there are 50,000 training and 10,000 testing images with 10 classes indicating different images. The bags are constructed as follows. First, we pick ‘automobile’, ‘bird’, and ‘dog’ related images as positive instances and the rest as negative. To construct a positive bag, we choose a random number from 1 to 3 and pick the positive instances equal to the randomly generated number. The rest of the instances are selected from a negative instances pool. For negative bags, all instances are selected from the negative instance pool. For each bag, we consider 32 instances.
- **Cifar100:** The dataset consists of 50,000 training and 10,000 testing images with 20 different superclasses indicating different species. Bag construction is similar to Cifar10, where images in superclass flowers are treated as positive and the rest as negative.
- **Pascal VOC:** This dataset consists of 2,913 images, where images are used for segmentation. Each image is treated as a bag and instances are obtained as follows. We define a grid size of 60×75 and partition the images. Depending on the image size, the number of instances may vary. We treat an instance as positive if at least 5% of the total pixels in a given instance are related to the object of interest otherwise negative. In our case, we consider the bird as the object of interest. All the images consisting of bird are regarded as positive bags and others as negative.

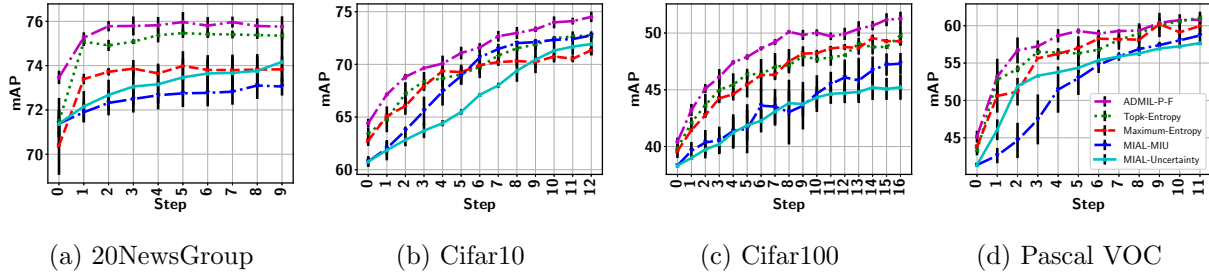


Figure 4.3: MI-AL performance

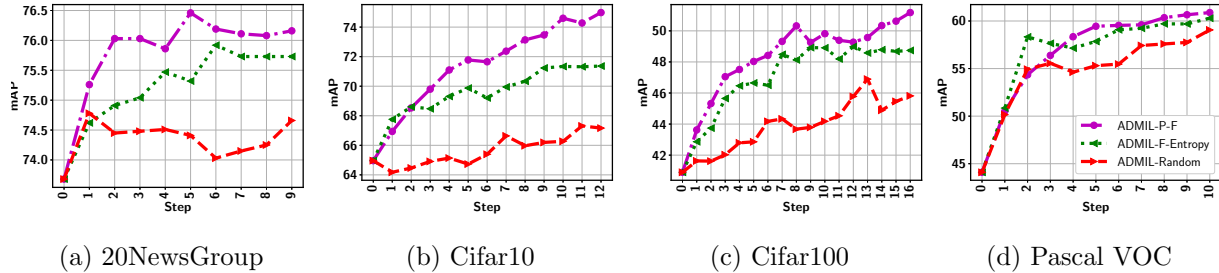


Figure 4.4: Effectiveness of P-F active sampling

Evaluation metric and model training. To assess the model performance, we report the instance-level mean average precision (mAP) score, which summarizes a precision-recall curve as a weighted mean of precision achieved at each threshold, with the increase in recall from the previous threshold as the weight. mAP explicitly places much stronger emphasis on the correctness of the few top ranked instances than other metrics (*e.g.*, AUC) [128]. This makes it particularly suitable for instance prediction evaluation as a small subset of instances with the highest prediction scores will eventually be identified as positive for further inspection (by human experts) with the rest being ignored. For Cifar10, Cifar100, and Pascal VOC datasets, we extract the visual features from the second-to-the last layer of a VGG16 network pre-trained using the imagenet dataset, yielding a 4,096 dimensional feature vector for each instance. For 20NewsGroup, we use the available 200-dimensional feature vector. In terms of network architecture, we use a 3-layer FC neural network. The first layer has 32 units followed by 16 units and 1 unit FC layers. We adopt 60% dropout between FC layers. ReLU and sigmoid activations are used for the first and last FC layers. Learning rate 0.01 is used for all dataset except for 20NewsGroup which is 0.1.

4.3.2 Performance Comparison

To demonstrate the instance prediction performance achieved by the proposed ADMIL model, we compare it with competitive baselines. First, the two MI-AL sampling strategies: MIAL-Uncertainty and MIAL-MIU [125], from the MI logistic model are included. Since our datasets

Table 4.2: MIL Performance in Passive Setting

Approach	20NewsGroup	Cifar10	Cifar100	Pascal VOC
Ilse et al. [52]	60.85	65.16	40.15	40.15
Hsu et al. [50]	42.08	63.84	41.57	34.83
ADMIL	73.47(75.42)	64.41(74.50)	40.41(51.26)	45.15(60.79)

involve high-dimensional data, we replace the original linear model by the exact DNN model used in our ADMIL so we can focus on comparing MI active sampling. The EGL sampling technique in [125] was not included due to the prohibitive computational cost to evaluate the gradient of each instance output with respect to the large number of DNN parameters. We also implement an MS-MIL model and its top- k variant with uncertainty sampling using entropy. Given the different sizes of the datasets, we query maximum 15 instances per step in 20NewsGroup, 30 instances in Pascal VOC, and 150 instances in Cifar10 and Cifar100. Figure 4.3 shows the MI-AL curves with one standard deviation (computed over three runs) represented by vertical black line for all four datasets. ADMIL achieves the best performance in all cases. For most datasets, it shows a much better initial performance, which results from the proposed DRBL-based MIL loss that significantly benefits MIL performance in passive learning. Overall the entire MI-AL process, ADMIL consistently stays the best and converges to a higher point in the end for all datasets. For the Pascal VOC, the top- k MIL model with entropy sampling achieves closer performance towards the end, which is mainly due to the limited positive instances in this dataset. Hence, no testing bags contain similar positive instances in the challenging bags that are explored by P-F sampling. While ADMIL achieves much better instance predictions in those bags, the advantage does not transfer to the testing bags. For reference, we also compare ADMIL with two recently developed MIL models, including Ilse et al. [52] and Hsu et al. [50], under the passive setting. As shown in Table 4.2, ADMIL achieves better or at least comparable performance as compared with these competitive baselines. This clearly justifies of using ADMIL as a base model for active sampling. After labeling a small set of actively sampled instances, the performance is significantly boosted (as shown in the parenthesis), which further justifies the benefits of combining AL with MIL. Our qualitative study will provide a more detailed analysis on this.

Effectiveness of active sampling. To demonstrate the effectiveness of the proposed P-F active sampling function, we compare it with two other sampling methods, F-Entropy and random sam-

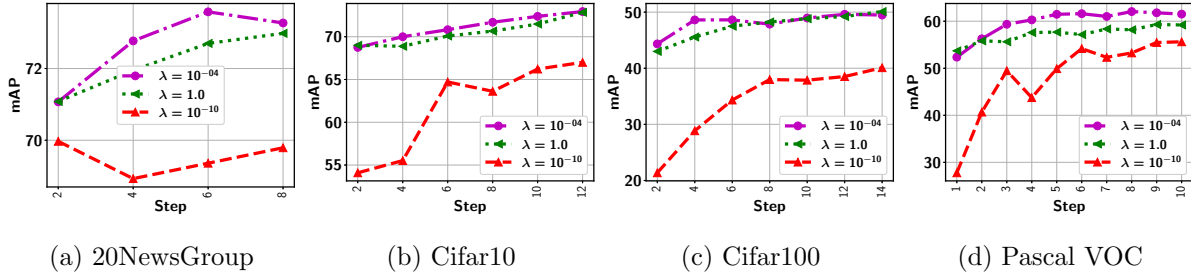


Figure 4.5: Impact of model parameter λ

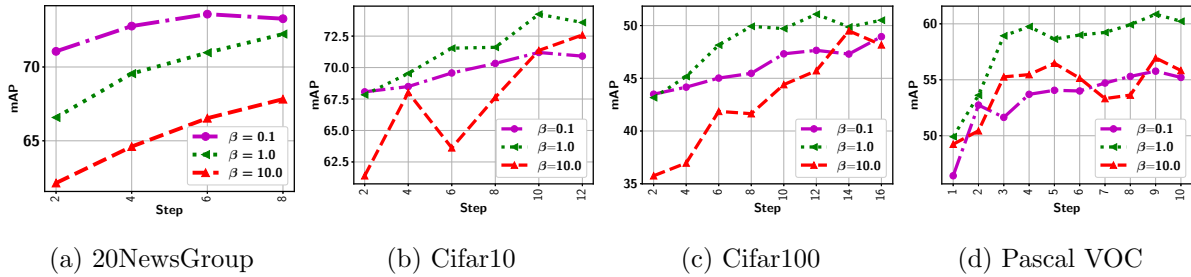


Figure 4.6: Impact of model parameter β

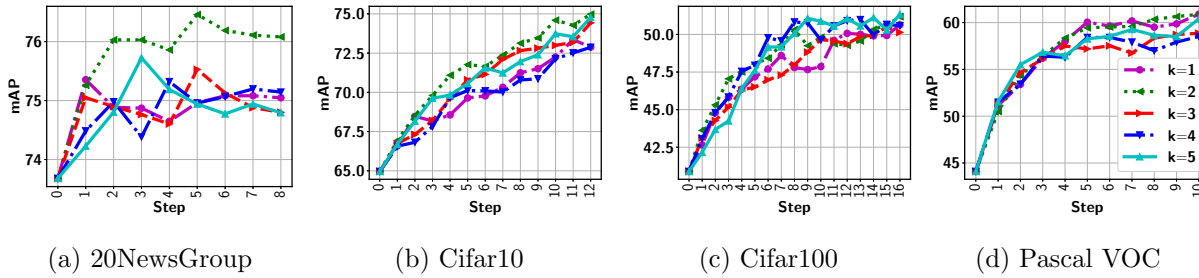


Figure 4.7: Impact of hyperparameter k

pling, while keeping all other parts of the model the same. As shown in Figure 4.4, P-F sampling clearly outperforms others with a large margin in the first three datasets. Its advantage over F-Entropy is smaller on Pascal VOC due to the same reason as explained above. The performance gain is mainly attributed to the effective exploration of P-F sampling over the most challenging bags.

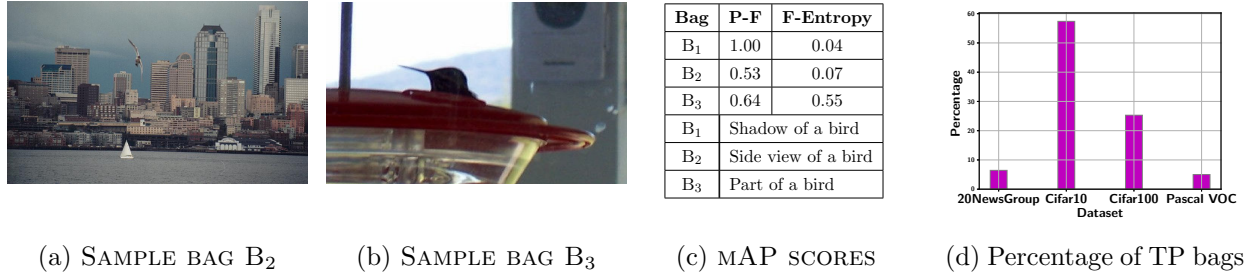


Figure 4.8: (a-b) Poorly explored bags in Pascal VOC; (c) Description of these bags and their mAP scores; (d) Additional true positive bags successfully explored by P-F sampling

4.3.3 Ablation Study

Impact of λ and β : Figures 4.5 and 4.6 demonstrate the impact of λ (with $\beta = 1$) and β (with $\lambda = 0.01$) to the model performance. In particular, λ can be set according to the imbalanced instance distribution within bags, where a larger λ corresponds to a higher imbalanced distribution. We vary λ in $[10^{-10}, 1]$ and since most bags in the MIL setting are highly imbalanced, relatively higher λ value gives very good performance in general. Figure 4.5 shows that $\lambda = 0.0001$ clearly outperforms too large (or small) λ values. As for β , placing less emphasis on an instance level loss (small β), we may not fully leverage labels of queried instances. Meanwhile, with too much emphasis on the instance level loss (large β), the model overly focuses on the limited queried instances with less attention to the bag labels. Therefore, a good balance results in an optimal performance, shown in Figures 4.6.

Impact of k : Figure 4.7 shows the impact of the hyperparameter k , which is the number of instances queried in each unexplored bag, on model performance. As can be seen, $k = 2$ achieves a generally decent performance across all the datasets. For datasets with a larger size (*e.g.*, Cifar100), a larger k leads to a slightly better performance.

4.3.4 Qualitative analysis

To further justify why the proposed ADMIL model and its P-F sampling function work better than other baselines, we provide a few illustrative examples to offer deeper insights on its good performance. First, we show two challenging bags in addition to the one shown in Figure 4.1 (a). As shown in Figure 4.8 (a-b), B₂ presents a side view of a bird while only a small portion of the bird is visible in B₃. For those difficult cases, the model originally predicts all instances as a negative with high confidence. However, by coupling the P-F sampling and the hybrid loss in (4.11), the

positive instances from those bags are successfully queried. Figure 4.8 (c) shows a clear advantage in the mAP scores between P-F sampling and F-Entropy. As a further evidence, we investigate the number of true positive (TP) bags being explored by both P-F sampling and F-Entropy. TP bags refer to those that the model is being able to query at least one true positive instance. Instead of reporting the actual number of bags, which is affected by the size of the dataset, we show the additional percentage TP bags being explored by P-F sampling in Figure 4.8 (d). It is worth to note that neither method tries to query the easy bags as their positive instances are correctly predicted with high confidence. The major difference is from the challenging bags and the percentage of these bags varies among different datasets. Nevertheless, P-F sampling consistently explores more effectively than F-Entropy across all datasets.

4.4 Conclusion

To tackle the low instance-level prediction performance of existing MIL models that is essential for many critical applications, we develop a novel MI-AL model to sample a small number of most informative instances, especially those from confusing and challenging bags, to enhance the instance-level prediction while keeping a low annotation cost. We propose to optimize a robust bag likelihood as a convex surrogate of a variance regularized MIL loss to identify a subset of potentially positive instances. Active sampling is conducted by properly balancing between exploring the challenging bags (through P-F sampling) and refining the model by sampling the most confusing instances (through F-Entropy). The design of the loss function naturally supports mini-batch training, which coupled with the batch-mode sampling, makes the MI-AL model work seamlessly with a deep neural network to support broader MIL applications that involve high-dimensional data. Our extensive experiments conducted on multiple MIL datasets show clear advantage over existing baselines.

Chapter 5

Anomaly Detection under Class Imbalanced Setting

Anomaly detection/open-set detection (OSD) under the class imbalanced problem poses a fundamental challenge as the model may be equally uncertain between minority class samples and openset (anomalous) samples. Despite the promising progress in OSD that focuses on differentiating samples from the close (normal classes) and open sets, respectively, limited attention has been devoted to the situation where the close set involves highly imbalanced classes, which may be quite common in many practical settings. For example, for anomaly detection, the known types of anomalies available for model training are usually unevenly distributed into multiple categories (*e.g.*, car accident vs. shooting). Similarly, for computer-aided medical diagnosis, the known diseases (to the model) may be highly imbalanced based on the available cases. Thus, following the standard Empirical Risk Minimization (ERM) framework for training, the model may not learn properly from the minority class due to the lack of positive samples. As a result, it is more likely to misidentify a minority-class sample as an unknown-class sample during OSD, leading to a high false-positive rate.

Distributionally Robust Optimization (DRO) offers an effective means to handle the imbalance class distribution in the closed set setting [108, 164]. In DRO, the worst case weighted loss is optimized, where the weights are searched in a given neighborhood (referred to as the uncertainty set) of the empirical sample distribution such that the overall loss is maximized. By expanding the uncertainty set, the model is encouraged to assign higher weights to difficult samples. As a result, samples from the minority class will be given more emphasis during model training if not properly learned

(which incurs a larger loss). Another common solution to handle imbalanced class distribution in the close set is through oversampling to achieve a more balanced class distribution [18]. While both oversampling and DRO may help to improve the close set performance, neither of them is adequate to address OSD from imbalanced data.

A fundamental challenge lies in the interplay between samples from the minority class and the difficult samples from the majority classes. As a result, simply oversampling the minority class may neglect these difficult samples. Similarly, applying DRO with a flexible uncertainty set may put too much emphasis on these difficult samples and ignore the minority class as well as some representative samples from the majority classes, which affects proper model training. In fact, directly applying these models for OSD may lead to even worse detection performance, which is evidenced by our experimental results. Recent approaches also try to leverage the visual similarity across the centroids of closed-set classes to allow more effective training from the minority class samples [85]. However, it is possible that the samples from the minority class may look quite different from most other samples, making such a strategy less effective.

To systematically tackle the fundamental challenge as outlined above, we propose Distributionally Robust Evidential Optimization (DREO) that offers a principled way to quantify sample uncertainty through evidential learning while optimally balancing the model training over all classes in the close set through adaptive DRO learning. To avoid the model from primarily focusing on the most difficult samples by following the standard DRO, the adaptive learning strategy gradually increases the size of the uncertainty set, which allows the model to learn from easy to hard samples. A class-ratio biased loss is further assigned to the minority class to ensure proper learning from its limited samples. Our main contribution is fourfold:

- a novel extension of DRO to evidential learning, which enables principled uncertainty quantification under the class imbalanced setting, critical for many applications, including OSD,
- adaptive DRO training governed by a uniquely designed multi-scheduler learning mechanism to ensure an optimal model training behavior that gives sufficient attention to the difficult samples and the minority class while capable of learning common patterns from the majority classes,
- theoretical connection to a boosting model (*i.e.*, AdaBoost), which ensures the nice convergence and generalization properties of DREO,
- state-of-the-art OSD performance on various datasets.

5.1 Related Work

Most related topics for our work Open set detection and Distributionally Robust Optimization (DRO) which are described in Related Work 2.

5.2 Methodology

Let $\mathcal{D}_N = \{\mathbf{X}, \mathbf{Y}\} = \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_N, \mathbf{y}_N)\}$ be a set of training samples in the close set. Each $\mathbf{x}_n \in \mathbb{R}^D$ is a D -dimensional feature vector and $\mathbf{y}_n \in \{0, 1\}^C$ indicates the one hot encoding associated with its class label: $y_{nj} = 1$ and $y_{nk} = 0$ for all $k \neq j$ with j being the true label.

Evidential Learning for OSD

Following the principle of Subjective Logic (SL) [58], we consider a total of $C + 1$ mass values with C being the number of classes. We assign a belief mass $b_c, \forall c \in [C]$, to each singleton, which corresponds to one class in the close set and the remaining mass is referred to as the uncertainty mass, denoted by u . Table 5.1 summarizes all the major symbols along with their descriptions.

The belief masses and the uncertainty mass are all non-negative and sum to one: $u + \sum_{c=1}^C b_c = 1, u \geq 0$ and $b_c \geq 0$. They can be evaluated as

$$b_c = \frac{e_c}{S}, \quad u = \frac{C}{S} \quad (5.1)$$

where $S = \sum_{c=1}^C (e_c + 1)$ with $e_c \geq 0$ being the evidence derived for the c^{th} singleton, which can be generated by a neural network enabled with a non-negative output. The belief mass assignment in the above expression corresponds to a Dirichlet distribution with the concentration parameters $\alpha_c = e_c + 1$:

$$\text{Dir}(\mathbf{p}|\boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{c=1}^C p_c^{\alpha_c - 1}, & \text{for } \mathbf{p} \in \mathcal{S}_C \\ 0, & \text{otherwise} \end{cases} \quad (5.2)$$

where \mathcal{S}_C is a $(C - 1)$ -simplex and $B(\boldsymbol{\alpha})$ is a beta function.

Given the evidences, the expected probability for the c^{th} singleton is given by

$$\mathbb{E}[p_c] = \frac{\alpha_c}{S} \quad (5.3)$$

Consider a sample \mathbf{x}_n and let $\mathbf{f}(\mathbf{x}_n, \Theta)$ denote the evidence vector generated by an evidential neural network parameterized by Θ . This allows us to fully characterize the Dirichlet distribution, whose

Table 5.1: Symbols with Descriptions

Notation	Description
b_c	Belief mass associated with class c
C	Total number of classes
e_c	Evidence for the c^{th} singleton
u	Uncertainty mass
α	Dirichlet Parameters
p_c	Probability for the c^{th} singleton
\mathbf{y}_n	One hot encoded C dimensional multinomial variable
y_{nc}	Class label for the n^{th} data sample for class c
p_{nc}	Probability of the n^{th} data sample belonging to class c
η_t	Uncertainty set size for DREO
β	Hyperparameter controlling the schedule
γ	Hyperparameter controlling the emphasis in a minority class
\mathbf{p}	Probability distribution in the DRO framework
\mathcal{P}^{DRO}	Uncertainty set
Θ	Evidential network parameters
$l_n^{EL}(\Theta)$	Evidential loss with the n^{th} data sample
$\mathcal{L}^{DREL}(\Theta)$	Distributionally robust evidential Loss
\mathcal{F}	Set of Different classifiers
σ_k	Weight associated with k^{th} weak learner
$p(c)$	Weight associated with the c^{th} class from Eq. (5.6)
$\widehat{p}(c)$	Readjusted weight associated with the c^{th} class from Eq. (5.9)
\mathbf{w}	Mixing weights associated with the MSF to control uncertainty set η_t
\mathbf{w}'	Mixing weights associated with MSF to readjust the class-specific weights
β	Set of Specific parameters for the SFs to control uncertainty set η_t
β'	Set of Specific parameters for the SFs to readjust the class-specific weights
\mathbf{W}	MSF Parameter sets associated in our model training
T	Total number of Epochs

mean vector gives rise to the probability of assigning \mathbf{x}_n to each class. There are multiple ways to design a loss function to train the evidential neural network [121]. A simple but effective option is the sum of square loss:

$$l_n^{EL}(\Theta) = \|\mathbf{y}_n - \mathbb{E}[\mathbf{p}_n]\|_2^2 = \sum_{c=1}^C (y_{nc}^2 - 2y_{nc}\mathbb{E}[p_{nc}] + \mathbb{E}[p_{nc}^2]) \quad (5.4)$$

Remark. Besides being used as a powerful model for close set classification, a unique benefit of evidential learning is that it offers a principled way to quantify the uncertainty mass, which is explicitly allocated to account for something that is ‘unknown’ to the model. Intuitively, a properly trained evidential model will output a high total evidence for data samples whose features are sufficiently exposed to the model during training. In contrast, it should predict a low total evidence for less representative samples in the training data. For these samples, their corresponding uncertainty mass u will be large (as the total mass sums to one). As a result, the uncertainty mass fits squarely for detecting open set samples, which have not been exposed to the model that is trained using the close set samples.

Robust Uncertainty Mass Quantification

The standard evidential learning does not explicitly consider an imbalanced class distribution. As a result, data samples from the minority class are usually assigned a higher uncertainty mass due to lack of sufficient training data. While this may not significantly impact the close set performance (*i.e.*, accuracy), it poses a more severe issue for OSD as the minority-class samples become equally uncertain as those open set samples. To address this challenge, we propose to integrate evidential learning with DRO for robust uncertainty mass quantification on the minority class samples in the close set. Since the model has less chance to learn from the minority class, DRO optimizes the worst cast loss by adjusting the weights assigned to each sample according to an uncertainty set:

$$\mathcal{P}^{DRO} := \left\{ \mathbf{p} \in \mathbb{R}^N : \mathbf{p}^\top \mathbf{1} = 1, \mathbf{p} \geq 0, D_f(\mathbf{p} \parallel \frac{\mathbf{1}}{N}) \leq \eta \right\} \quad (5.5)$$

where $D_f(\mathbf{p} \parallel \mathbf{q})$ is f -divergence between two distributions \mathbf{p} and \mathbf{q} and η controls the size of the uncertainty set. This allows us to define a distributionally robust evidential loss (DREL) as

$$\mathcal{L}^{DREL}(\Theta) = \max_{\mathbf{p} \in \mathcal{P}^{DRO}} \sum_{n=1}^N p_n l_n^{EL}(\Theta) \quad (5.6)$$

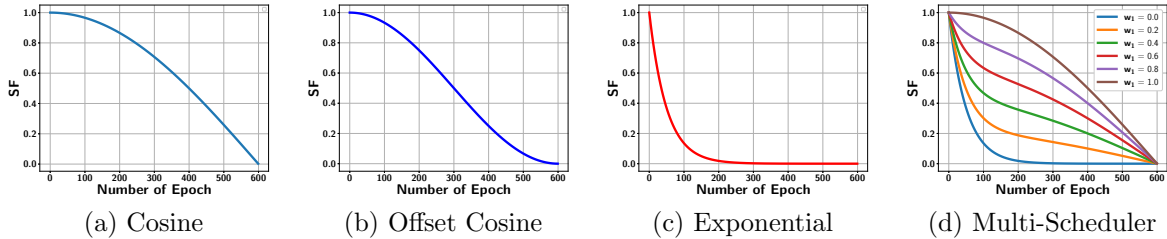


Figure 5.1: Examples of Scheduler Functions

Depending on η in the uncertainty set, we can decide whether we want to assign an equal weight to all data samples or focus on the most difficult ones. The lemma below reveals the relationship between DREL and the standard evidential loss.

Lemma 5.1. *With $\eta \rightarrow 0$, the EDL loss under DRO reduces to the standard EDL loss.*

When η is set to be very small, the model gives similar weights to all samples, which allows them to participate equally in the training process. At another extreme, we can direct the model to fully focus on the most difficult sample with the maximum loss, as summarized in the lemma below.

Lemma 5.2. *With $\eta \rightarrow \infty$, the loss under DRO becomes equivalent to a maximum loss based approach focusing only on the hardest sample.*

The above lemma implies that a highly flexible uncertainty set may cause the model to put too much emphasis on the difficult samples. Since these difficult samples may come from the majority classes, simply setting a large η will not be necessary to direct the model’s attention to the samples from the minority class. Furthermore, using a flexible uncertainty set in the initial phase of the model training may misguide the model to neglect a large number of representative data samples. As a result, the model will not be able to capture the common patterns that exhibit in most of the training samples. The proposed DREO model aims to address this issue by optimally balancing the model training over all classes in the close set through adaptive DRO learning.

Adaptive DRO Training

The key idea of adaptive DRO training is to gradually increase the size of the uncertainty set, which allows the model to learn from easy to hard samples from the close set classes. Scheduler functions (SF) provide a natural way to achieve the desired training behavior. Figure 5.1 (a-c) shows three typical SFs, including cosine in (a): $\cos\left(\frac{\pi t}{2T}\right)$, cosine in (b): $\frac{1}{2}\cos\left(\frac{\pi t}{T}\right) + \frac{1}{2}$, and exponential in (c): $\exp\left(-\frac{t}{\beta}\right)$, where t denotes the index of the training epoch, T is the terminating epoch, and β is

a specific parameter of the exponential function. It can be seen that while the general trends of different SFs are similar, they exhibit some key differences that may lead to quite distinct model training behaviors. For example, a cosine function can help to ensure the uncertainty set to stay small for a relatively longer time in the beginning of model training. This ensures the model to learn from the representative samples in the majority classes (according to Lemma 5.1). In contrast, an exponential function can change the size of the uncertainty set very rapidly, which can give the model more time to learn from the difficult samples at the later phase (according to Lemma 5.2). The offset cosine function can offer both a relatively long initial learning and later learning phases. However, choosing a SF that best matches the nature of a given dataset poses a key challenge. Furthermore, a single SF may not be rich enough to express the desired training behavior of a complex dataset.

To address this key challenge, we propose to conduct multi-scheduler learning to automatically construct a composite scheduler function that can be automatically learned for each given dataset to deliver the optimal training behavior. More specifically, the multi-scheduler function (MSF) is formulated as a convex combination of a set of atomic SFs:

$$\text{MSF}(\mathbf{w}, \boldsymbol{\beta}, t, T) = \sum_{m=1}^M w_m \text{SF}_m(\beta_m, t, T), \sum_{m=1}^M w_m = 1, w_m \geq 0 \quad \forall m \in [M] \quad (5.7)$$

where \mathbf{w} are the mixing weights and $\boldsymbol{\beta}$ is a set of specific parameters for the atomic SFs. Figure 5.1 (d) visualizes an example MSF that combines a cosine and exponential functions with different mixing weights and fixed $\beta = 20, T = 600$. As can be seen, the MSF is much more expressive than either its component SF, which makes it capable to represent a much broader range of training behaviors.

By leveraging the proposed MSF to control the size of the uncertainty set, we can achieve data adaptive DRO training. Let η_0 be the initial size of the uncertainty set and the size of the set at epoch t is

$$\eta_t = \frac{\eta_{t-1}}{\text{MSF}(\mathbf{w}, \boldsymbol{\beta}, t, T)} \quad (5.8)$$

As η_t increases, the model gradually shifts its focus from easier samples to the more difficult ones. In this way, the model can be trained to first capture the common patterns in the data and then conduct fine-tuning by attending to those difficult samples. However, for imbalanced classes, there may be a good number of difficult samples from the majority classes. Therefore, solely controlling the size of the uncertainty set does not guarantee a sufficient training over the minority class. To address this, we further leverage the label of the minority-class c to formulate a ratio biased weight

augmentation on samples from this class. Let $p(c) = \sum_{y_{nc}=1} p_n$ be the total weight for minority class c obtained by solving (5.6). Then, the weights for the minority class samples are adjusted as:

$$\widetilde{p(c)} = \begin{cases} p(c), & \text{if } p(c) \geq \frac{1}{C} \\ \min\left(\frac{1}{C}, p(c)^{\text{MSF}(\mathbf{w}', \beta', t, T)}\right), & \text{otherwise} \end{cases} \quad \widetilde{p}_n = \begin{cases} \frac{\widetilde{p(c)}}{p(c)} p_n, & \text{if } y_{nc} = 1 \\ \frac{1 - \widetilde{p(c)}}{1 - p(c)} p_n, & \text{otherwise} \end{cases} \quad (5.9)$$

As the MSF monotonically decreases over the training epochs, the total weight for the minority class samples will eventually reach $\frac{1}{C}$, making it equally weighted as the other $(C - 1)$ classes.

The adaptive DRO training is achieved through a bi-level optimization, where the inner loop optimizes the the model parameters (Θ) and the outer loop optimizes the MSF parameters $\mathbf{W} = \{\mathbf{w}, \mathbf{w}', \beta, \beta'\}$:

$$\min_{\mathbf{W}} \mathcal{L}_{val}^{DREL}(\Theta^*, \mathbf{W}), \text{ s.t. } \Theta^* = \arg \min_{\Theta} \mathcal{L}_{train}^{DREL}(\Theta, \mathbf{W}) \quad (5.10)$$

where $\mathcal{L}_{train}^{DREL}$, \mathcal{L}_{val}^{DREL} are training and validation losses, respectively. The outer loop optimization can be solved by computing the Hypergradients [91, 104] or through a population based methods [53], where the former may easily get stuck in local optimum [133]. To this end, we extend the existing population based method to learn an optimal MSF and the details are given in Appendix.

Theoretical Analysis

We establish the key theoretical properties of DREO, including the convergence speed in model training and the generalization capability by formally demonstrating the equivalence between DREO and AdaBoost. The key idea is to leverage the equivalence between AdaBoost and the gradient descent search of an optimal function from a linear combination of a set of (weak) learners [9, 95].

Let $\mathcal{F} = \{f_1, \dots, f_K\}$ be a set of different classifiers, and the linear span generated by the set \mathcal{F} is

$$\text{LS}(\mathcal{F}) = \left\{ f : f = \sum_{k=1}^K \sigma_k f_k, 1 \leq k \leq K \right\} \quad (5.11)$$

DREO training consists of two alternative updates between optimizing the worst case probability and predicting function f . The update in function prediction can be regarded as finding a sub-gradient $\mathcal{G}_t \in \partial \mathcal{L}_{DRO}(f_t)$ and updating with $\prod_{\text{LS}(\mathcal{F})}(\mathcal{G}) = \arg \min_{f \in \text{LS}(\mathcal{F})} \|f - \mathcal{G}_t\|_{\mathcal{D}_N}$ where \mathcal{D}_N is the training data. Letting $\mathcal{L}_n(f_t)$ be the loss associated with the data sample \mathbf{x}_n , the update of \mathbf{p} involves the optimization of the following objective with f_t being fixed:

$$\mathcal{L}^{DREO}(f_t) = \max_{\mathbf{p} \in \mathcal{P}^{DRO}} \sum_{n=1}^N p_n \mathcal{L}_n(f_t) \quad (5.12)$$

where the uncertainty set is given by (5.5). The corresponding Lagrangian of the above optimization problem is given by

$$\max_{\mathbf{p} \geq 0, \mathbf{p}^\top \mathbf{1} = 1} \sum_{n=1}^N p_n \mathcal{L}_n(f_t) - \alpha \left[\left(\sum_{n=1}^N p_n \log p_n \right) - \eta_t \right] \quad (5.13)$$

The theorem below shows the equivalence between DREO and Adaboost.

Theorem 5.3. *Under the assumption of finite exponential moment for $\mathcal{L}_n(f)$, with $\alpha \geq 0$ being sufficiently large and*

$$\eta_t = \beta^* \psi'(\beta^*) - \psi(\beta^*) \quad (5.14)$$

the worst case probability \mathbf{p}^* is given by

$$p_n^* = \frac{\exp\left(\frac{\mathcal{L}_n(f_t)}{\alpha}\right)}{\sum_{j=1}^N \exp\left(\frac{\mathcal{L}_j(f_t)}{\alpha}\right)} \quad (5.15)$$

where $\beta^* = \frac{1}{\alpha^*}$, $\alpha^* \geq 0$ be the optimal α , and $\psi(\beta) = \log \left[\frac{\sum_{n=1}^N \exp(\beta \mathcal{L}_n(f_t))}{N} \right]$. The alternative optimization between f and with above worst case probability solution exactly recovers the AdaBoost algorithm proposed in [38].

Remark. There are several key benefits of connecting DREO with AdaBoost. First, AdaBoost is less prone to overfitting even running for a large number of iterations [94]. Inheriting such a property is crucial for OSD as an overfitted evidential model can produce highly confident wrong predictions. This implies that a low uncertainty may be predicted for samples that the model is less familiar with, resulting in a false negative detection of an open set sample. Furthermore, since the target function is expressed as a linear combination of a set of weak learners, the optimal function can be regarded as maximizing the l_1 geometric margin among the training samples to ensure good generalization capability like other maximum-margin classifiers [95]. This ensures a decent close set performance from DREO (as shown by our experiments). The proof of Theorem 5.3 is provided in Appendix.

5.3 Experiments

We perform extensive experimentation to evaluate the effectiveness of the proposed DREO model. We first describe five real-world image datasets where a minority class is introduced to create an imbalanced setting. We then assess the OSD performance of the proposed technique by comparing with competitive baselines. Finally, we conduct some qualitative analysis, which uncovers deeper insights on the performance advantage of the proposed model.

Datasets

Our experiments involve five real-world image datasets: Cifar10, Cifar100 [67], ImageNet [24], MNIST [25], and Architecture Heritage Elements Dataset (AHED) [86]. In our experimentation, model training is performed solely based on the closed set samples. During the detection phase, the testing samples corresponding to the closed set classes will be assessed against the samples from open set classes. For all datasets, for the hyperparameter optimization, randomly selected 20% of the training set is used. The brief description for each dataset is given below. For the detailed description and data sample distribution in majority and minority classes, please refer to the Appendix.

- **MNIST:** Five classes are treated as open set classes and the rest as the class set. To make the dataset imbalanced, we consider class ‘3’ as a minority class and randomly select 30% data samples as compared with other majority classes. The same imbalanced ratio is applied to both training and testing sets. In addition to the MNIST open set classes as described above, we follow other existing works [130] and further test the OSD performance on additional open set samples from three more sources: (1) MNIST-Noise, (2) Noise, and (3) Omniglot [70].
- **Cifar10:** Five classes are assigned as open set and close set, respectively. Bird’ is made as the minority class using the same strategy introduced above. In addition to the open set classes from Cifar10 itself, we further assess the OSD performance with Cifar+10 and Cifar+50.
- **Cifar100:** ‘Living being’ related super classes are assigned as the closed set and the remaining super classes are assigned as the open set. We make ‘insect’ related classes as the minority one.
- **ImageNet:** Five classes are assigned as open set and close set, respectively. We make ‘king crab’ as the minority class.
- **Architectural Heritage Elements Dataset (AHED):** Five classes are assigned as open set and close set, respectively. This is inherently highly imbalanced dataset where number of data points are unevenly distributed across different classes. The class ‘portal’ is the minority one.

Experimental Settings

Evaluation metric. To assess the model performance, we report mean average precision (MAP) score which summarizes the precision-recall curve as a weighted mean of precision achieved at each threshold, with the increase in recall from previous threshold as the weight. Specifically, in the OSD, we treat the open set samples as positive and close set samples as negative and compute the MAP score based on the uncertainty score produced by the trained model. Different from AUROC, MAP places more emphasis on initial part of the ROC curve, which gives preference if

model can rank the openset samples on the top based on their predicted uncertainty scores. This MAP metric works well in practice as the main focus may be devoted to the first few predicted candidate samples, especially when there is a long candidate list. The theoretical result shows that MAP is approximately the AUROC times the initial precision of the model [128]. Therefore, we focus on reporting the MAP performance and leave the AUROC results in Appendix. It is worth to note that our AUROC results also show a consistent trend as the MAP results.

Network architecture. In terms of the architecture of the evidential neural network, for all datasets, we use an LeNet5 network with tanh activation in the feature extractor and ReLU in the fully connected layers. For training, we use the Adam optimizer with a learning rate of 0.001 and l_2 regularization with a coefficient of 0.001. The detailed hyperparameter setting is provided in Appendix.

Performance Comparison

In our comparison study, we include baselines that are most relevant to our model, including EDL, EDL augmented with oversampling using SMOTE [18] (referred to as AEDL), and EDL with standard DRO training (referred to as DRO). Further, we also compare the performance with the Posterior networks [17] and its robust form, PostNet (RS), proposed by Kopetzki et al. [65]. In addition, we also compare with representative baselines with outstanding OSD performance: OpenMAX [7], CGDL [130], and OLTR [85]. Please refer to the Appendix for the more detailed description of the baselines used in our comparison study along with additional results and an ablation study.

Tables 5.2 presents the OSD performance comparison between different models for all five datasets. DREO consistently outperforms all the baseline models across all the datasets. For certain datasets, the performance advantage over the second best model is more than or close to 10%. This clearly demonstrates the benefits of conducting evidential learning through adaptive DRO training to achieve an optimal balanced learning from all classes and different types of data samples. We also observe that EDL consistently performs better than other non-evidential learning based models, such as OpenMAX, in most cases. The better OSD performance from EDL is attributed to its explicit modeling of the uncertainty mass that works naturally for detecting the open set samples. In contrast, directly applying DRO with a flexible uncertainty set, which aims to address the imbalanced class distribution, leads to a rather poor OSD performance due to the reasons as analyzed in prior sections. Similarly, AEDL does not perform better than the standard EDL due to the lack of fine-tuning of the difficult examples from the majority classes that become inseparable

Table 5.2: OSD (MAP) performance on all datasets

Approach	Cifar10			Cifar100	ImageNet
	Cifar10	Cifar+10	Cifar+50		
EDL	62.42 ± 0.31	29.23 ± 0.38	65.75 ± 1.11	52.00 ± 2.40	55.93 ± 4.30
AEDL	54.19 ± 0.77	26.21 ± 0.68	63.04 ± 0.70	52.79 ± 0.91	57.94 ± 0.07
DRO	57.86 ± 2.94	18.35 ± 0.40	56.04 ± 1.63	50.78 ± 4.44	55.67 ± 3.86
OpenMAX	59.65 ± 1.03	24.48 ± 1.34	62.80 ± 2.08	50.88 ± 0.60	53.24 ± 0.39
CGDL	54.27 ± 2.06	16.83 ± 0.20	50.15 ± 1.08	50.59 ± 4.56	55.47 ± 1.53
PostNet	56.71 ± 6.08	25.71 ± 5.39	62.51 ± 4.68	53.85 ± 2.76	56.83 ± 1.52
PostNet (RS)	51.54 ± 11.32	18.28 ± 1.50	53.13 ± 4.33	51.75 ± 0.98	56.21 ± 0.75
OLTR	56.37 ± 0.25	19.59 ± 0.49	53.98 ± 0.68	48.48 ± 0.29	50.71 ± 0.66
DREO	72.48 ± 4.08	37.14 ± 2.06	73.87 ± 1.42	57.52 ± 1.60	62.02 ± 1.11
Approach	MNIST				AHED
	MNIST	Noise	MNIST-Noise	Omniglot	
EDL	87.32 ± 4.01	82.16 ± 8.74	82.89 ± 8.06	77.62 ± 6.79	50.23 ± 1.84
AEDL	75.37 ± 11.14	71.90 ± 11.45	76.23 ± 12.67	67.29 ± 10.77	52.22 ± 0.26
DRO	63.25 ± 4.32	46.78 ± 1.22	49.59 ± 3.98	48.15 ± 1.70	42.28 ± 0.18
OpenMAX	84.11 ± 1.55	83.03 ± 1.71	78.31 ± 2.745	81.14 ± 0.89	48.13 ± 0.19
CGDL	61.33 ± 1.53	74.88 ± 8.42	73.92 ± 8.41	90.72 ± 2.16	48.57 ± 1.39
PostNet	55.58 ± 9.12	47.53 ± 13.88	43.94 ± 8.00	72.79 ± 4.24	46.69 ± 1.90
PostNet (RS)	49.3 ± 4.071	36.14 ± 0.59	39.96 ± 2.20	77.43 ± 9.80	46.10 ± 4.37
OLTR	86.38 ± 0.51	90.61 ± 1.43	83.75 ± 1.28	55.27 ± 3.04	49.26 ± 1.66
DREO	90.80 ± 0.058	94.24 ± 0.32	94.18 ± 0.21	93.80 ± 0.2	53.21 ± 0.65

Table 5.3: Closed set performance (MAP) on all datasets

Approach	Cifar10	Cifar100	ImageNet	MNIST	AHED
EDL	55.39 ± 3.78	31.80 ± 2.37	55.84 ± 1.60	99.58 ± 0.26	40.48 ± 2.65
AEDL	54.98 ± 0.63	36.11 ± 0.08	55.62 ± 0.58	99.62 ± 0.23	41.36 ± 3.98
DRO	27.16 ± 5.94	10.50 ± 0.25	20.71 ± 1.00	90.87 ± 3.89	30.02 ± 0.93
DREO	54.65 ± 1.02	36.44 ± 0.23	55.31 ± 1.11	99.88 ± 0.01	49.68 ± 1.58

from the open set samples with a high predicted uncertainty score.

Table 5.3 also shows the closed set performance as a reference. It is interesting to see that DRO with a flexible uncertainty set performs the worst in the close set setting as it does not learn properly from the most representative samples in the training data while only focusing on the difficult ones. AEDL performs very competitively and achieves the best performances on two datasets. This is

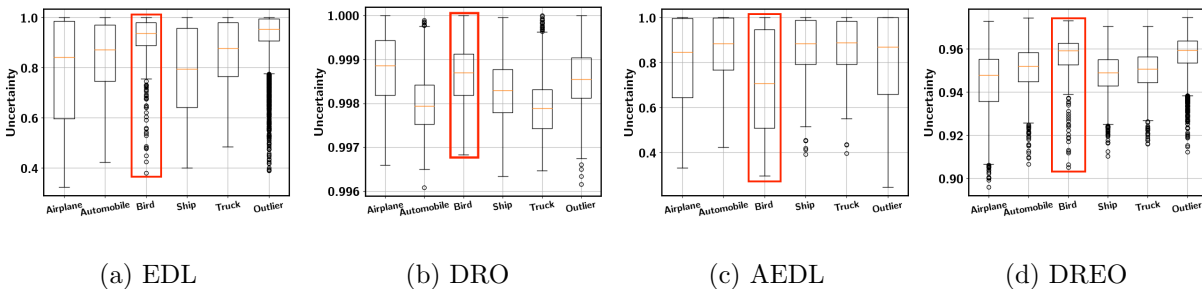


Figure 5.2: OSD performance comparison from imbalanced Cifar10 dataset.

partly because we are evaluating MAP by treating the minority class as positive and oversampling helps to improve the prediction on the minority class quite significantly. DREO also performs competitively and achieves the best performance on the other two datasets. The good close set performance further confirms our theoretical result that proves the equivalence between DREO and AdaBoost, which justifies its strong generalization capability.

Figure 5.2 provides a deeper insight on the superior OSD performance of DREO than other competitive baselines, including EDL, DRO, and AEDL. Cifar10 is used as an illustrative example and similar patterns are obtained on other datasets. First, while EDL is able to separate outliers from most samples in the majority classes based on their predicted uncertainty scores, it assigns much higher uncertainty scores to samples from the minority class, making them hard to be separated from the outliers. Second, the uncertainty scores for the majority classes span a wide range, which implies that several (difficult) samples from these classes have also been assigned very high uncertainty scores. If the goal is to ensure that most top-ranked samples are true outliers for effective detection in practice, these highly uncertain close-set samples may significantly affect the detection effectiveness. Third, while oversampling can help to better detect the samples from the minority class, which is indicated by lower uncertainty scores achieved by AEDL, most majority classes become much more uncertain and some of them have even a higher average uncertainty score than the outliers. Furthermore, the uncertainty scores from most classes also span a wide range. Finally, DRO effectively narrows down the range of the uncertainty scores as it allows the model to focus more on the difficult samples. However, it does not effectively bring down the high uncertainty scores of the minority class, either, which is still higher than outliers. Similar to DRO, the proposed DREO also manages to keep the uncertainty scores of data samples from the majority classes low so that even the difficult samples are unlikely to be mis-identified as outliers. Meanwhile, it effectively lowers the uncertainty scores of the minority-class examples so that they can better separated from the outliers.

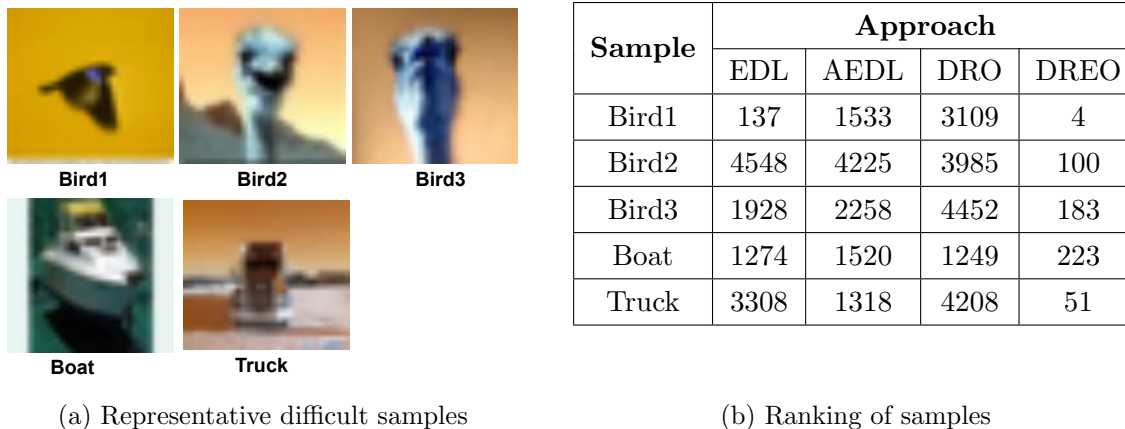


Figure 5.3: (a) Top row: minority class; bottom-row: majority classes; (b) sample ranking.

Qualitative Examples

We perform a qualitative analysis to further assess the effectiveness of DREO. Figure 5.3 (a) top row shows representative testing samples from the minority class (‘bird’) in Cifar10. These images appear to be difficult even for the humans to identify the bird as only a small part is visible. Thus, EDL, AEDL, and DRO assign a relative higher uncertainty score for them. As a result, many open set samples may be assigned a relatively lower uncertainty score, leading to false negative detection on these samples. Figure 5.3 (b) shows the ranking of these samples according to the uncertainty scores (a lower ranking indicates a lower uncertainty). In contrast, DREO assigns much lower rankings for these birds objects. This analysis justifies the effectiveness of DREO for detecting minority class data samples in the close set. Similarly, Figure 5.3 (a) bottom row show representative images from some majority classes. Again, DREO is able to recognize these difficult samples and assign a relatively low uncertainty score to avoid them being mis-identified as open set samples as shown by Figure 5.3 (b).

5.4 Conclusion

In summary, we focus on open set detection from imbalanced close set data. To address the fundamental challenge due to the interplay between the minority-class samples and difficult samples from the majority classes, we propose an important extension of DRO to the evidential learning setting, leading to a novel Distributionally Robust Evidential Optimization (DREO) model. As an evidential learning model, DREO effectively breaks the closed set assumption by explicitly modeling the uncertainty mass that is uniquely suitable for detecting open set samples. An adaptive

DRO training process is achieved through multi-scheduler learning to achieve an optimal training behavior. The experimentation conducted on five real-world datasets with diverse types of open set data samples justifies the effectiveness of the proposed model.

Chapter 6

Anomaly Detection under Few-Shot Learning Settings

Various learning strategies have been explored to reduce label dependency, including semi-supervised learning [16, 101] and weakly supervised learning [52, 116]. Few shot learning (FSL) offers another promising direction by assuming that only limited labeled data samples are available for model training [56]. Once trained, the model is expected to perform well on unseen data samples. While existing FSL models achieve promising results, most of them primarily focus on the closed-set setting, where both the training and test samples are assumed to be from the same data distribution over a common set of classes [89]. Nevertheless, when deployed in a practical setting, the model may very likely be exposed to samples from unknown classes, which are not part of the training distribution. In this case, it is ideal that the model can detect these samples as unknown. The open-set recognition (OSR) problem has been studied in the general setting with ample training data samples [8, 41, 130, 155]. However, the few-shot setting poses unique challenges, making existing solutions inadequate. There have been few attempts to address few-shot open-set recognition (FSOSR). For example, PEELER is designed to learn a high-entropy posterior distribution for samples from the open-set classes [80]. SnaTCHer further improves PEELER by leveraging transformer consistency [56]. It considers a set as a whole that includes all the prototypes of closed-set classes to detect the open-set ones. Because of the attention-based transformation of the entire set, this approach can provide a compact representation for the entire closed-set classes, leading to improved detection performance. However, when facing more challenging scenarios, where open-set classes share some similarities with closed-set ones, existing techniques become less effective.

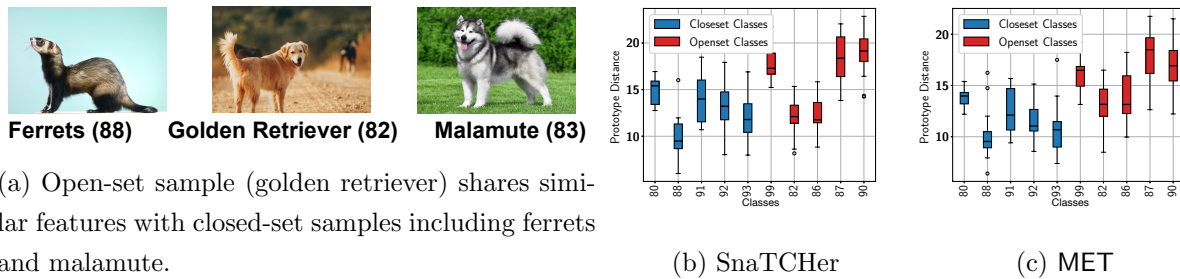


Figure 6.1: OSR performance (AUROC) of a difficult task consisting of similar closed- and open-set images with examples shown in (a). (b) SnaTCHer (72.84%) and (c) MET (83.34%) that uses Class Malamute (83) serves as the opponent class for better separation between Ferrets (closed) and Golden Retriever (open).

As shown in Figure 6.1 (a), Golden Retriever (Class ID: 82) shares some feature similarity (*e.g.*, body structure, whiskers, and tail) with the closed-set class Ferrets (Class ID: 88). As such, when a golden retriever is evaluated, it may be predicted as a Ferrets. In this case, the distance between the altered prototype (where the class Ferrets prototype is replaced by a golden retriever sample) and the original prototype will be very small, resulting in the misclassification of an open-set sample as a closed-set one with high confidence. As illustrated in Figure 6.1 (b), the mean prototype distance from this open-set class (*i.e.*, 82) is smaller than most other closed-set classes (*e.g.*, 80, 91, 92), leading to a relatively low detection rate with a 72.84 AUROC score. It is noted in the figure, prototype distance is the distance that tells how far the original prototype is from the altered prototype (the closest class prototype is replaced by a test sample) (see (6.9) for a definition).

Similar cases as described above may commonly occur in an open world. This makes it inherently challenging to detect open-set classes similar to certain closed-set ones but with subtle and important differences (*e.g.*, 82 and 86). Since recognizing open-set samples that are very different from their closed-set counterparts is relatively trivial with promising results achieved by existing methods, we will focus on attacking the more challenging cases. Since we have limited control over the open-set samples, the goal is to *learn a more compact representation of closed-set classes*. To this end, we propose a novel Meta Evidential Transformer (MET) that integrates uniquely designed training and inference modules to address the central challenges in FSOSR.

During training, MET leverages the power of similar closed-set classes playing a role as open-set samples (referred to as opponent classes) for improved model training. MET assigns a high uncertainty to the *opponent classes* that serve as training-time open-set samples. This will help the

model make (relatively) more confident predictions on the closed-set samples while being uncertain of unseen open-set samples that may share similar features as the opponent classes. To achieve this, a straightforward way would be enforcing the model to produce high uncertainty on the opponent class samples through the entropy maximization technique [80]. However, a high entropy cannot tell whether a sample is close to multiple closed-set classes or far away from all of them [126], where the former corresponds to a confusing closed-set sample and the latter is a true open-set one. To address this issue, we propose to integrate evidential learning [122], which allows us to design an evidence-based loss function to guide model training. Intuitively, a data sample with a small sum of evidence from all closed-set classes is more likely to be from the open-set while one with strong conflicting evidence from multiple classes should be a confusing closed-set sample. Figure 6.1 (b) shows an improved OSR performance by MET as compared with SnaTCHer.

While the use of a transformer coupled with the special training process allows us to improve the overall compactness of entire closed-set samples, one challenging issue still remains when a certain class is very different from others in the closed-set. Given an open-set sample that is relatively similar to this special closed-set class, it will also be very different from other classes (and their prototypes). Due to the normalization effect during transformation, this data sample will likely be assigned to the special closed-set class. We propose a novel *evidence-to-variance ratio (EVR)* to identify such cases during inference time. The inference module then conducts evidence-guided cross-attention in the transformer to improve detection performance with theoretical guarantees. Our main contributions are summarized below:

- a MET model that uses an evidential open-set loss to learn more compact closed-set representations by leveraging similar closed-set classes as opponent open-set classes,
- a novel evidence-to-variance ratio (EVR) to identify challenging open-set samples by collectively considering both the predicted evidence and their distribution over all closed-set classes,
- a uniquely designed evidence-based cross-attention mechanism to form a more accurate representation of the prototypes for improved OSR performance,

We conduct extensive experiments on real-world datasets and the results clearly demonstrate the outstanding OSR performance from the proposed MET model.

6.1 Related Work

We discuss representative works that are most relevant to ours. More related works are also covered in the Appendix.

Few-shot Open-set Recognition. There are recent OSR models specifically developed for few-shot learning under the meta-learning setting. Liu et al. propose an oPen set mEta LEaRning (PEELER) model that leverages ProtoNet for few shot open-set recognition [80], which makes an assumption that the unknown samples are available during the training process. The key limitation of this approach is the learned embedding representation is not task adaptive and the open-set detection process heavily depends on the used open-set samples during the training process. To address those limitations, Jeong et al. propose the SnaTCHer model based on FEAT [152], which makes the embedding task specific by leveraging different transformer functions [56]. While the training paradigm is very similar to FEAT (not requiring unknown samples), SnaTCHer proposes a unique process to detect unknown samples during testing by leveraging the transformed set of prototypes to represent all closed-set classes. However, SnaTCHer may suffer from more challenging open-set samples and the normalization effect may miss detecting open-set samples with strong confidence. Similarly, Huang et al. [51] leverage task-adaptive negative class prototype to learn dynamic rejection boundaries for FSOSR tasks. However, learning from negative samples generated from closed-set prototypes may not help to deal with challenging open-set samples.

Uncertainty-aware Open-set Recognition. Multiple approaches have been developed that explicitly consider uncertainty during model training [17, 92, 121]. For instance, Sensoy et al. propose an evidential deep learning (EDL) model that leverages the subjective logic principle to learn the evidence and uncertainty explicitly based on the training data samples [121]. Similarly, Malinin et al. propose a Prior Network (PN) that uses an explicit mechanism to quantify the distributional uncertainty coming from the distributional mismatch [92]. However, this approach requires unknown samples during the training time and therefore limits its applicability in practical settings. Considering this limitation, Charpentier et al. propose the posterior network that leverages the normalizing flows to estimate the density in the latent space in order to predict the posterior distribution based upon the in-distribution samples [17]. The proposed METmodel extends these approaches to the few-shot setting through seamless and novel integration with a transformer architecture for effective open-set recognition.

6.2 Methodology

6.2.1 Preliminaries

Meta learning. Meta learning refers to learning to learn where a meta-learner learns a learning algorithm by exploiting many learning tasks. Meta learning splits data into two sets: *meta-train* and *meta-test* considering distinct training and test classes. Meta-train $\mathcal{MS} = \{(\mathcal{S}_i^{tr}, \mathcal{Q}_i^{tr})\}_{i=1}^{N^{tr}}$ includes support (\mathcal{S}_i^{tr}) and query (\mathcal{Q}_i^{tr}) sets for the i^{th} task and N^{tr} is a number of training tasks. Similarly, meta-test $\mathcal{MT} = \{(\mathcal{S}_i^{te}, \mathcal{Q}_i^{te})\}_{i=1}^{N^{te}}$ includes support (\mathcal{S}_i^{te}) and query (\mathcal{Q}_i^{te}) sets for the i^{th} task and N^{te} is a number of test tasks. Meta-learning performs training by minimizing the error of label prediction for the query set \mathcal{Q}^{tr} conditioned on the support set \mathcal{S}^{tr} . Specifically, the meta-training objective is

$$\theta^* = \arg \min_{\theta} \sum_{(\mathbf{x}_j, y_j) \in T_i^{tr}} \mathcal{L}(y_j, P_{\theta}(\cdot | \mathbf{x}_j, \mathcal{S}_i^{tr})) \quad (6.1)$$

where \mathcal{L} is a loss function (*e.g.*, cross-entropy and mean-square error), which is suitable for the optimization procedure, and $P_{\theta}(\cdot)$ is a parametric neural network or other models to make predictions. Meta-learning is a popular approach for few-shot learning. It forms support and query sets by sampling N -classes from the set of classes with few training samples (*e.g.*, K -shot examples per class) commonly referred to as a N -way K -shot problem.

Evidential learning. Theory of evidence and subjective logic (SL) [22, 58] are utilized to address inexact and expensive posterior inference of Bayesian and Monte-Carlo approximation. It also provides predictive uncertainty, including both aleatoric and epistemic uncertainty. In particular, *evidence* provides a measure of the number of supportive observations from data for each class and let e_k denote the evidence for a class k . Then, the Dirichlet concentration parameter α_k for each class $k \in \mathbb{Y}$ can be calculated as: $\alpha_k = e_k + a_k W$, where $e_k \geq 0$. The belief mass and uncertainty mass (*a.k.a.*, vacuity) is computed as:

$$b_k = \frac{e_k}{S}, \quad u = \frac{K}{S} \text{ with } S = \sum_{k=1}^K (e_k + 1) \quad (6.2)$$

Evidential learning essentially places a Dirichlet prior $\text{Dir}(p_i | \alpha_i)$ on a multinomial likelihood $\text{Mult}(y_i | p_i)$ and then uses the negative log-likelihood to train the model:

$$\mathcal{L}_{\text{EDL}} = \sum_{k=1}^K y_{ik} (\log(S_i) - \log(\alpha_{ik})) \quad (6.3)$$

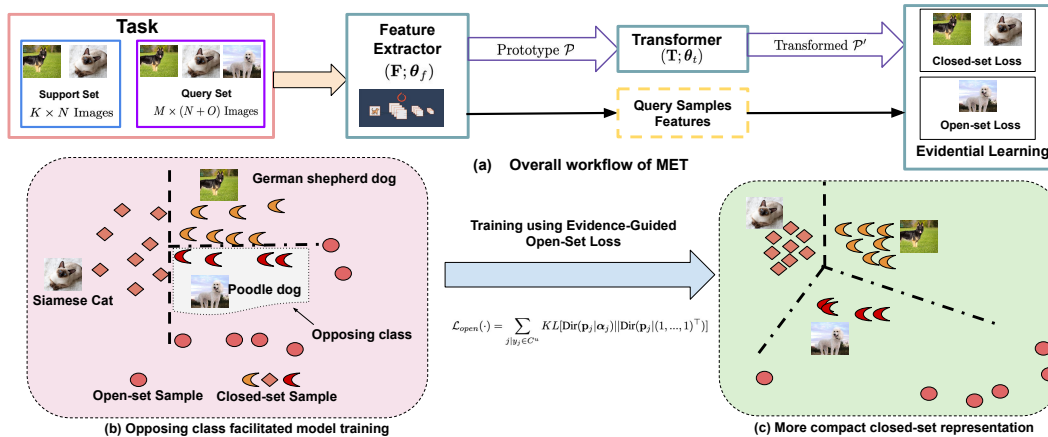


Figure 6.2: (a) MET training pipeline and opposing class selection (c) for compact closed-set representation learning (c).

where y_{ik} is an one-hot encoding of ground truth label y_i of a data sample \mathbf{x}_i , α_{ik} is a corresponding Dirichlet parameter and S_i is the total Dirichlet strength.

6.2.2 Transformer based FSOSR

Transformers leverage the similarity among the closed-set classes through the attention mechanism, which results in a more compact representation of the entire closed-set classes. As such, the open-set sample representation can stay away from all of the closed-set class representations, improving the openset detection capability. Let $F(\cdot)$ be the feature extractor and we can define the class-representation (*i.e.*, prototype) of closed-set class n as follow:

$$\mathbf{p}_n = \frac{1}{K} \sum_{\mathbf{x} \in \text{class } n} F(\mathbf{x}; \theta_f) \quad (6.4)$$

where K is the total number of samples belonging to class n in the support set, θ_f denotes the parameters associated with feature extractor F , \mathbf{x} represents a data sample belonging to class n , and N is the total number of closed-set classes for a given task. The overall prototype representation can then be formed as a concatenation of N closed-set class prototypes:

$$\mathcal{P} = \{\mathbf{p}_n\}_{n=1}^N \quad (6.5)$$

The above prototype representation does not leverage the similarity among closed-set classes. For a more compact representation, we can transform the prototype using the transformation function

$\mathsf{T}(\cdot)$. Specifically, we transform the prototype (\mathcal{P}) in the form of a triplet [*key* (\mathcal{K}), *value* (\mathcal{V}), *query* (\mathcal{Q})] with trainable transformer weight matrices and then the transformed prototype is achieved by

$$\mathcal{P}' = \mathsf{T}(\mathcal{P}; \boldsymbol{\theta}_t) = \text{LayerNorm}(\mathcal{P} + \frac{1}{N}(W_{\mathcal{V}}\mathcal{P})) \left[\text{softmax} \left(\frac{(W_{\mathcal{K}}\mathcal{P})^\top (W_{\mathcal{Q}}\mathcal{P})}{\sqrt{d}} \right)^\top \right] \quad (6.6)$$

where $\boldsymbol{\theta}_t = \{W_{\mathcal{K}}, W_{\mathcal{V}}, W_{\mathcal{Q}}\} \in \mathbb{R}^{d \times d}$ are learnable transformation matrices and d is the feature dimension.

During the training process, we aim to minimize the distance between a closed-set query sample feature representation with its respective transformed prototype class while maximizing with the rest. To achieve this, we can leverage cross-entropy loss with the following inverse of distance as the logits for that loss.

$$o_{jn} = [d(F(\mathbf{x}_j), \mathbf{p}'_n)]^{-1} = \left[\sqrt{(F(\mathbf{x}_j) - \mathbf{p}'_n)^\top (F(\mathbf{x}_j) - \mathbf{p}'_n)} \right]^{-1} \quad (6.7)$$

where \mathbf{x}_j is the j^{th} query sample, \mathbf{p}'_n is the transformed prototype of class n , and $d(\cdot, \cdot)$ is the Euclidean distance.

During the inference process, for open-set detection, one straightforward process would be computing the distance using (6.7) and deciding the sample as open-set or closed-set based on its value. Compared to this, SnaTCHer considers a set as a whole that includes all the prototypes of the closed-set classes to detect the open-set samples which results in better open-set detection performance. Specifically, let \mathbf{x}_j be a query sample, and c be the closest closed-set class with this sample, then we alter the prototype in (6.5) as

$$\mathcal{P}_a = \mathcal{P} - \{\mathbf{p}_c\} + F(\mathbf{x}_j) \quad (6.8)$$

Next, the altered prototype is passed through the transformer using (6.6). This yields the transformed representation of altered prototype represented as \mathcal{P}'_a . Finally, we compute the distance between transformed prototype and the altered transformed prototype which is given as

$$\delta(\mathbf{x}_j) = d(\mathcal{P}'_a, \mathcal{P}') \quad (6.9)$$

For an open-set sample, the transformed \mathcal{P}'_a is expected to be very different from the original \mathcal{P}' which has a compact representation, leading to an improved OSR performance.

Remarks. As described in the introduction, in case of more challenging scenarios where open-set classes share some similarities with closed-set classes, the existing techniques like SnaTCHer become less effective. Because of the feature similarity of open-set class sample with one of the closed-set sample, \mathcal{P}'_a can become similar to \mathcal{P}' , which compromises the open-set detection ability.

6.2.3 Meta Evidential Transformer (MET)

MET is designed to attack the most challenging few-shot open-set detection tasks that the state-of-the-art FSOSR techniques are less effective to handle. It integrates uniquely designed training and inference modules to achieve significantly improved OSR performance in these challenging settings. Specifically, training of MET is guided by a novel *evidential open-set loss* that learns a more compact closed-set representation by leveraging similar closed-set classes playing the role as open-set classes (referred to as opponent classes). As a result, open-set samples can be more easily separated from the closed-set ones falling into this compact representation. Another difficulty arises when the closed-set involves classes that are very different from all other closed-set classes. In this case, learning a compact representation that covers all the closed-set classes becomes challenging due to the large difference within the set. We propose a novel evidence-to-variance ratio (EVR) to identify such cases during the inference time using the predicted evidence by the trained evidential transformer. The inference module then conducts evidential cross-attention in the transformer to improve detection performance.

MET Training via Evidence-Guided Open-Set Loss. We first construct a meta-training (\mathcal{MS}) set consisting of only training classes so there is no overlap with samples from meta-test (\mathcal{MT}) classes.

Furthermore, we choose a set of opponent classes from the existing known closed-set classes to serve as open-set classes, aiming to learn a more compact representation of the known classes. Figure 6.2 (b) and (c) provides an illustrative example. We develop a unique mechanism to select opponent classes that are similar to some other closed-set classes. Specifically, within a training set, we perform semantic analysis at the class level to identify groups of semantically relevant classes (*e.g.*, different categories of dogs). For datasets with a relatively small number of classes, this introduces minimal overhead. For larger datasets, we can usually benefit from some existing hierarchical structure among the classes. If a hierarchical structure is unavailable, similar classes can be identified based on their semantic similarity. A neural network trained on a training dataset with a cross-entropy loss can be used to form a feature representation for each sample that could be used to pick similar classes.

Evidence Guided Evidential Loss. The overall training pipeline of MET is shown in Figure 6.2 (a) and the respective training algorithm is presented in Appendix. We proceed to conduct episodic training, where the test procedure mimics the training procedure. In training, we sample $(K + M)$

instances from N closed-set classes and form a support set (S_i^{tr}) utilizing K instances from each closed-set classes and a query set (Q_i^{tr}) with the M closed-set samples as well as O samples from open-set classes (*i.e.*, the chosen opponent classes) for the i^{th} training task $T_i^{tr}=(S_i^{tr}, Q_i^{tr})$. In this way, MET has a similar procedure as standard meta-learning and we propose the following learning process:

$$\boldsymbol{\theta}^* = \arg \min_{\boldsymbol{\theta}} \left\{ \sum_{(\mathbf{x}_j, y_j) \in T_i^{tr} | y_j \in C^s} \mathcal{L}_{close}(y_j, P_{\boldsymbol{\theta}}(\cdot | \mathbf{x}_j, S_i^{tr})) + \lambda \sum_{(\mathbf{x}_j, y_j) \in T_i^{tr} | y_j \in C^u} \mathcal{L}_{open}(P_{\boldsymbol{\theta}}(\cdot | \mathbf{x}_j, S_i^{tr})) \right\} \quad (6.10)$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_f, \boldsymbol{\theta}_t\}$ indicates total parameters consisting of both feature extractor parameter $\boldsymbol{\theta}_f$ and transformer parameter $\boldsymbol{\theta}_t$. \mathcal{L}_{close} is a closed-set loss and suitable loss functions include cross-entropy and mean-square error. However, our method is based on evidential learning so it leverages the evidential loss given in (6.3). Similarly, \mathcal{L}_{open} is an open-set loss applied to the open-set classes introduced into the training process and we will discuss our novel approach next. Also, C^s and C^u are the sets of closed-set classes and open-set classes of few-shot training task T_i^{tr} .

During the meta-update, the model uses the query set, which includes samples from those opponent classes chosen from the closed-set playing the role of challenging open-set classes. Our goal is to shrink the total evidence towards zero for these samples that effectively learn a more compact representation for the closed-set classes. To this end, we utilize KL-divergence between the predictive distribution on these open-set classes and a uniform distribution that indicates a maximum uncertainty mass (*i.e.*, $u = 1$):

$$\mathcal{L}_{open}(\cdot) = \sum_{j | y_j \in C^u} KL[\text{Dir}(\mathbf{p}_j | \boldsymbol{\alpha}_j) || \text{Dir}(\mathbf{p}_j | (1, \dots, 1)^\top)] \quad (6.11)$$

where \mathbf{p}_j represents the class probabilities of sample \mathbf{x}_j , Dir represents Dirichlet distribution, $\boldsymbol{\alpha}_j$ is the Dirichlet parameter given by represented as follow

$$\alpha_{jn} = \{e_{jn} + 1\}, \quad e_{jn} = o_{jn} \quad (6.12)$$

where o_{jn} is defined in (6.7), which is non-negative.

Evidential Cross-Attention. When the closed-set involves classes that are inherently different from other classes, learning a compact closed-set representation is more difficult. Using a loose representation may cause trouble during inference especially when evaluating a test open-set sample that is similar to one of the closed-set classes. Figure 6.3 (a) provides an illustrative example on

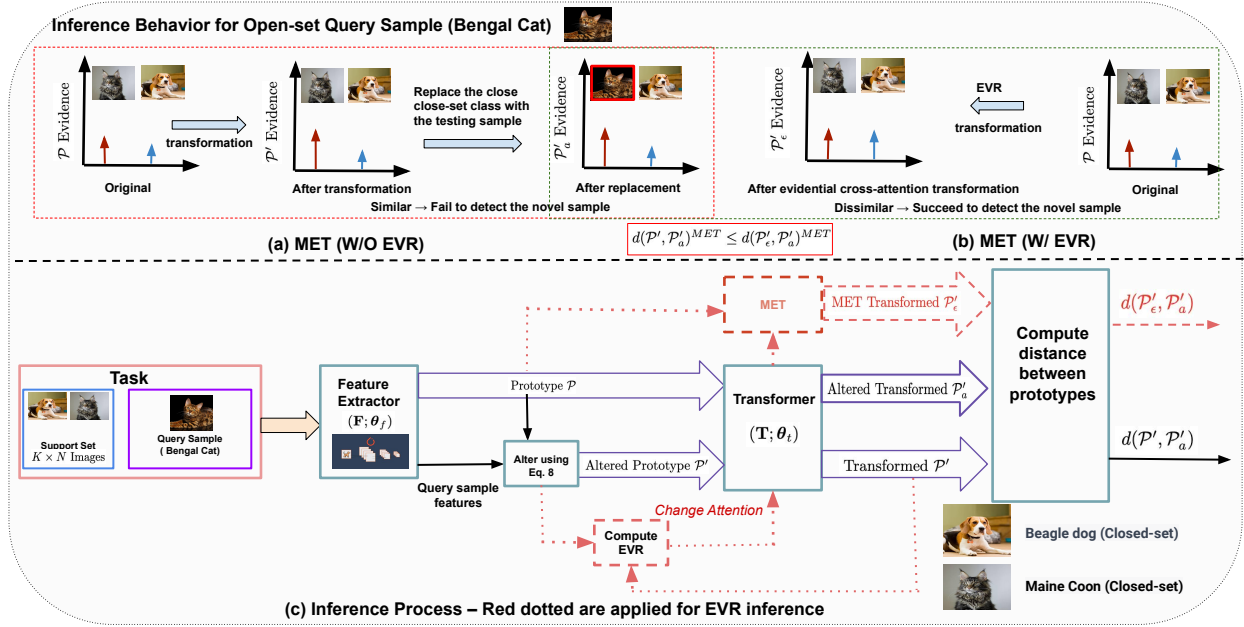


Figure 6.3: (a) A model trained using evidential loss fails to identify the Bengal Cat as open-set because it shares feature similarity with Maine Coon and is very distinct from the Dog class. (b) EVR and evidential cross-attention help to recognize the open-set sample. (c) Overall inference process that integrates EVR and evidential cross-attention.

this scenario. Due to the similarity between the open-set Bengal Cat and the original closed-set Maine Coon, after the latter is replaced by the former using (6.8), the transformed representation of the altered prototype (\mathcal{P}'_a) and that of the original prototype (\mathcal{P}') remains similar to each other, resulting in a low distance $d(\mathcal{P}', \mathcal{P}'_a)$. As a result, the model fails to recognize the open-set Bengal Cat. To improve the detection performance, the proposed inference module leverages the predicted evidence by MET to detect the challenging FSOSR tasks using a uniquely designed *Evidence-to-Variance Ratio (EVR)* metric. Once detected, it further performs *evidential cross-attention*, leading to a transformed representation (\mathcal{P}'_e) of the original prototypes that is more compact than \mathcal{P}' as shown in Figure 6.3 (b). Thus, the distance $d(\mathcal{P}'_e, \mathcal{P}'_a)$ becomes much larger, which results in a successful detection of the open-set Bengal Cat.

Let e_{jn} denote the predicted evidence on class n for the j -th sample in the query set of a meta-test task, *i.e.*, $j \in \mathcal{Q}_i^{te}$. The EVR metric is designed based on the following two properties of the predicted evidence:

- (P1) If j is an open-set sample, $\max_n [e_{jn}]$ is not high (since the model has not learned from

the same class); if j is a closed-set sample, $\max_n[e_{jn}]$ is high.

- **(P2)** For a challenging FSOSR task, if j is an open-set sample similar to some closed-set class n' , $\text{var}_{n \in N}[e_{jn}]$ is high (because a very low evidence for all other classes while a relative higher evidence for n'); if j is a closed-set sample, $\text{var}_{n \in N}[e_{jn}]$ is even higher (because a high evidence to the true closed-set class).

Guided by these properties, EVR is defined as

$$\text{EVR}_i = \frac{1}{|\mathcal{Q}_i^{te}|} \sum_{j \in \mathcal{Q}_i^{te}} \frac{\max_{n \in N}[e_{jn}]}{\text{var}_{n \in N}[e_{jn}]} \quad (6.13)$$

Lemma 6.1. *Consider a properly trained MET model that can predict evidence satisfying the two properties **(P1, P2)**. Given two FSOSR test tasks a and b , with a being a challenging task having a loose transformed closed-set representation \mathcal{P}' , and b being a regular task, we have $\text{EVR}_a < \text{EVR}_b$.*

Leveraging the key result in the lemma, we propose an evidential cross-attention mechanism to improve the OSR performance. Let c be the class nearest to the given query point \mathcal{Q}_{ij}^{te} , and $A_i \in \mathbb{R}^{N \times N}$ be the attention matrix obtained from the transformer network then, we update the attention as

$$A_i[c_1, c_2] = \begin{cases} A_i[c_1, c_2] \times \frac{\epsilon}{\text{EVR}_i} & \text{if } \text{cond} == \text{true} \\ A_i[c_1, c_2] & \text{else} \end{cases} \quad (6.14)$$

$$\text{cond} = \{(c_1 == c | | c_2 == c) \& c_1 \neq c_2\}$$

where ϵ is the threshold which is chosen in a way that $\epsilon > \text{EVR}_i$ for all tasks. The ratio $\frac{\epsilon}{\text{EVR}_i}$ should be large for challenging FSOSR tasks whereas small for easier tasks. As such, the ratio $\frac{\epsilon}{\text{EVR}_i}$ will have a minimal impact on the easier tasks whereas drastic effect on challenging tasks.

Theorem 6.2. *Consider a challenging FSOSR testing task and a properly trained MET model that can predict evidence satisfying the two properties **(P1, P2)**. Let \mathcal{P}' , \mathcal{P}'_a denote the transformed representations of the original prototypes and altered prototypes, respectively; let \mathcal{P}'_ϵ be the transformed representation of the original prototypes augmented through the evidential cross-attention. Then, with high probability the following holds true*

$$d(\mathcal{P}'_a, \mathcal{P}') \leq d(\mathcal{P}'_a, \mathcal{P}'_\epsilon) \quad (6.15)$$

Remarks. The evidential cross-attention essentially for the (originally) different closed-set classes to attend to each other through (6.14). This leads to a more compact representation \mathcal{P}'_ϵ . On the

other hand, the representation of the altered prototypes \mathcal{P}'_a is much less compact as it involves an open-set sample, leading to a large distance $d(\mathcal{P}'_a, \mathcal{P}'_e)$ that helps to improve the detection performance. Figure 6.3 (b) shows the impact of evidential cross-attention. The overall inference process is illustrated in Figure 6.3 (c), which integrates the EVR metric and the evidential cross-attention. For our inference algorithm along with the proof of Theorem 6.2 please refer to Appendix.

6.3 Experiments

We conduct experiments to evaluate the effectiveness of the proposed MET model. Through these experiments, we aim to demonstrate: (i) state-of-the-art open-set detection performance in comparison to existing competitive baselines, and (ii) deeper insights on better detection performance through a qualitative and ablation study.

Datasets and Evaluation Metrics. For the evaluation we conducted experimentation on multiple datasets, including MiniImageNet [141], TiredImageNet [112], Cifar100 [68], and Caltech101 [33]. Table C.2 in the Appendix shows the data split for each dataset. We present the results for MiniImageNet and TiredImageNet in the main chapter and leave the results from the other two datasets in the Appendix. Both MiniImageNet and TiredImageNet are subsets of ImageNet [23]. In MiniImageNet, there are a total of 100 classes with each class consisting of 600 low-resolution 84×84 RGB images. In the case of TiredImageNet, there are a total of 608 classes under 34 superclasses.

As our goal is open-set recognition, we use the Area Under the ROC curve (AUROC) as the detection performance metric. AUROC measures unseen class instance detection capability using both seen and unseen class samples. We set five classes as known classes and the other non-overlapped five classes as unknown classes to compose a single 5-way classification problem during the experiments. We collected 15 instances for each class as queries, which leads to 75 known queries and 75 unknown queries for a 5-way classification problem. We use 1 shot and 5 shot indicating the number of examples per class in the support set.

Comparison Baselines. We compare our method with the state-of-the-art few-shot learning and open-set recognition methods. For the few-shot learning method, we use a metric-based meta-learning method FEAT [152] due to their setting being close to our method. Similarly, we use PEELER [80], SnaTCHer [56], and TANE [51] as the state-of-the-art open-set recognition methods.

We also include a classical open-set detection baseline, *i.e.*, OpenMax [8]. Additionally, we have included two standard metric-based few-shot learning models: Prototypical Network [127] and Relation Network [131].

6.3.1 Results and Discussion

The comparison results are calculated over 1,000 evaluation episodes with 1,000 tasks per episode and computed the mean over 1,000 episodes. It is worth mentioning that we achieve a comparable closed-set accuracy with regard to competitive baselines as demonstrated in the Table C.3 (Appendix). Table 6.1 shows the open-set performance comparison between different competitive models proposed MET. As demonstrated, our approach has a much superior performance compared to the second-best TANE or SnaTCHer. In the case of MiniImageNet, the performance improvement is more than 3% in both settings compared to TANE. Furthermore, compared to SnaTCHer, our approach has more than 9% performance improvement in 1-shot and more than 6% in case of 5-shot setting.

Similarly for TieredImagenet, the performance gain over TANE for both 5-shot and 1-shot is around 4%. Compared to SnaTCHer, the performance improvement is more than 7% in the 1-shot and around 5% in the 5-shot setting. This justifies the effectiveness of our proposed technique. For other standard few-shot learning baselines, we leverage distance and relation scores between the prototype and the query sample to perform open-set recognition in prototypical and relational networks, respectively. Also, they don't have information about opponent classes and we compute the naive distance between prototypes and the query sample to provide its corresponding score. This results in poor performance of those models in all datasets. Similarly, another standard open-set recognition baseline called OpenMax has better results than other baselines in 1-shot miniImageNet but has poor performance compared to the proposed MET model in all cases.

6.3.2 Ablation Study

In this section, we first demonstrate the effectiveness of each proposed component. Next, we perform qualitative analysis to further justify the effectiveness of our proposed technique. Because of the limited space, we move some of the ablation studies to the Appendix. This includes the robustness of the

Table 6.2: Ablation study results on Mini-ImageNet.

	Transformer	Evidential Loss	EVR	AUROC	
				1-shot	5-shot
✓				63.56	77.99
		✓		74.35	81.47
		✓	✓	76.93	84.90

Table 6.1: OSD (AUROC) performance on different datasets.

Approaches	MiniImageNet 5-way		TieredImageNet 5-way	
	1-shot	5-shot	1-shot	5-shot
<i>ProtoNet</i>	51.63 \pm 0.47	60.26 \pm 0.56	58.48 \pm 0.50	63.46 \pm 0.24
<i>RelationNet</i>	53.14 \pm 0.67	62.22 \pm 0.78	60.85 \pm 0.68	64.42 \pm 0.57
<i>OpenMAX</i>	71.67 \pm 0.87	76.75 \pm 0.80	62.27 \pm 0.55	70.92 \pm 0.52
<i>FEAT (Probability)</i>	45.00 \pm 0.70	53.82 \pm 0.78	57.14 \pm 0.57	63.94 \pm 0.52
<i>Feat (Distance)</i>	67.71 \pm 0.92	75.32 \pm 0.84	61.52 \pm 0.58	70.77 \pm 0.52
<i>PEELER</i>	60.36 \pm 0.72	68.45 \pm 0.78	58.24 \pm 0.65	66.14 \pm 0.74
<i>SnaTCHer</i>	67.37 \pm 0.91	77.99 \pm 0.76	71.00 \pm 0.66	79.49 \pm 0.47
<i>TANE</i>	73.23 \pm 0.25	81.15 \pm 0.18	74.89 \pm 0.64	80.45 \pm 0.49
MET	76.93 \pm 0.59	84.90 \pm 0.41	78.77 \pm 0.46	84.37 \pm 0.35

proposed methodology for different backbones along with elaborative qualitative analysis. Further, we will also show in the Appendix that even in the original data split, the performance of the proposed technique is comparable/better than the existing techniques.

Different Proposed Components.

We conduct an ablation study to justify the effectiveness of each component, including the evidential loss and EVR-based detection. Table 6.2 shows the effectiveness of each component on the MiniImageNet dataset in the 5-Way 1-shot setting. As can be seen, the performance using proposed evidential loss (second row, w/o EVR) yields better compared to without using it (first row). Further combining both the evidential loss and EVR significantly boosts the performance as demonstrated in the third row of the table.

Qualitative Analysis

In addition to the ablation study, we perform a qualitative analysis to show the effectiveness of each proposed component (evidential loss and EVR). It should be noted that using SnaTCHer,

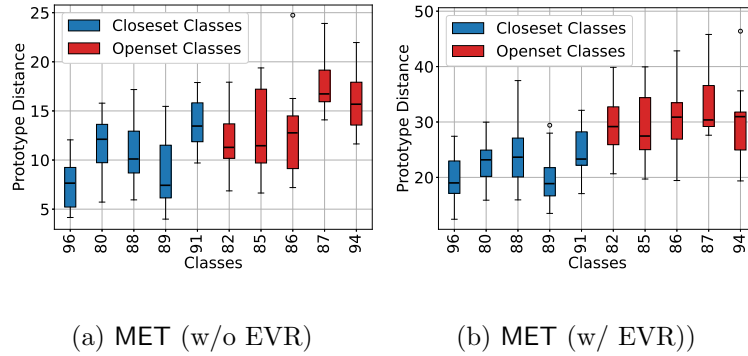


Figure 6.4: OSR performance comparison on MiniImageNet.

challenging open-set classes 82 and 85 have low prototype distance making them overlap with many closed-set classes. This is because, 82 and 85 share similarities with closed-set class 88. The proposed evidential loss helps to increase the separation between open-set and closed-set class samples. Specifically, as shown in 6.4 (a), the loss helps to push difficult open-set classes, including 82 and 85 upwards. As demonstrated in 6.4 (b), by leveraging EVR, we can further push those samples upward and thereby creating a bigger separation between open-set and closed-set samples. In terms of detection performance on AUROC, SnaTCHer, MET (w/o EVR), and MET (w/ EVR) achieve 65.16%, 73.40%, and 85.53%, respectively. For more detailed qualitative analysis along with visualization, please refer to the Appendix.

6.4 Conclusion

To tackle the FSOSR task, we propose a novel meta evidential transformer (MET) that uses an evidential open-set loss during training to learn more compact closed-set representation by leveraging similar closed-set classes. Furthermore, MET integrates an evidence-to-variance ratio to detect fundamentally challenging open-set samples by using an evidence-guided cross-attention mechanism. Experimental results on multiple real-world datasets demonstrate the effectiveness of the proposed technique over existing competitive methods in terms of better recognizing unseen class samples without deteriorating closed-set performance.

Chapter 7

Anomaly Detection under Sparse Network Training

While there is remarkable progress in developing deep neural networks with densely connected layers, most of these dense networks have poor calibration performance [45], limiting their applicability in safety-critical domains like self-driving cars [11] and medical diagnosis [57]. The poor calibration is mainly due to the fact that there exists a good number of wrongly classified data samples (*i.e.*, low accuracy) with high confidence resulting from the memorization effect introduced by an over-parameterized architecture [114]. Recent sparse network training methods, such as Lottery Ticket Hypothesis (LTH) [37] and its variants [6, 73, 81, 153, 161] generally assume that there exists a sparse sub-network (*i.e.*, lottery ticket) in a randomly initialized dense network, which could be trained in isolation and also match the performance of its dense counterpart network in terms of accuracy. While these methods may, to some extent, alleviate the overconfident issue, two key challenges remain to be addressed: (i) most of sparse network training methods require pre-training of a dense network followed by multi-step iterative pruning, making the overall training process highly costly, especially for large dense networks; (ii) even for techniques that do not rely on pre-training and iterative pruning (*e.g.*, Edge Popup or EP [110]), their learning goal focuses on pushing the accuracy up to the original dense networks and hence may still exhibit a severely over-fitting behavior, leading to a poor calibration performance as demonstrated in Figure 7.1 (b).

Inspired by the recent success of using ensembles to estimate uncertainties [71, 148], a potential solution to realize well-calibrated predictions would be training multiple sparse sub-networks and building an ensemble from them. As such, by leveraging accurate uncertainty quantification, the

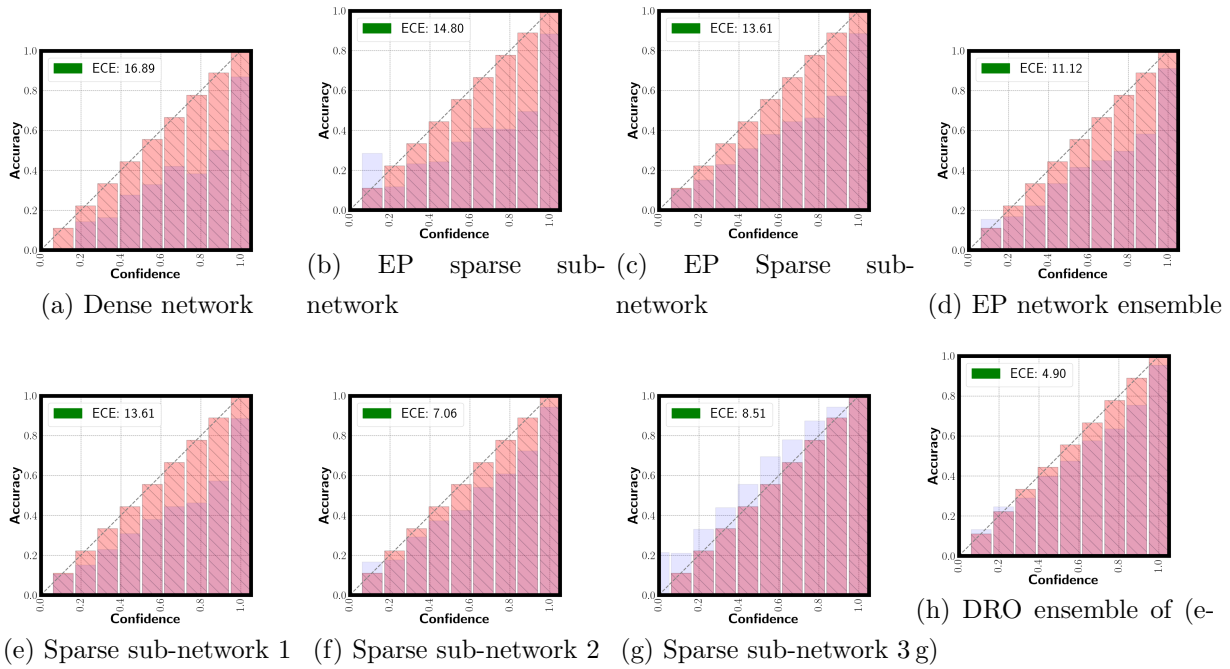


Figure 7.1: Calibration performance by expected calibration error (ECE) on Cifar100 dataset with ResNet101 architecture with density $\mathcal{K} = 15\%$. EP refers to the Edge Popup algorithm [110].

ensemble is expected to achieve better calibration. However, existing ensemble models of sparse networks rely on pre-training and iterative fine-tuning for learning each sub-network [81, 153], leading to a significant overhead for building the entire ensemble. Furthermore, an ensemble of independently trained sparse sub-networks does not necessarily improve the calibration performance. Since these networks are trained in a similar fashion from the same training data distribution, they could be strongly correlated such that the ensemble model will potentially inherit the overfitting behavior of each sub-network as shown in Figure 7.1(c). Therefore, the calibration capacity of sparse sub-network ensemble can be compromised as shown empirically in Figure 7.1 (d).

To further enhance the calibration of the ensemble, it is critical to ensure sufficient diversity among sparse sub-networks so that they are able to complement each other. One natural way to achieve diversity is to allow each sparse sub-network (ticket) to primarily focus on a specific part of training data distribution. This inspires us to leverage the AdaBoost [117] framework that sequentially finds tickets by manipulating training data distribution based on errors. By this means, the AdaBoost facilitates the training for a sequence of complementary sparse sub-networks. However, the empirical analysis (see Table 7.1) reveals that in the AdaBoost ensemble, most sub-networks (except for the first one) severely under-fit data leading to poor generalization ability. This is mainly because of the overfitting behavior of the first sub-network, which assigns very low training losses to the majority of data samples, making the subsequent sub-networks concentrate on very rare difficult

samples that are likely to be outliers or noises. Hence, directly learning from these difficult samples without having global knowledge of the entire training distribution will result in the failure of subsequent training tickets and also hurt the overall calibration.

To this end, we need a more robust learning process for proper training of complementary sparse sub-networks, each of which can be learned in an efficient way to ensure the cost-effective construction of the entire ensemble. We propose a Distributionally Robust Optimization (DRO) framework to schedule learning an ensemble of lottery tickets (sparse sub-networks) with complimentary calibration behaviors that contribute to an overall well-calibrated ensemble as shown in Figure 7.1 (e-h). Our technique directly searches sparse sub-networks in a randomly initialized dense network without pre-training or iterative pruning. Unlike the AdaBoost ensemble, the proposed ensemble ticket method starts from the original training distribution and eventually allows learning each sub-network from different parts of the training distribution to enrich diversity. This is also fundamentally different from existing sparse ensemble models [81, 153], which attempt to obtain diverse sub-networks in a heuristic way by relying on different learning rates. As a result, these models offer no guaranteed complementary behavior among sparse sub-networks to cover a different part of training data, which is essential to alleviate the overfitting behavior of the learned sparse sub-networks. In contrast, we realize a principled scheduling process by changing the uncertainty set of DRO, where a small set pushes sub-networks learning with easy data samples and a large set focuses on the difficult ones (see Figure 7.2). By this means, the ticket ensemble governed by our DRO framework could work complementary and lead to much better calibration ability as demonstrated in Figure 7.1(h). On the one hand, we hypothesize that the ticket found with easy data samples will tend to be learned and overfitted easily, resulting in overconfident predictions (Figure 7.1(e)). On the other hand, the ticket focused on more difficult data samples will be less likely to overfit and may become conservative and give under-confident predictions. Thus, it is natural to form an ensemble of such lottery tickets to complement each other in making calibrated predictions. As demonstrated in Figure 7.1 (h), owing to the diversity in the sparse sub-networks (e-g), the DRO ensemble exhibits better calibration ability. It is also worth noting that under the DRO framework, our sparse sub-networks already improve the calibration ability as shown in Figure 7.1 (f-g), which is further confirmed by our theoretical results.

Experiments conducted on three benchmark datasets demonstrate the effectiveness of our proposed technique compared to sparse counterparts and dense networks. Furthermore, we show through the experimentation that because of the better calibration, our model is being able to perform well on the distributionally shifted datasets [37] (CIFAR10-C and CIFAR100-C). The experiments also demonstrate that our proposed DRO ensemble framework can better detect open-set samples on

varying confidence thresholds. The contribution of this work can be summarized as follows:

- a new sparse ensemble framework that combines multiple sparse sub-networks to achieve better calibration performance without dense network training and iterative pruning.
- a distributionally robust optimization framework that schedules the learning of an ensemble complementary sub-networks (tickets),
- theoretical justification of the strong calibration performance by showing how the proposed robust training process guarantees to lower the confidence of incorrect predictions in Theorem 7.2,
- extensive empirical evidence on the effectiveness of the proposed lottery ticket ensemble in terms of competitive classification accuracy and improved open-set detection performance.

7.1 Related Work

Sparse networks training. Sparse network training has received increasing attention in recent years. Representative techniques include lottery ticket hypothesis (LTH) [37] and its variants [19, 140]. To avoid training a dense network, supermasks have been used to find the winning ticket in the dense network without training network weights [161]. Edge-Popup (EP) extends this idea by leveraging training scores associated with the neural network weights and only weights with top scores are used for predictions. There are two key limitations to most existing LTH techniques. First, most of them require pre-training of a dense network followed by multi-step iterative pruning making the overall training process expensive. Second, their learning objective remains as improving the accuracy up to the original dense networks and may still suffer from over-fitting (as shown in Figure 7.1).

Sparse network ensemble. There are recent advancements in building ensembles from sparse networks. A pruning and regrowing strategy has been developed in a model, called CigL [73], where dropout serves as an implicit ensemble to improve the calibration performance. CigL requires weight updates and performs pruning and growing for multiple rounds, leading to a high training cost. Additionally, dropping many weights may lead to a performance decrease, which prevents building highly sparse networks. This idea has been further extended by using different learning rates to generate different typologies of the network structure for each sparse network [81, 153]. While diversity among sparse networks can be achieved, there is no guarantee that this can improve the calibration performance of the final ensemble. In fact, different networks may still learn from the training data in a similar way. Hence, the learned networks may exhibit similar overfitting behavior with a high correlation, making it difficult to generate a well-calibrated ensemble. In contrast, the proposed DRO ensemble schedules different sparse networks to learn from complementary parts of

the training distribution, leading to improved calibration with theoretical guarantees.

Model calibration. Various attempts have been proposed to make the deep models more reliable either through calibration [45, 106, 140] or uncertainty quantification [40, 123]. Post-calibration techniques have been commonly used, including temperature scaling [45, 106], using regularization to penalize overconfident predictions [105]. Recent studies show that post-hoc calibration falls short of providing reliable predictions [103]. Most existing techniques require additional post-processing steps and an additional validation dataset. In our setting, we aim to improve the calibration ability of sparse networks without introducing additional post-calibration steps or validation dataset.

7.2 Methodology

Let $\mathcal{D}_N = \{\mathbf{X}, \mathbf{Y}\} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$ be a set of training samples where each $\mathbf{x}_n \in \mathbb{R}^D$ is a D -dimensional feature vector and $y_n \in [1, C]$ be associated label with C total classes. Let M be the total number of base learners used in the given ensemble technique. Further, consider \mathcal{K} to be the density ratio in the given network, which denotes the percentage of weights we keep during the training process. The major notations are summarized in the Appendix.

7.2.1 Preliminaries

Edge-Popup (EP) [110]. EP finds a lottery ticket (sparse sub-network) from a randomly initialized dense network based on the score values learned from training data. Specifically, to find the sub-network with density \mathcal{K} , the algorithm optimizes the scores associated with each weight in the dense network. During the forward pass, the top- \mathcal{K} weights in each layer are selected based on their scores. During the backward pass, scores associated with all weights are updated, which allows potentially useful weights that are ignored in previous forward passes to be re-considered.

Expected calibration error. Expected Calibration Error (ECE) measures the correspondence between predicted probability and empirical accuracy [97]. Specifically, mis-calibration is computed based on the difference in expectation between confidence and accuracy: $\mathbb{E}_{\hat{p}} [|\mathbb{P}(\hat{y} = y | \hat{p} = p) - p|]$.

In practice, we approximate the expectation by partitioning confidences into T bins (equally spaced) and take the weighted average on the absolute difference between each bins' accuracy and confidence. Let B_t denote the t -th beam and we have $\text{ECE} = \sum_{t=1}^T \frac{|B_t|}{N} |\text{acc}(B_t) - \text{conf}(B_t)|$.

7.2.2 Distributionally Robust Ensemble (DRE)

As motivated in the introduction, to further enhance the calibration of a deep ensemble, it is instrumental to introduce sufficient diversity among the component sparse sub-networks so that they can complement each other when forming the ensemble. One way to achieve diversity is to allow each sparse sub-network to primarily focus on a specific part of the training data distribution. Figure 7.2 provides an illustration of this idea, where the training data can be imagined to follow a multivariate Gaussian distribution with the red dot representing its mean. In this case, the first sub-network will learn the most common patterns by focusing on the training data close to the mean. The subsequent sub-networks will then learn relatively rare patterns by focusing on other parts of the training data (*e.g.*, two or three standard deviations from the mean).

AdaBoost ensemble. The above idea inspires us to leverage the AdaBoost framework [117] to manipulate the training distribution that allows us to train a sequence of complementary sparse sub-networks. In particular, we train the first sparse sub-network from the original training distribution, where each data sample has an equal probability to be sampled. In this way, the first sparse sub-network can learn the common patterns from the most representative training samples. Starting from the second sub-network, the training distribution is changed according to the losses suffered from the previous sub-network during the last round of training. This allows the later sub-networks to focus on the difficult data samples by following the spirit of AdaBoost.

However, our empirical results reveal that in the AdaBoost ensemble, most sub-networks (except for the first one) severely underfit the training data, leading to a rather poor generalization capability. This is caused by the overfitting behavior of the first sparse sub-network, which assigns very small training losses to a majority of data samples. As a result, the subsequent sub-networks can only focus on a limited number of training samples that correspond to relatively rare patterns (or even outliers and noises) in the training data. Directly learning from these difficult data samples without a general knowledge of the entire training distribution will result in the failure of training the sub-networks.

Distributionally robust ensemble (DRE). To tackle the challenge as outlined above, we need a more robust learning process to ensure proper training of complementary sparse sub-networks. Different from the AdaBoost ensemble, the training of all sub-networks starts from the original training distribution in the DRO framework. Meanwhile, it also allows each sub-network to eventually focus on learning from different parts of the training distribution to ensure the desired diverse and complementary behavior. Let $l(\mathbf{x}_n, \Theta)$ denote the loss associated with the n^{th} data sample with

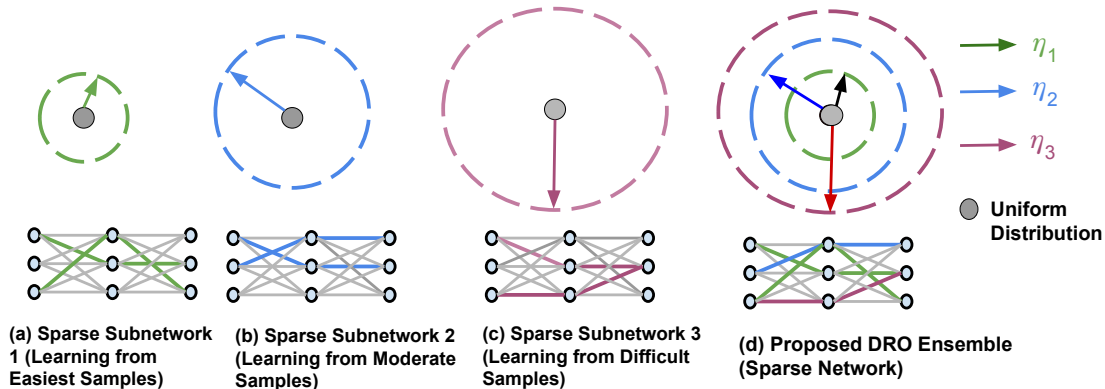


Figure 7.2: Robust ensemble where η defines the size of an uncertainty set with $\eta_1 \leq \eta_2 \leq \eta_3$.

Θ being the parameters in the sparse sub-network. Then, the total loss is given by

$$\mathcal{L}^{\text{Robust}}(\Theta) = \max_{\mathbf{z} \in \mathcal{U}^{\text{Robust}}} \sum_{n=1}^N z_n l(\mathbf{x}_n, \Theta) \quad (7.1)$$

The uncertainty set defined to assign weights \mathbf{z} is given as

$$\mathcal{U}^{\text{Robust}} := \left\{ \mathbf{z} \in \mathbb{R}^N : \mathbf{z}^\top \mathbf{1} = 1, \mathbf{z} \geq 0, D_f(\mathbf{z} \| \frac{\mathbf{1}}{N}) \leq \eta \right\} \quad (7.2)$$

where $D_f(\mathbf{z} \| \mathbf{q})$ is f -divergence between two distributions \mathbf{z} and \mathbf{q} and η controls the size of the uncertainty set and $\mathbf{1} \in \mathbb{R}^N$ is N -dimensional unit vector. Depending on the η value, the above robust framework instantiates different sub-networks. For example, by making $\eta \rightarrow \infty$, we have $\mathcal{U}^{\text{Robust}} = \{ \mathbf{z} \in \mathbb{R}^N : \mathbf{z}^\top \mathbf{1} = 1, \mathbf{z} \geq 0, D_f(\mathbf{z} \| \frac{\mathbf{1}}{N}) \leq \infty \}$. In this case, we train a sub-network by only using the most difficult sample in the training set. On the other extreme with $\eta \rightarrow 0$, we have $\mathcal{U}^{\text{Robust}} = \{ \mathbf{z} \in \mathbb{R}^N : \mathbf{z}^\top \mathbf{1} = 1, \mathbf{z} \geq 0, D_f(\mathbf{z} \| \frac{\mathbf{1}}{N}) \leq 0 \}$, which assigns equal weights to all data samples. So, the sub-network learns from the original training distribution.

To fully leverage the key properties of the robust loss function as described above, we propose to perform distributionally robust ensembling learning to generate a diverse set of sparse sub-networks with well-controlled overfitting behavior that can collectively achieve superior calibration performance. The training process starts with a relatively small η value to ensure that the initially generated sub-networks can adequately capture the general patterns from the most representative data samples in the original training distribution. The training proceeds by gradually increasing the η value, which allows the subsequent sub-networks to focus on relatively rare and more difficult data samples. As a result, the later generated sub-networks tend to produce less confident predictions that complement the sub-networks generated in the earlier phase of the training process. This diverse and complementary behavior among different sparse sub-networks is clearly illustrated in

Figure 7.1 (e)-(g). During the ensemble phase, we combine the predictions of different sub-networks in the logit space by taking the mean and then performing the softmax. In this way, the sparse sub-networks with high η values help to lower the overall confidence score, especially those wrongly predicted data samples. Furthermore, the sub-networks with lower η values help to bring up the confidence score of correctly predicted data samples. Thus, the overall confidence score will be well compensated, resulting in a better calibrated ensemble.

7.2.3 Theoretical Analysis

In this section, we theoretically justify why the proposed DRE framework improves the calibration performance by extending the recently developed theoretical framework on multi-view learning [1]. In particular, we will show how it can effectively lower the model’s false confidence on its wrong predictions resulting from spurious correlations. For this, we first define the problem setup that includes some key concepts used in our theoretical analysis. We then formally show that DRO helps to decorrelate the spurious correlation by learning from less frequent features that characterize difficult data samples in a training dataset. This important property further guarantees better calibration performance of DRO as we show in the main theorem.

Problem setup. Assume that each data sample $\mathbf{x}_n \in \mathbb{R}^D$ is divided into P total patches, where each patch is a d -dimensional vector. For the sake of simplicity, let us assume each class $c \in [1, C]$ has two characterizing (major) features $\mathbf{v}_c = \{\mathbf{v}_{c,l}\}_{l=1}^L$ with $L = 2$. For example, the features for **Cars** could be **Headlights** and **Tires**. Let \mathcal{D}_N^S and \mathcal{D}_N^M denote the set of *single-view* and *multi-view* data samples, respectively, which are formally defined as

$$\begin{cases} \{\mathbf{x}_n, y_n\} \in \mathcal{D}_N^S & \text{if one of } \mathbf{v}_{c,1} \text{ or } \mathbf{v}_{c,2} \text{ appears along with some noise features} \\ \{\mathbf{x}_n, y_n\} \in \mathcal{D}_N^M & \text{if both } \mathbf{v}_{c,1} \text{ and } \mathbf{v}_{c,2} \text{ appears along with some noise features} \end{cases} \quad (7.3)$$

The noise features (also called minor features) refer to those that do not characterize (or differentiate) a given class c (*e.g.*, being part of the background). In important applications like computer vision, images supporting such a "multi-view" structure is very common [1]. For example, for most car images, we can observe all main features, such as **Wheels**, **Tires**, and **Headlights** so they belong to \mathcal{D}_N^M . Meanwhile, there may also be images, where multiple features are missing. For example, if the car image is taken from the front, the tire and wheel features may not be captured. In most real-world datasets, such single-view data samples are usually much limited as compared to their multi-view counterparts. The Appendix provides concrete examples of both single and multi-view images. Let us consider $(\mathbf{x}, y) \in \mathcal{D}_N^S$ with the major feature $\mathbf{v}_{c,l}$ where $y = c$. Then

each patch $\mathbf{x}^p \in \mathbb{R}^d$ can be expressed as

$$\mathbf{x}^p = a^p \mathbf{v}_{c,l} + \sum_{\mathbf{v}' \in \cup \setminus \mathbf{v}_c} \alpha^{p,\mathbf{v}'} \mathbf{v}' + \epsilon^p \quad (7.4)$$

where $\cup = \{\mathbf{v}_{c,1}, \mathbf{v}_{c,2}\}_{c=1}^C$ is collection of all features, $a^p > 0$ is the weight allocated to feature $\mathbf{v}_{c,l}$, $\alpha^{p,\mathbf{v}'} \in [0, \gamma]$ is the weight allocated to the noisy feature \mathbf{v}' that is not present in feature set \mathbf{v}_c *i.e.*, $\mathbf{v}' \in \cup \setminus \mathbf{v}_c$, and $\epsilon^p \sim \mathcal{N}(0, (\sigma^p)^2 \mathbb{1})$ is a random Gaussian noise. In (7.4), a patch \mathbf{x}^p in a single-view sample \mathbf{x} also contains set of minor (noise) features presented from other classes *i.e.*, $\mathbf{v}' \in \cup \setminus \mathbf{v}_c$ in addition to the main feature $\mathbf{v}_{c,l}$. Since $\mathbf{v}_{c,l}$ characterizes class c , we have $a^p > \alpha^{p,\mathbf{v}'}; \forall \mathbf{v}' \in \cup \setminus \mathbf{v}_c$. However, since the single-view data samples are usually sparse in the training data, it may prevent the model from accumulating a large a^p for $\mathbf{v}_{c,l}$ as shown Lemma 7.1 below. In contrast, some noise \mathbf{v}' may be selected as the dominant feature (due to spurious correlations) to minimize the errors of specific training samples, leading to potential overfitting of the model.

We further assume that the network contains H convolutional layers, which outputs $F(\mathbf{x}; \Theta) = (F_1(\mathbf{x}), \dots, F_C(\mathbf{x})) \in \mathbb{R}^C$. The logistic output for the c^{th} class can be represented as

$$F_c(\mathbf{x}) = \sum_{h \in [H]} \sum_{p \in [P]} \text{ReLU}[\langle \Theta_{c,h}, \mathbf{x}^p \rangle] \quad (7.5)$$

where $\Theta_{c,h}$ denote the h^{th} convolution layer (feature map) associated with class c . Under the above data and network setting, we propose the following lemma.

Lemma 7.1. *Let $\mathbf{v}_{c,l}$ be the main feature vector present in the single-view data \mathcal{D}_N^S . Assume that number of single-view data samples containing feature $\mathbf{v}_{c,l}$ is limited as compared with the rest, *i.e.*, $N_{\mathbf{v}_{c,l}} \ll N_{\cup \setminus \mathbf{v}_{c,l}}$. Then, at any iteration $t > 0$, we have*

$$\langle \Theta_{c,h}^{t+1}, \mathbf{v}_{c,l} \rangle = \langle \Theta_{c,h}^t, \mathbf{v}_{c,l} \rangle + \beta \max_{\mathbf{z} \in \mathcal{U}} \sum_{n=1}^N z_n [\mathbb{1}_{y_j=c}(V_{c,h,l}(\mathbf{x}_n) + \kappa)(1 - \text{SOFT}_c(F(\mathbf{x}_n)))] \quad (7.6)$$

where κ is a dataset specific constant, β is the learning rate, SOFT_c is the softmax output for class c , and $V_{c,h,l}(\mathbf{x}_j) = \sum_{p \in \mathcal{P}_{\mathbf{v}_{c,l}}(\mathbf{x}_j)} \text{ReLU}(\langle \Theta_{c,h}, \mathbf{x}_j^p \rangle a^p)$ with $\mathcal{P}_{\mathbf{v}_{c,l}}(\mathbf{x}_j)$ being the collection of patches containing feature $\mathbf{v}_{c,l}$ in \mathbf{x}_j . The set \mathcal{U} is an uncertainty set that assigns a weight to each data sample based on its loss. In particular, the uncertainty set under DRO is given as in (7.2) and we further define the uncertainty set under ERM: $\mathcal{U}^{ERM} := \{\mathbf{z} \in \mathbb{R}^N : z_n = \frac{1}{N}; \forall n \in [1, N]\}$.

Learning via the robust loss in (7.1) leads to a stronger correlation between the network weights $\Theta_{c,h}$ and the single-view data feature $\mathbf{v}_{c,l}$:

$$\{\langle \Theta_{c,h}^t, \mathbf{v}_{c,l} \rangle\}_{\text{Robust}} > \{\langle \Theta_{c,h}^t, \mathbf{v}_{c,l} \rangle\}_{\text{ERM}}; \forall t > 0 \quad (7.7)$$

Remark. The robust loss $\mathcal{L}^{\text{Robust}}$ forces the model to learn from the single-view samples (according to the loss) by assigning a higher weight. As a result, the network weights will be adjusted to increase the correlation with the single-view data features $\mathbf{v}_{c,l}$ due to Lemma 7.1.

In contrast, for standard ERM, weight is uniformly assigned to all samples. Due to the sparse single-view data features (which also makes them more difficult to learn from, leading to a larger loss), the model does not grow sufficient correlation with $\mathbf{v}_{c,l}$. In this case, the ERM model instead learns to memorize some noisy feature \mathbf{v}' introduced through certain spurious correlations. For a testing data sample, the ERM model may confidently assign it to an incorrect class k according to the noise feature \mathbf{v}' . In the theorem below, we show how the robust training process can effectively lower the confidence of incorrect predictions, leading to an improved calibration performance.

Theorem 7.2. *Given a new testing sample $\mathbf{x} \in \mathcal{D}_S^N$ containing $\mathbf{v}_{c,l}$ as the main feature and a dominant noise feature \mathbf{v}' that is learned due to memorization, we have*

$$\{\text{SOFT}_k(\mathbf{x})\}_{\text{Robust}} < \{\text{SOFT}_k(\mathbf{x})\}_{\text{ERM}} \quad (7.8)$$

where \mathbf{v}' is assumed to be a main feature characterizing class k .

Remark. For ERM, due to the impact of the dominate noise feature \mathbf{v}' , it assigns a large probability to class k since \mathbf{v}' is one of its major features, leading to high confidence for an incorrect prediction. In contrast, the robust learning process allows the model to learn a stronger correlation with the main feature $\mathbf{v}_{c,l}$ as shown in Lemma 7.1. Thus, the model is less impacted by the noise feature \mathbf{v}' , resulting in reduced confidence in predicting the wrong class k . Such a key property guarantees an improved calibration performance, which is clearly verified by our empirical evaluation. It is also worth noting that Theorem 7.2 does not necessarily lead to better classification accuracy. This is because (7.8) only ensures that the false confidence is lower than an ERM model, but there is no guarantee that $\{\text{SOFT}_k(\mathbf{x})\}_{\text{Robust}} < \{\text{SOFT}_c(\mathbf{x})\}_{\text{Robust}}$. It should be noted that our DRE framework ensures diverse sparse sub-network focusing on different single-view data samples from different classes. As such, an ensemble of those diverse sparse subnetworks provides maximum coverage of all features (even the weaker one) and therefore can ultimately improve the calibration performance. The detailed proofs are provided in the Appendix.

7.3 Experiments

We perform extensive experimentation to evaluate the distributionally robust ensemble of sparse sub-networks. Specifically, we test the ability of our proposed technique in terms of calibration and

classification accuracy. For this, we consider three settings: (a) general classification, (b) out-of-distribution setting where we have in-domain data but with different distributions, and (c) open-set detection, where we have unknown samples from new domains.

7.3.1 Experimental Settings

Dataset description. For the general classification setting, we consider three real-world datasets: Cifar10, Cifar100 [67], and TinyImageNet [72]. For the out-of-distribution setting, we consider the corrupted version of the Cifar10 and Cifar100 datasets which are named Cifar10-C and Cifar100-C [49]. It should be noted that in this setting, we train all models in clean dataset and perform testing in the corrupted datasets. For open-set detection, we use the SVHN dataset [100] as the open-set dataset and Cifar10 and Cifar100 as the close-set data. A more detailed description of each dataset is presented in the Appendix.

Evaluation metrics. To assess the model performance in the first two settings, we report the classification accuracy (\mathcal{ACC}) along with the Expected Calibration Error (\mathcal{ECE}). In the case of open-set detection, we report open-set detection for different confidence thresholds.

Implementation details. In all experiments, we use a family of ResNet architectures with two density levels: 9% and 15%. To construct an ensemble, we learn 3 sparse sub-networks each with a density of 3% for the total of 9% density and that of 5% density for the total of density 15%. All experiments are conducted with the 200 total epochs with an initial learning rate of 0.1 and a cosine scheduler function to decay the learning rate over time. The last-epoch model is taken for all analyses. For the training loss, we use the EP-loss in our DRO ensemble that optimizes the scores for each weight and finally selects the sub-network from the initialized dense network for the final prediction. The selection is performed based on the optimized scores. More detailed information about the training process and hyperparameter settings can be found in the Appendix.

7.3.2 Performance Comparison

In our comparison study, we include baselines that are relevant to our technique and therefore we primarily focus on the LTH-based techniques. Specifically, we include the initial lottery ticket hypothesis (LTH) [37] that iteratively performs pruning from a dense network until the randomly initialized sub-network with a given density is reached. Once the sub-network is found, the model trains the sub-network using the training dataset. Similarly, we also include L1 pruning [75]. We also include three approaches CigL [73], Sup-ticket [153], DST Ensemble [81] which are based on

Table 7.1: Accuracy and ECE performance with 9% density for Cifar10 and Cifar100.

Training Type	Approach	Cifar10				Cifar100			
		ResNet50		ResNet101		ResNet101		ResNet152	
		<i>ACC</i>	<i>ECE</i>	<i>ACC</i>	<i>ECE</i>	<i>ACC</i>	<i>ECE</i>	<i>ACC</i>	<i>ECE</i>
	<i>Dense</i> [†]	94.82	5.87	95.12	5.99	76.40	16.89	77.97	16.73
Dense Pre-training	<i>L1 Pruning</i> [75]	93.45	5.31	93.67	6.14	75.11	15.89	75.12	16.24
	<i>LTH</i> [37]	92.65	3.68	92.87	6.02	74.09	15.45	74.41	16.12
	<i>DLTH</i> [6]	93.27	5.87	95.12	7.09	77.29	16.64	77.86	17.26
	<i>Mixup</i> [140]	92.86	3.68	93.06	6.01	74.15	15.41	74.28	16.05
Sparse Training	<i>CigL</i> [73]	92.39	5.06	93.41	4.60	76.40	9.30	76.46	9.91
	<i>DST Ensemble</i> [81]	88.87	2.02	84.93	0.8	63.57	7.23	63.22	6.18
	<i>Sup-ticket</i> [153]	94.52	3.30	95.04	3.10	78.28	10.20	78.60	10.50
Mask Training	<i>AdaBoost</i>	93.12	5.13	94.15	5.46	75.15	22.96	75.89	24.54
	<i>EP</i> [110]	94.20	3.97	94.35	4.03	75.05	14.62	75.68	14.41
	<i>SNE</i>	94.70	2.51	94.48	3.51	75.69	9.02	75.22	10.89
	DRE (Ours)	94.60	0.7	94.28	0.7	74.68	1.20	74.37	2.09

Table 7.2: Accuracy and ECE on TinyImageNet.

Training Type	Approach	$\mathcal{K} = 9\%$				$\mathcal{K} = 15\%$			
		ResNet101		WideResNet101		ResNet101		WideResNet101	
		<i>ACC</i>	<i>ECE</i>	<i>ACC</i>	<i>ECE</i>	<i>ACC</i>	<i>ECE</i>	<i>ACC</i>	<i>ECE</i>
	<i>Dense</i> [†]	71.28	15.58	72.57	16.96	71.28	15.58	72.57	16.96
Dense Pre-training	<i>L1 Pruning</i> [75]	68.85	14.72	69.78	16.38	70.24	14.24	70.98	15.36
	<i>LTH</i> [37]	69.23	13.97	69.13	15.34	70.16	13.63	70.25	14.24
	<i>DLTH</i> [6]	70.12	16.15	71.36	18.35	71.68	15.88	72.97	17.21
	<i>Mixup</i> [140]	69.34	14.24	69.25	15.59	70.28	14.31	70.39	14.57
Mask Training	<i>AdaBoost</i>	69.52	17.23	68.66	19.46	70.12	16.57	70.24	18.35
	<i>EP</i> [110]	69.88	10.78	71.57	9.82	70.46	11.99	70.71	12.41
	<i>SNE</i>	71.28	4.64	73.32	5.48	72.20	6.57	74.56	6.55
	DRE (Ours)	71.68	3.48	74.04	2.82	72.00	1.52	73.72	1.08

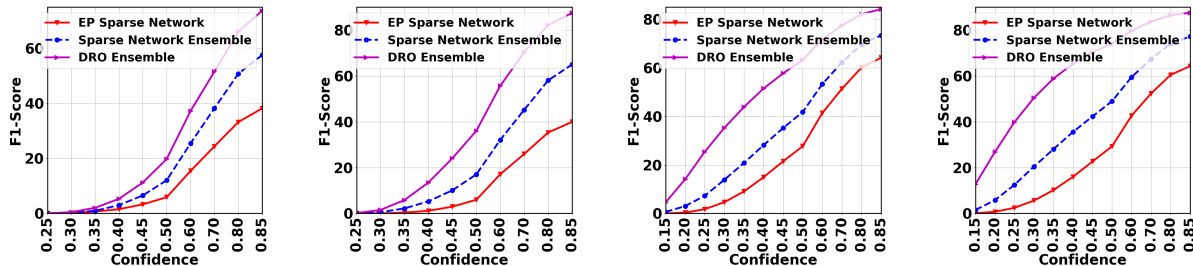
the pruning and regrowing sparse network training strategies. From Venkatesh et al. [140] we consider MixUp strategy as a comparison baseline as it does not require multi-step forward passes. A dense network is also included as a reference (denoted as *Dense*[†]). Furthermore, we report the performance obtained using the EP algorithm [110] on a single model with a given density. Finally, we also include the deep ensemble technique (*i.e.*, Sparse Network Ensemble (SNE)), where each base model is randomly initialized and independently trained. The approaches that require pre-training of a dense network are categorized under the *Dense Pre-training* category. Those performing sparse

Table 7.3: Accuracy and ECE performance on out-of-distribution datasets.

Training Type	Approach	Cifar10				Cifar100			
		ResNet50		ResNet101		ResNet101		ResNet152	
		<i>ACC</i>	<i>ECE</i>	<i>ACC</i>	<i>ECE</i>	<i>ACC</i>	<i>ECE</i>	<i>ACC</i>	<i>ECE</i>
	<i>Dense</i> [†]	79.65	19.63	79.65	19.63	54.75	35.32	54.75	35.32
Dense Pre-training	<i>L1 Pruning</i> [75]	77.34	17.95	76.39	17.89	52.06	31.45	51.67	30.98
	<i>LTH</i> [37]	75.85	17.88	76.15	17.62	50.79	31.23	51.35	30.56
	<i>DLTH</i> [6]	79.67	21.74	80.12	20.31	54.82	37.55	55.12	35.74
	<i>Mixup</i> [140]	76.35	17.74	76.88	17.55	51.36	31.12	51.92	30.35
Sparse Training	<i>CigL</i> [73]	70.80	21.04	69.84	21.42	49.42	25.86	51.49	24.13
	<i>Sup-ticket</i> [153]	72.89	17.80	73.01	18.82	48.80	24.99	48.81	25.62
Mask Training	<i>AdaBoost</i>	75.94	22.96	74.55	21.46	51.36	38.45	51.25	38.34
	<i>EP</i> [110]	77.58	17.82	77.73	17.46	52.18	30.60	52.14	29.48
	<i>SNE</i>	78.93	15.73	78.61	15.56	54.74	24.22	54.00	20.54
	<i>DRE (Ours)</i>	78.57	10.92	78.00	10.19	54.11	14.28	53.21	8.13

network training but actually updating the network parameters are grouped as *Sparse Training*. It should be noted that sparse training techniques still require iterative pruning and regrowing. Finally, techniques that attempt to search the best initialized sparse sub-network through mask update (*e.g.*, EP) are grouped as *Mask Training*.

General classification setting. In this setting, we consider clean Cifar10, Cifar100, and Tiny-ImageNet datasets. Tables 7.1, 7.2, and D.2 (in the Appendix) show the accuracy and calibration error for different models with density 9% and 15%. It should be noted that for the TinyImageNet dataset, we could not run the Sparse Training techniques due to the computation issue (*i.e.*, memory overflow). This may be because sparse training techniques require maintaining additional parameters for the pruning and regrowing strategy. In the Appendix, we have made a comparison of the proposed DRE with those baselines on a lower architecture size. There are three key observations we can infer from the experimental results. First, sparse networks are able to maintain or improve the generalization performance (in terms of accuracy) with better calibration, which can be seen by comparing dense network performance with the edge-popup algorithm. Second, the ensemble in general helps to further lower the calibration error (lower the better). For example, in all datasets, standard ensemble (SNE) consistently improves the EP model. Finally, the proposed DRE significantly improves the calibration performance by diversifying base learners and allow each sparse sub-network to focus on different parts of the training data. The strong calibration performance provides clear empirical evidence to justify our theoretical results.



(a) CIFAR10 ($\mathcal{K} = 15\%$) (b) CIFAR10 ($\mathcal{K} = 9\%$) (c) CIFAR100 ($\mathcal{K} = 15\%$) (d) CIFAR100 ($\mathcal{K} = 9\%$)

Figure 7.3: Open-set detection performance on different confidence thresholds.

Out-of-distribution classification setting. In this setting, we assess the effectiveness of the proposed techniques on out-of-distribution samples. Specifically, [49] provide the Cifar10-C and Cifar100-C validation datasets which are different than that of the original clean datasets. They apply different corruptions (such as blurring noise, and compression) to shift the distribution of the datasets. We assess those corrupted datasets using the models trained using the clean dataset. Table 7.3 shows the performance using different architectures. In this setting, we have not included DST Ensemble, because: (a) its accuracy is far below the SOTA performance, and (b) same training mechanism as that of the Sup-ticket, whose performance is reported. As shown, the proposed DRE provides much better calibration performance even with the out of distribution datasets.

Open-set detection setting. In this setting, we demonstrate the ability of our proposed DRO ensemble in detecting open-set samples. For this, we use the SVHN dataset as an open-set dataset. Specifically, if we have a better calibration, we would be able to better differentiate the open-set samples based on the confidence threshold. For this, we randomly consider 20% of the total testing in-distribution dataset as the open-set samples from the SVHN dataset. The reason for only choosing a subset of the dataset is to imitate the practical scenario where we have very few open-set samples compared to the close-set samples. We treat the open-set samples as the positive and in-distribution (close-set) ones as the negative. Since this is a binary detection problem, we compute the F-score [42] at various thresholds, which considers both precision and recall. Figure 7.3 shows the performance for the proposed technique along with comparative baselines. As shown, our proposed DRE (referred as DRO Ensemble) always stays on the top for various confidence thresholds which demonstrates that strong calibration performance can benefit DRE for open-set detection as compared to other baselines.

7.3.3 Additional Results, Ablation Study, and Qualitative Analysis

Limited by space, we have reported additional results in the Appendix. Specifically, we compare the proposed DRE with other standard calibration techniques commonly used in dense networks. In addition, we have performed an ablation study to investigate the impact of parameter η and different backbones (*i.e.*, ViT and WideResNet). We present a qualitative analysis to further justify the effectiveness of our proposed technique. Finally, we report the parameter size and inference speed (FLOPS) of DRE and compare it with existing baselines.

7.4 Conclusion

In this chapter, we proposed a novel DRO framework, called DRE, that achieves an ensemble of lottery tickets towards calibrated network sparsification. Specifically, with the guidance of uncertainty sets under the DRO framework, the proposed DRE aims to learn multiple diverse and complementary sparse sub-networks (tickets) where uncertainty sets encourage tickets to gradually capture different data distributions from easy to hard and naturally complement each other. We have theoretically justified the strong calibration performance by demonstrating how the proposed robust training process guarantees to lower the confidence of incorrect predictions. The extensive evaluation shows that the proposed DRE leads to significant calibration improvement without sacrificing the accuracy and burdening inference cost. Furthermore, experiments on OOD and Open-set datasets show its effectiveness in terms of generalization and novelty detection capability, respectively.

Chapter 8

Conclusion and Future Works

8.1 Conclusion

Despite having increased attention from diverse domains, anomaly detection is inherently challenging because of the rare and unbounded nature of anomalous activities. Multiple unsupervised and semi-supervised learning models have been used in the past but those techniques are sensitive to outliers (*i.e.*, normal samples that look different from other normal ones) and multi-modalities (*i.e.*, existence of multiple types of anomalies). As a result, the existing techniques yield much worse detection performance under these situations. Also, the imbalanced class distribution further poses a challenge as the model may be confused between normal samples from the minority class and true anomalies. Finally, the presence of the spurious correlation worsens the situation as the model is heavily biased toward majority groups while misidentifying the minority samples as anomalies. To tackle these challenges, we propose a novel Robust Weakly Supervised Learning (RWSL) framework that provides fundamental support for real-world anomaly detection using only weak and/or sparse learning signals. Our proposed RWSL framework constitutes three tightly coupled primary components. In the first component (Chapter 3), we integrate the Robust DRO with Bayesian learning, called Bayesian DRO, that achieves robust detection performance under the weak learning signals in the presence of outliers and multimodal anomalies. Next, in Chapter 4 we further augment Bayesian DRO with non-parametric submodular optimization and active instance sampling to improve both the reliability as well as accuracy. In the second component, we leverage evidential theory and fine-grained uncertainty formulation to tackle anomaly detection under the imbalanced class distribution of normal data samples. Specifically, in Chapter 5 we propose an adaptive Dis-

tributionally Robust Evidential Optimization (DREO) training process that boosts the anomaly detection performance by accurately differentiating the minority class samples and true anomalies using evidential uncertainty. Furthermore, in Chapter 6 we integrate evidential learning with a transformer architecture to result in Evidential Meta Transformer (MET) that provides reliable anomaly detection in the few-shot learning setting. Finally, in Chapter 7, we leverage the idea of the sparse network training along with DRO to propose the Distributionally Robust Ensemble (DRE) that helps to avoid: (1) learning from the spurious correlation, (2) overfitting resulting from memorization effect. As such, the resulting model is unbiased and has better calibration leading to better identifying minority class samples from true anomalies.

8.2 Future Works

Fairness AI. Bias can exist in various forms including biases including data bias, algorithmic bias etc. This results in a high prediction error in the minority groups. For example, in the natural language processing task, although SOTA speech recognizers achieve high overall accuracy, they fail to perform well on the minority group with slightly different accents compared to the majority one [2]. Furthermore, those approaches may even treat minority class samples as anomaly samples. As such, the minority group people may be discouraged from using the existing biased techniques resulting in a further reduction in the minority class data making the existing model even more biased as time progresses. Therefore, it is important to develop the fair model to work well on all groups considering the important applications such as face recognition [44], language identification [10], video captioning [134] etc. Most of the previous approaches, however, have the assumption that group-level information is available for each datapoint which may not be feasible in all practical scenarios. Previous literature has demonstrated that DRO has been a natural choice to deal with the class-imbalance problem. To tackle this, in the future, we plan to devise a technique that ensures high accuracy on minority groups without accessing the group assignment information. Our technique will leverage the DRO techniques to make the existing techniques unbiased by ensuring high accuracy across different groups without their information during the training process.

The planned DRO-enabled framework is motivated by the following facts

- DRO provides a natural way to learn an unbiased model by ensuring the minimization of the minority group data samples.

- As the DRO performs the optimization based on the loss, it does not require explicit group-level information for each data point.
- DRO is grounded on rich statistical properties that may be helpful in developing a tight bound to ensure the minimization of the worst group loss.

To accomplish this, in the future, we plan to conduct the research in the following two ways

1. Dataset Exploration and Design: To validate our approach, we need a dataset that requires explicit group-level information at least in the evaluation set. For this, we will be exploring the different datasets that may serve our purpose. This includes datasets from natural language processing, image classification, etc. Further, while selecting the dataset, we need to make sure that the dataset contains the imbalanced scenario in the training set which contains both minority group as well as majority group samples with the severe imbalanced scenario. Also, if needed, we will be designing a dataset to serve our purpose in order to validate the proposed technique.

2. Devise Methodology: Our proposed technique will leverage the DRO and may require the non-trivial extension of the existing DRO to work well in order to make the machine learning models unbiased. For this, we will be proposing the DRO-enabled technique along with the strong mathematical formulation which ensures the model capable of being able to minimize the worst group error without explicitly having the group-level information.

Trustworthy, sparse, and uncertainty aware Generative Language Model. There has been increasing attention on Generative Language Models especially in the large language models (LLMs) domain [84, 102, 145]. However, despite LLMs popularity, those models are extremely difficult to finetune because of the demand for extensive resources [55]. As such, the progress in generative models like LLM relies heavily on the contribution from the industry as the majority of academic researchers may not have extensive resources to conduct research on generative language models. Furthermore, due to extensive parameters, the LLMs may be overfitted leading to poorly calibrated, biased, and untrustworthy models. Additionally, due to the overfitting phenomenon, the model may confidently identify the anomalous sample as the known sample with high confidence and also minority group samples as unknowns. We plan to solve the aforementioned shortcomings in two ways.

In the first direction, inspired by Chapter 7, we plan to propose a way to find the sparse network *i.e.*, lottery ticket for the LLMs. The discovered lottery ticket will have very few parameters but will perform similar to that of LLMs. This technique of LLMs sparsification will have nu-

merous benefits. First, it makes the finetuning of the LLM process extremely efficient. As such, academic researchers can actively contribute (in addition to the industry) in the LLM domain to make remarkable progress in this field. Second, because of the low computational and memory cost, the sparse LLMs could be easily deployed in low-capacity devices. Third, as evidenced in Chapter 7, the overparameterized networks such as LLMs could be poorly calibrated because of the overfitting phenomenon resulting from the memorization effect. Furthermore, the models may confidently predict unknown samples as known because of the spurious correlation. To tackle this, the sparsification of LLMs could be an important step.

In the second direction, we seek to systematically quantify the uncertainty in the LLMs. As shown in Chapter 5 and Chapter 6, it is crucial to quantify the uncertainty in an effective way to better detect anomalous signals. Also, having an uncertainty score will facilitate the LLM to tell 'I do not know' instead of being incorrectly wrong. In case of an unknown situation, we can bring humans into the loop to further improve the LLM. Despite its need for dire attention, uncertainty quantification is poorly explored in the LLM domain. There has been very little research in this direction but is mostly focused on aggregating the token-level uncertainty while completely missing the overall semantics of the sentence [69, 93]. To tackle this, we aim to extend the evidential learning in the LLM by making it semantically meaningful. As such, the semantically meaningful uncertainty can be used to perform various tasks such as anomaly detection, model finetuning in limited data by the human in the loop, etc.

Chapter 9

List of Publications

9.1 Published

- [1] Hitesh Sapkota, Dingrong Wang, Zhiqiang Tao, and Qi Yu. Distributionally Robust Ensemble of Lottery Tickets Towards Calibrated Sparse Network Training. In *NeurIPS*, 2023.
- [2] Hitesh Sapkota and Qi Yu. Adaptive Robust Evidential Optimization For Open Set Detection from Imbalanced Data. In *ICLR*, 2023.
- [3] Hitesh Sapkota and Qi Yu. Balancing Bias and Variance for Active Weakly Supervised Learning. In *KDD*, 2022.
- [4] Hitesh Sapkota and Qi Yu. Bayesian Nonparametric Submodular Video Partition for Robust Anomaly Detection. In *CVPR*, 2022.
- [5] Dingrong Wang, Hitesh Sapkota, Xumin Liu, and Qi Yu. Deep Reinforced Attention Regression for Partial Sketch Based Image Retrieval. In *ICDM*, 2021.
- [6] Hitesh Sapkota, Yiming Ying, Feng Chen, and Qi Yu. Distributionally Robust Optimization for Deep Kernel Multiple Instance Learning. In *AISTATS*, 2021.
- [7] Moayad Alshangiti, Hitesh Sapkota, Pradeep K Murukannaiah, Xumin Liu, and Qi Yu. Why is developing machine learning applications challenging? a study on stack overflow posts. In *ESEM*, 2019.

9.2 Submitted (Preprints)

- [1] Hitesh Sapkota, Krishna Neupane, and Qi Yu. Meta Evidential Transformer for Few-Shot Open-Set Recognition. In *Submission*, 2024.
- [2] Dingrong Wang, Hitesh Sapkota, and Qi Yu. Adaptive Important Region Selection with Reinforced Hierarchical Searching for Dense Object Detection. In *Submission*, 2024.

Bibliography

- [1] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations*, 2023.
- [2] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jin Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Gregory Frederick Diamos, Erich Elsen, Jesse Engel, Linxi (Jim) Fan, Christopher Fougner, Awni Y. Hannun, Billy Jun, Tony Xiao Han, Patrick LeGresley, Xiangang Li, Libby Lin, Sharan Narang, A. Ng, Sherjil Ozair, Ryan J. Prenger, Sheng Qian, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Anuroop Sriram, Chong-Jun Wang, Yi Wang, Zhiqian Wang, Bo Xiao, Yan Xie, Dani Yogatama, Junni Zhan, and Zhenyao Zhu. Deep speech 2 : End-to-end speech recognition in english and mandarin. In *ICML*, 2016.
- [3] S. Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2002.
- [4] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2002.
- [5] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv*, 2019.
- [6] Yue Bai, Huan Wang, ZHIQIANG TAO, Kumpeng Li, and Yun Fu. Dual lottery ticket hypothesis. In *International Conference on Learning Representations*, 2022.
- [7] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1563–1572, 2016.
- [8] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016.

- [9] José H. Blanchet, Yang Kang, Fan Zhang, and Zhangyi Hu. A distributionally robust boosting algorithm. *2019 Winter Simulation Conference (WSC)*, pages 3728–3739, 2019.
- [10] Su Lin Blodgett, Lisa Green, and Brendan T. O’Connor. Demographic dialectal variation in social media: A case study of african-american english. In *EMNLP*, 2016.
- [11] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Praseon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars, 2016.
- [12] Marc-André Carbonneau, V. Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognit.*, 77:329–353, 2018.
- [13] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 4724–4733, 2017.
- [14] Arantxa Casanova, Pedro H. O. Pinheiro, Negar Rostamzadeh, and Christopher Joseph Pal. Reinforced active learning for image segmentation. *ArXiv*, abs/2002.06583, 2020.
- [15] Hakan Cevikalp and Hasan Serhan Yavuz. Fast and accurate face recognition with image sets. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1564–1572, 2017.
- [16] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2006.
- [17] Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *ArXiv*, abs/2006.09239, 2020.
- [18] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, June 2002.
- [19] Tianlong Chen, Zhenyu Zhang, Jun Wu, Randy Huang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Can you win everything with a lottery ticket? *Transactions on Machine Learning Research*, 2022.
- [20] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, pages 3449–3456, 2011.

- [21] J. Cunningham, K. Shenoy, and M. Sahani. Fast gaussian process methods for point process intensity estimation. In *ICML '08*, 2008.
- [22] Arthur P Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968.
- [23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [25] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.
- [26] A. Dezfouli and Edwin V. Bonilla. Scalable inference for gaussian process models with black-box likelihoods. In *NIPS*, 2015.
- [27] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2):31–71, January 1997.
- [28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.
- [29] John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *J. Mach. Learn. Res.*, 20(1):2450–2504, January 2019.
- [30] John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.
- [31] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015.
- [32] Yanbo Fan, Siwei Lyu, Yiming Ying, and Baogang Hu. Learning with average top-k loss. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 497–505. Curran Associates, Inc., 2017.

- [33] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.
- [34] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.
- [35] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. *arXiv preprint arXiv:1806.02817*, 2018.
- [36] Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. A sticky hdp-hmm with application to speaker diarization. *The Annals of Applied Statistics*, 5:1020–1056, 2011.
- [37] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. 2018.
- [38] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *COLT 1997*, 1997.
- [39] Yarín Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv*, 2015.
- [40] Yarín Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2015.
- [41] ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*, 2017.
- [42] Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In David E. Losada and Juan M. Fernández-Luna, editors, *Advances in Information Retrieval*, pages 345–359, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.
- [43] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*, 2018.
- [44] Patrick Grother, George W. Quinn, and P. Jonathon Phillips. Report on the evaluation of 2d still-image face recognition algorithms. 2011.

- [45] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org, 2017.
- [46] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 733–742, 2016.
- [47] Manuel Hausmann, Fred A. Hamprecht, and Melih Kandemir. Variational bayesian multiple instance learning with gaussian processes. In *CVPR*, pages 810–819, 2017.
- [48] Chengkun He, Jie Shao, and Jiayu Sun. An anomaly-introduced learning method for abnormal event detection. *Multimedia Tools Appl.*, 77(22):29573–29588, November 2018.
- [49] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.
- [50] Yen-Chi Hsu, Cheng-Yao Hong, Ming-Sui Lee, and Tyng-Luh Liu. Query-driven multi-instance learning. In *AAAI*, 2020.
- [51] Shiyuan Huang, Jiawei Ma, Guangxing Han, and Shih-Fu Chang. Task-adaptive negative envision for few-shot open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7171–7180, 2022.
- [52] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, 2018.
- [53] Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M. Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, Chrisantha Fernando, and Koray Kavukcuoglu. Population based training of neural networks. *ArXiv*, abs/1711.09846, 2017.
- [54] Lalit P. Jain, Walter J. Scheirer, and Terrance E. Boult. Multi-class open set recognition using probability of inclusion. In *ECCV*, 2014.
- [55] Ajay Jaiswal, Shiwei Liu, Tianlong Chen, Ying Ding, and Zhangyang Wang. Instant soup: Cheap pruning ensembles in a single pass can draw lottery tickets from large models, 2023.
- [56] Minki Jeong, Seokeon Choi, and Changick Kim. Few-shot open-set recognition by transformation consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12566–12575, 2021.

- [57] Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association : JAMIA*, 19:263 – 274, 2011.
- [58] Audun Jøsang. *Subjective logic*. Springer, 2016.
- [59] Pedro Ribeiro Mendes Júnior, Roberto Medeiros de Souza, Rafael de Oliveira Werneck, Bernardo V. Stein, Daniel V. Pazinato, Waldir R. de Almeida, Otávio Augusto Bizetto Penatti, Ricardo da Silva Torres, and Anderson Rocha. Nearest neighbors distance ratio open-set classifier. *Machine Learning*, 106:359–386, 2016.
- [60] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv*, 2015.
- [61] R. Keys. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(6):1153–1160, 1981.
- [62] Minyoung Kim and F. Torre. Gaussian processes multiple instance learning. In *ICML*, 2010.
- [63] Minyoung Kim and F. Torre. Gaussian processes multiple instance learning. In *ICML*, 2010.
- [64] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *arXiv*, 2019.
- [65] Anna-Kathrin Kopetzki, Bertrand Charpentier, Daniel Zügner, Sandhya Giri, and Stephan Günnemann. Evaluating robustness of predictive uncertainty estimation: Are dirichlet-based models reliable? In *ICML*, 2021.
- [66] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, pages 1446–1453, 2009.
- [67] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.
- [68] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [69] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023.
- [70] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350:1332 – 1338, 2015.

- [71] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, page 6405–6416, Red Hook, NY, USA, 2017. Curran Associates Inc.
- [72] Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015.
- [73] Bowen Lei, Ruqi Zhang, Dongkuan Xu, and Bani Mallick. Calibrating the rigged lottery: Making all tickets reliable. In *The Eleventh International Conference on Learning Representations, 2023*.
- [74] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1):1–14, 2017.
- [75] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets, 2016.
- [76] Nannan Li, Xinyu Wu, Huiwen Guo, Dan Xu, Yongsheng Ou, and Yen-Lun Chen. Anomaly detection in video surveillance via gaussian process. *Int. J. Pattern Recognit. Artif. Intell.*, 29:1555011:1–1555011:25, 2015.
- [77] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. *CoRR*, abs/2007.01162, 2020.
- [78] Weixin Li and N. Vasconcelos. Multiple instance learning for soft bags via top instances. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4277–4285, 2015.
- [79] Hui Lin and Jeff Bilmes. How to select a good training-data subset for transcription: Submodular active selection for sequences. Technical report, WASHINGTON UNIV SEATTLE DEPT OF ELECTRICAL ENGINEERING, 2009.
- [80] Bo Liu, Hao Kang, Haoxiang Li, Gang Hua, and Nuno Vasconcelos. Few-shot open-set recognition using meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2020.
- [81] Shiwei Liu, Tianlong Chen, Zahra Atashgahi, Xiaohan Chen, Ghada Sokar, Elena Mocanu, Mykola Pechenizkiy, Zhangyang Wang, and Decebal Constantin Mocanu. Deep ensembling with no overhead for either training or testing: The all-round blessings of dynamic sparsity, 2022.

- [82] W. Liu, W. Luo, D. Lian, and S. Gao. Future frame prediction for anomaly detection - a new baseline. In *CVPR*, pages 6536–6545, 2018.
- [83] Wen Liu, Weixin Luo, Zhengxin Li, Peilin Zhao, and Shenghua Gao. Margin learning embedded prediction for video anomaly detection with a few anomalies. In *IJCAI*, 2019.
- [84] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.
- [85] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2532–2541, 2019.
- [86] Jose Llamas. Architectural heritage elements image dataset, 2017.
- [87] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *2013 IEEE International Conference on Computer Vision*, pages 2720–2727, 2013.
- [88] David G. Luenberger. *Optimization by Vector Space Methods*. John Wiley and Sons, Inc., USA, 1st edition, 1997.
- [89] Tiange Luo, Aoxue Li, Tao Xiang, Weiran Huang, and Liwei Wang. Few-shot learning with global class representations, 2019.
- [90] W. Luo, W. Liu, and S. Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *ICCV*, pages 341–349, 2017.
- [91] Dougal Maclaurin, David Duvenaud, and Ryan P. Adams. Gradient-based hyperparameter optimization through reversible learning. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML’15*, page 2113–2122. JMLR.org, 2015.
- [92] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS’18*, page 7047–7058, Red Hook, NY, USA, 2018. Curran Associates Inc.
- [93] Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction, 2021.
- [94] David Mease and Abraham J. Wyner. Evidence contrary to the statistical view of boosting. *J. Mach. Learn. Res.*, 9:131–156, 2008.

- [95] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.
- [96] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *International Conference on Machine Learning*, pages 2554–2563. PMLR, 2017.
- [97] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI’15, page 2901–2907. AAAI Press, 2015.
- [98] Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.
- [99] Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [100] Yuval Netzer, Tao Wang, Adam Coates, A. Bissacco, Bo Wu, and A. Ng. Reading digits in natural images with unsupervised feature learning. 2011.
- [101] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D. Cubuk, and Ian J. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS’18, page 3239–3250, Red Hook, NY, USA, 2018. Curran Associates Inc.
- [102] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.
- [103] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift, 2019.
- [104] Fabian Pedregosa. Hyperparameter optimization with approximate gradient. *ArXiv*, abs/1602.02355, 2016.

- [105] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions, 2017.
- [106] John Platt and Nikos Karampatziakis. Probabilistic outputs for svms and comparisons to regularized likelihood methods. 2007.
- [107] Qi Qi, Yan Yan, Zixuan Wu, X. Wang, and Tianbao Yang. A simple and effective framework for pairwise deep metric learning. *ArXiv*, abs/1912.11194, 2019.
- [108] Qi Qi, Yan Yan, Zixuan Wu, X. Wang, and Tianbao Yang. A simple and effective framework for pairwise deep metric learning. *ArXiv*, abs/1912.11194, 2020.
- [109] Joaquin Quiñero Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *J. Mach. Learn. Res.*, 6:1939–1959, December 2005.
- [110] Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What’s hidden in a randomly weighted neural network?, 2019.
- [111] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.
- [112] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, H. Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. *ArXiv*, abs/1803.00676, 2018.
- [113] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018.
- [114] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations, 2020.
- [115] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR, 2016.
- [116] Hitesh Sapkota, Yiming Ying, Feng Chen, and Qi Yu. Distributionally robust optimization for deep kernel multiple instance learning. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2188–2196. PMLR, 13–15 Apr 2021.
- [117] Robert E Schapire. Explaining adaboost. In *Empirical inference*, pages 37–52. Springer, 2013.

- [118] Walter J. Scheirer, Lalit P. Jain, and Terrance E. Boult. Probability models for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:2317–2324, 2014.
- [119] Walter J. Scheirer, Anderson Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1757–1772, 2013.
- [120] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv*, 2017.
- [121] M. Sensoy, Melih Kandemir, and Lance M. Kaplan. Evidential deep learning to quantify classification uncertainty. In *NeurIPS*, 2018.
- [122] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31, 2018.
- [123] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty, 2018.
- [124] Burr Settles. Active learning literature survey. 2009.
- [125] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. *NIPS*, 20:1289–1296, 2007.
- [126] Weishi Shi, Xujiang Zhao, Feng Chen, and Qi Yu. Multifaceted uncertainty estimation for label-efficient deep learning. *Advances in Neural Information Processing Systems*, 33:17247–17257, 2020.
- [127] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.
- [128] Wanhua Su, Yan Yuan, and Mu Zhu. A relationship between the average precision and the area under the roc curve. In *ICTIR*, pages 349–352, 2015.
- [129] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, 2018.
- [130] Xin Sun, Zhenning Yang, Chi Zhang, Guohao Peng, and Keck-Voon Ling. Conditional gaussian distribution learning for open set recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13477–13486, 2020.

- [131] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.
- [132] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.
- [133] Zhiqiang Tao, Yaliang Li, Bolin Ding, Ce Zhang, Jingren Zhou, and Yun Fu. Learning to mutate with hypergradient guided population. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS’20*, Red Hook, NY, USA, 2020. Curran Associates Inc.
- [134] Rachael Tatman. Gender and dialect bias in youtube’s automatic captions. In *EthNLP@EACL*, 2017.
- [135] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581, 2006.
- [136] Kai Tian, Shuigeng Zhou, Jianping Fan, and Jihong Guan. Learning competitive and discriminative reconstructions for anomaly detection. In *AAAI*, pages 5167–5174. AAAI Press, 2019.
- [137] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4975–4986, October 2021.
- [138] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV ’15, page 4489–4497, USA, 2015. IEEE Computer Society.
- [139] R. Turner. Statistical models for natural sounds. 2010.
- [140] Bindya Venkatesh, Jayaraman J. Thiagarajan, Kowshik Thopalli, and Prasanna Sattigeri. Calibrate and prune: Improving reliability of lottery tickets through prediction calibration, 2020.
- [141] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29:3630–3638, 2016.

- [142] Hung Vu, T. Nguyen, Trung Le, Wei Luo, and Dinh Q. Phung. Robust anomaly detection in videos using multilevel representations. In *AAAI*, 2019.
- [143] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016.
- [144] X. Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.
- [145] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.
- [146] X. Wei, J. Wu, and Z. Zhou. Scalable multi-instance learning. In *2014 IEEE International Conference on Data Mining (ICDM)*, pages 1037–1042, 2014.
- [147] A. Wilson, Zhiting Hu, R. Salakhutdinov, and E. Xing. Stochastic variational deep kernel learning. *ArXiv*, abs/1611.00336, 2016.
- [148] Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4697–4708, 2020.
- [149] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378, 2016.
- [150] Andrew Gordon Wilson and Hannes Nickisch. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML’15, page 1775–1784. JMLR.org, 2015.
- [151] Huan Yang, Baoyuan Wang, Stephen Lin, David Wipf, Minyi Guo, and Baining Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *ICCV, ICCV ’15*, page 4633–4641, USA, 2015. IEEE Computer Society.
- [152] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8808–8817, 2020.

- [153] Lu Yin, Vlado Menkovski, Meng Fang, Tianjin Huang, Yulong Pei, Mykola Pechenizkiy, Decibal Constantin Mocanu, and Shiwei Liu. Superposing many tickets into one: A performance booster for sparse neural network training, 2022.
- [154] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Nae-mura. Classification-reconstruction learning for open-set recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4011–4020, 2019.
- [155] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Nae-mura. Classification-reconstruction learning for open-set recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4011–4020, 2019.
- [156] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. *CoRR*, abs/2104.02324, 2021.
- [157] He Zhang and Vishal M. Patel. Sparse representation-based open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1690–1696, 2017.
- [158] Jize Zhang, Bhavya Kailkhura, and T. Yong-Jin Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning, 2020.
- [159] J. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1237–1246, 2019.
- [160] Jia-Xing Zhong, Nannan Li, Weijie Kong, S. Liu, Thomas H. Li, and G. Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1237–1246, 2019.
- [161] Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask, 2019.
- [162] Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. Multi-instance learning by treating instances as non-i.i.d. samples. In *ICML, ICML '09*, page 1249–1256, New York, NY, USA, 2009. Association for Computing Machinery.
- [163] Zhi-Hua Zhou and Min-Ling Zhang. Neural networks for multi-instance learning. *Proceedings of the International Conference on Intelligent Information Technology*, 11 2002.

- [164] Dixian Zhu, Zhe Li, Xiaoyu Wang, Boqing Gong, and Tianbao Yang. A robust zero-sum game framework for pool-based active learning. In *AISTATS*, 2019.
- [165] Dixian Zhu, Zhe Li, Xiaoyu Wang, Boqing Gong, and Tianbao Yang. A robust zero-sum game framework for pool-based active learning. In *AISTATS*, volume 89, pages 517–526, 2019.

Appendices

Appendix A

In this section, we will mainly provide the additional information required to support the Sections that are present in the Chapter 3. Mostly we provide the proofs for the proposed theorems used in the respective section.

A.1 Distributionally Robust Optimization for Deep Kernel Multiple Instance Learning

This section will provide the additional information to support Chapter 3. Specifically, first we provide the detailed proof for Lemma 3.1 and Lemma 3.2 in Chapter 3. Next, we provide parameter update mechanism of the parameters introduced in the same chapter.

A. Detailed Proofs: In this section, we show the detailed steps of the proofs for Lemma 3.1 and Lemma 3.2.

Proof of Lemma 3.1 : We have the bag level likelihood from Eqs. (3.1) and (3.2) as follows:

$$p(y_b | \mathbf{f}_b, \mathbf{z}_b) = \prod_{i=1}^n \left\{ \frac{1}{1 + \exp(-t_b f_{bi})} \right\}^{z_{bi}} \quad (\text{A.1})$$

$$p(\mathbf{z}_b | \boldsymbol{\pi}_b) = \prod_{i=1}^n \pi_{bi}^{z_{bi}}, \pi_{bi} \geq 0, \sum_i \pi_{bi} = 1 \quad (\text{A.2})$$

Marginalizing over \mathbf{z}_b , we have the following expression:

$$p(y_b|\mathbf{f}_b, \boldsymbol{\pi}_b) = \sum_{i=1}^n p(\mathbf{y}_b|\mathbf{f}_b, z_{bi} = 1)p(z_{bi} = 1|\pi_b) = \sum_{i=1}^n \pi_{bi} \frac{1}{1 + \exp(-t_b f_{bi})}$$

Let's denote $p(f_{bi}) = \frac{1}{1 + \exp(-t_b f_{bi})}$, which yields,

$$p(y_b|\mathbf{f}_b, \boldsymbol{\pi}_b) = \sum_{i=1}^n \pi_{bi} p(f_{bi})$$

In Lemma 3.1, we maximize the above likelihood over $\boldsymbol{\pi}_b$ with respect to the following uncertainty set:

$$\mathcal{P}_{\boldsymbol{\pi}_b, n}^{max} := \{\boldsymbol{\pi}_b \in R^n : \boldsymbol{\pi}_b^T \mathbf{1} = 1, 0 \leq \boldsymbol{\pi}_b\}$$

The resulting optimization becomes:

$$\max_{\boldsymbol{\pi}_b} \sum_{i=1}^n \pi_{bi} p(f_{bi}) \quad \text{s.t.} \quad \sum_{i=1}^n \pi_{bi} = 1, \pi_{bi} \geq 0, \forall i \in [1, n]$$

Adding the Lagrange multipliers $u_i \geq 0$ and λ , we get:

$$L(\boldsymbol{\pi}_b, \mathbf{u}, \lambda) = \sum_{i=1}^n [\pi_{bi} p(f_{bi}) + u_i \pi_{bi}] + \lambda \left[\sum_{i=1}^n \pi_{bi} - 1 \right]$$

Taking derivative with respect to π_{bi} and setting to zero yields

$$p(f_{bi}) + u_i + \lambda = 0 \tag{A.3}$$

The corresponding KKT conditions are:

$$u_i \geq 0, \quad u_i \pi_{bi} = 0, \quad \forall i \in [1, n] \tag{A.4}$$

Considering $j = \arg \max_{i \in b} p(f_{bi})$, we have the following condition

$$p(f_{bj}) + u_j + \lambda = 0 \tag{A.5}$$

Combining Eqs. (A.3) and (A.5) results in:

$$p(f_{bj}) + u_j = p(f_{bi}) + u_i, \quad \text{s.t.} \quad j = \arg \max_i p(f_{bi}), \forall i \in [1, n], i \neq j \tag{A.6}$$

Since $p(f_{bj}) > p(f_{bi})$, we have $u_j < u_i$. As $u_i \geq 0, \forall i \in [1, n]$, we have $u_i \neq 0, \forall i \neq j$. By leveraging the complementary slackness condition, we have

$$u_i \neq 0, \quad \pi_{bi} = 0, \quad \forall i \in [i, n], i \neq j \tag{A.7}$$

Further using summation constraint, i.e., $\sum_{i=1}^n \pi_{bi} = 1$, we have the following

$$\pi_{bi} = \begin{cases} 1, & \text{if } p(f_{bi}) = \max_i p(f_{bi}) \\ 0, & \text{otherwise} \end{cases} \quad (\text{A.8})$$

In case of equality condition with $p(f_{bj}) = p(f_{bi})$ with $i \neq j$, we randomly select one and assign $\pi_{bi} = 1$ for one instance whereas 0 for others.

Proof of Lemma 3.2: We have the following marginalized likelihood function (from Lemma 3.1 proof):

$$p(y_b | \mathbf{f}_b \pi_b) = \sum_{i=1}^n \pi_{bi} p(f_{bi}) \quad \forall i \in [1, n]$$

In Lemma 3.2, we maximize the above likelihood function with respect to the following uncertainty set:

$$\mathcal{P}_{\pi_b, n}^{\text{top-}k} := \{ \pi_b \in R^n : \pi_b^T \mathbf{1} = 1, 0 \leq \pi_{bi} \leq \frac{1}{k} \} \quad (\text{A.9})$$

The resulting optimization becomes

$$\max_{\pi_b} \sum_{i=1}^n \pi_{bi} p(f_{bi}), \quad \text{s.t.} \quad \sum_{i=1}^n \pi_{bi} = 1, \quad 0 \leq \pi_{bi} \leq \frac{1}{k}, \quad \forall i \in [1, n]$$

Adding the Lagrange multipliers $u_i \geq 0$, $v_i \geq 0$, and λ we get:

$$L(\pi_b, \mathbf{u}, \mathbf{v}, \lambda) = \sum_{i=1}^n \left[\pi_{bi} p(f_{bi}) + u_{bi} \pi_{bi} + v_{bi} \left(\frac{1}{k} - \pi_{bi} \right) \right] + \lambda \left(\sum_{i=1}^n \pi_{bi} - 1 \right)$$

Taking the derivative with respect to π_{bi} yields the following:

$$p(f_{bi}) + u_{bi} - v_{bi} + \lambda = 0$$

Considering $p(f_{b[1]}) > p(f_{b[2]}) > \dots > p(f_{b[n]})$ with $p(f_{b[i]})$ be the i^{th} highest probability score, we have the following conditions:

$$\begin{aligned} & p(f_{b[1]}) + u_{b[1]} - v_{b[1]} + \lambda \\ & = p(f_{b[2]}) + u_{b[2]} - v_{b[2]} + \lambda \\ & \dots \\ & = p(f_{b[n]}) + u_{b[n]} - v_{b[n]} + \lambda \end{aligned}$$

Removing the $p(f_{b[i]})$ and λ terms, we get the following inequalities:

$$u_{b[1]} - v_{b[1]} < u_{b[2]} - v_{b[2]} < \dots < u_{b[k]} - v_{b[k]} < \dots < u_{b[n]} - v_{b[n]} \quad (\text{A.10})$$

Consider the following KKT conditions $\forall i \in [1, n]$

$$\sum_{i=1}^n \pi_{bi} = 1 \quad (\text{A.11})$$

$$u_{b[i]} \pi_{bi} = 0 \quad (\text{A.12})$$

$$v_{b[i]} \left(\frac{1}{k} - \pi_{bi} \right) = 0 \quad (\text{A.13})$$

$$u_{b[i]} \geq 0 \quad (\text{A.14})$$

$$v_{b[i]} \geq 0 \quad (\text{A.15})$$

Case 1: Assume $\pi_{b[1]} = 0$, so $v_{b[1]} = 0$ according to (A.13). This implies $u_{b[1]} < u_{b[2]} - v_{b[2]}$. Using KKT condition (A.14), we can write the following:

$$u_{b[2]} - v_{b[2]} > 0 \Rightarrow u_{b[2]} > v_{b[2]} \Rightarrow u_{b[2]} > 0 \text{ (according to KKT condition (A.15))}$$

This means, to satisfy the constraint $u_{b[2]} \pi_{b[2]} = 0$ we need to have the following:

$$\pi_{b[2]} = 0$$

Again $\pi_{b[2]}=0$ makes $\pi_{b[3]} = 0$ and so on. As $\pi_{b[i]} = 0 \forall i \in [1, n]$, and therefore violating the summation constraint $\sum_{i=1}^n \pi_{bi} = 1$. **Therefore, $\pi_{b[1]}$ can not be 0.**

Case 2: Assume $0 < \pi_{b[1]} < \frac{1}{k}$, then $v_{b[1]} = 0$, $u_{b[1]} = 0$, We have the following expression:

$$u_{b[2]} > v_{b[2]}$$

Using KKT condition $v_{b[2]} \geq 0$ we can write:

$$u_{b[2]} > 0$$

Now again using KKT condition $u_{b[2]} \pi_{b[2]} = 0$, we write:

$$\pi_{b[2]} = 0$$

Using Case 1, once $\pi_{b[2]}$ becomes 0 all of the proceeding values also become zero:

$$\pi_{b[3]} = 0, \dots, \pi_{b[n]} = 0$$

This again violates the summation constraint $\sum_{i=1}^n \pi_{bi} = 1$. **Therefore, $\pi_{b[1]}$ can not be less than $\frac{1}{k}$.** This leads to the following conclusion:

$$\pi_{b[1]} = \frac{1}{k}$$

The process of having $\pi_{b[i]} = \frac{1}{k}$ continues until $\pi_{b[k]} = \frac{1}{k}$. As long as we reach to the k^{th} highest element, $\sum_{i=1}^n \pi_{bi} = 1$. This implies the following:

$$\pi_{b[i]} = 0, \quad \forall i > k$$

In conclusion, we can write the following:

$$\pi_{bi} = \begin{cases} \frac{1}{k}, & \text{if } p(f_{bi}) \geq p(f_{b[k]}) \\ 0, & \text{otherwise} \end{cases} \quad (\text{A.16})$$

which proves Lemma 3.2.

B. Parameter Update: We provide the parameter update procedures for parameters introduced in Chapter 3.1. To update the parameters, we take the derivative of the lower bound $L(q)$ with respect to the parameter we want to update. We then, use SGD to update the parameter with a given learning rate. Therefore, in this section, we show the computation of derivative of $L(q)$, w.r.t. each parameter.

Derivative of $L(q)$ w.r.t. Base Kernel Hyperparameters We have following expression for a lower bound

$$L(q) \approx \mathbb{E}_{q(\mathbf{U})q(\mathbf{Z})} [\log p(\mathbf{y}|\mathbf{F}, \mathbf{Z})] - KL[q(\mathbf{U})||p(\mathbf{U})] - KL[q(\mathbf{Z})||p(\mathbf{Z})]$$

In the above equation, the base kernel hyperparameters θ are only involved in the second term i.e., $KL[q(\mathbf{U})||p(\mathbf{U})]$. Before taking its derivative let us simplify it further,

$$KL[q(\mathbf{U})||p(\mathbf{U})] = \frac{1}{2} \{ \log |K| - \log |\mathbf{S}| - D + \text{tr}(K^{-1}\mathbf{S}) + \boldsymbol{\mu}^T K^{-1} \boldsymbol{\mu} \}$$

Taking derivative w.r.t. θ , it gives:

$$\frac{\partial L(q)}{\partial \theta} = -\frac{\partial KL(q(\mathbf{U})||p(\mathbf{U}))}{\partial \theta} = -\frac{1}{2} \left\{ \text{tr}(K^{-1} \frac{\partial K}{\partial \theta}) - \text{tr}(K^{-1} \frac{\partial K}{\partial \theta} K^{-1} \mathbf{S}) - \boldsymbol{\mu}^T K^{-1} \frac{\partial K}{\partial \theta} K^{-1} \boldsymbol{\mu} \right\}$$

Depending on type of a given kernel, we can find $\frac{\partial K}{\partial \theta}$ in the above equation. The corresponding matrix inversions and traces can be computed efficiently by using Kronecker product decomposition [147].

Derivative of $L(q)$ w.r.t. variational parameters of distribution $q(\mathbf{U})$ Both variational parameters $\boldsymbol{\mu}$ and \mathbf{L} depend on the first bag-level likelihood term and $KL[q(\mathbf{U})||p(\mathbf{U})]$. The first term in $L(q)$ is given as:

$$\log p(y_b|\mathbf{f}_b, \mathbf{z}_b) = \sum_{i=1}^n z_{bi} \log \frac{1}{1 + \exp(-t_b f_{bi})} = \sum_{i=1}^n z_{bi} p(f_{bi})$$

Taking the derivative of i^{th} instance with respect to (p, q) -th element of $L_d^{(j)}$, i.e. λ , where $j \in [1, J]$ indicates the j^{th} base GP:

$$\nabla_{\lambda} \log p(y_b|\mathbf{f}_b, \mathbf{z}_b) = z_{bi} \frac{\partial \log p(f_{bi})}{\partial f_{bi}} \frac{\partial f_{bi}}{\partial \lambda}$$

In the above equation f_{bi} is defined as:

$$f_{bi} = \sum_{j=1}^J A_j f_{bi}^j \quad (\text{A.17})$$

As we are taking with respect to the j^{th} GP, it means:

$$\frac{\partial f_{bi}}{\partial \lambda} = A_j \frac{\partial f_{bi}^j}{\partial \lambda}$$

As we know we have the following relationship $f = M(\boldsymbol{\mu} + \mathbf{L}\boldsymbol{\epsilon})$. Using this relationship we get:

$$\frac{\partial f_{bi}^j}{\partial \lambda} = A_j M_i^j \nabla_{\lambda} L^{(j)} \boldsymbol{\epsilon}$$

where M_i^j is the i^{th} row of the j^{th} GP. Now the first term can be computed as:

$$\frac{\partial \log p(f_{bi})}{\partial f_{bi}} = \frac{t_b}{1 + \exp(t_b f_{bi})}$$

Combining both and considering all elements present in a bag b , we have the following update rule:

$$\nabla_{\lambda} \log p(y_b|\mathbf{f}_b, \mathbf{z}_b) = \mathbb{E}_{p(\boldsymbol{\epsilon})q(\mathbf{Z})} \left[z_{bi} \frac{t_b A_j M_i^j \nabla_{\lambda} L^{(j)} \boldsymbol{\epsilon}}{1 + \exp(t_b f_{bi})} \right]$$

We can write down the derivatives w.r.t. the whole matrix $\mathbf{L}^{(j)}$ which is efficient for computing:

$$\nabla_{\mathbf{L}^{(j)}} \log p(y_b | f_b, z_b) = \mathbb{E}_{p(\boldsymbol{\epsilon})q(\mathbf{Z})} \left[\begin{array}{c} t_b A_j \left(\boldsymbol{\epsilon} M_i^j \right)^T \\ z_{bi} \frac{1}{1 + \exp(t_b f_{bi})} \end{array} \right]$$

Now the derivative of $KL[q(\mathbf{U})||p(\mathbf{U})]$ can be written as:

$$\nabla_{\lambda} KL[q(\mathbf{U})||p(\mathbf{U})] = -\frac{1}{2} \frac{\partial[-\log |S| + \text{tr}(K^{-1}S)]}{\partial \lambda}$$

We can efficiently compute the above by using Kronecker decomposition matrix. The derivatives w.r.t. the variational mean μ can be computed similarly as that of \mathbf{L} .

Derivative w.r.t. other parameters The mixing weight only depends on the likelihood function. Therefore, it can be easily computed as:

$$\nabla_{A_j} \log p(y_b | \mathbf{f}_b, \mathbf{z}_b) = \mathbb{E}_{p(\boldsymbol{\epsilon})q(\mathbf{Z})} \left[\frac{z_{bi} t_b f_{bi}^j}{1 + \exp(t_b f_{bi})} \right] \quad (\text{A.18})$$

To update the neural network parameters \mathbf{w} , we take the derivative of $L(q)$ with respect to \mathbf{w} . As the network parameters are related to $L(q)$ through the kernel matrix K , our update procedure is given as:

$$\frac{\partial L(q)}{\partial \mathbf{w}} = \frac{\partial L(q)}{\partial K} \frac{\partial K}{\partial h(\mathbf{x}, \mathbf{w})} \frac{\partial h(\mathbf{x}, \mathbf{w})}{\partial \mathbf{w}}$$

In the above expression, the term $\frac{\partial K}{\partial h(\mathbf{x}, \mathbf{w})}$ is the implicit derivative of the deep kernel with respect to $h(\mathbf{x}, \mathbf{w})$, holding base kernel hyperparameters θ fixed. The derivatives with respect to network weight variables $\frac{\partial h(\mathbf{x}, \mathbf{w})}{\partial \mathbf{w}}$ are computed using the standard back-propagation techniques.

A.2 Bayesian Nonparametric Submodular Video Partition for Robust Anomaly Detection

In this section, we provide an detailed proof of the Theorem 3.3 introduced in Chapter 3.2.

A. Proof of Theorem 3.3:

In this section, we provide the detailed proof of Theorem 3.3. We first show that the representative set based MIL loss given by (3.32) is equivalent to the submodularity diversified MIL loss given by

Equation (3.28) with a specific λ to balance the MIL loss and the diversity of the set. We then show that greedy algorithm to locate the κ representative segments provides a κ -constrained greedy approximation to the maximization of the submodular set function $F(\mathcal{C})$ with the solution to be no worse than $(1 - e^{-1})$ of the optimal solution.

Proof of representative set based MIL loss in (3.32) is a special case of the submodular diversified MIL loss in (3.28) We first present a lemma, which is used in the proof.

Lemma A.1. *Assume that $\widetilde{\mathcal{C}}^+$ with size κ is a solution that maximizes $F(\mathcal{C})$ in (3.27). Then, $\widetilde{\mathcal{C}}^+$ should contain one segment from each mixture component (i.e., sub-scene).*

Proof. The lemma can be proved by following the definition of the BN-SVP induced pairwise similarity between segments given by (3.26) and then use proof by contradiction. Assume that at least two segments, say $\mathbf{x}_i^{(t)}, \mathbf{x}_j^{(t)}$, are chosen from the same component t . Then, there will be at least one component, say t' , where no segments are chosen by $\widetilde{\mathcal{C}}^+$. Given the definition of $F(\mathcal{C})$ in (3.27), for each segment in t , either $\mathbf{x}_i^{(t)}$ or $\mathbf{x}_j^{(t)}$ could be used to compute the pairwise similarity based on their closeness to that segment. Since the cohesiveness of each component is guaranteed through the BN-SVP process, both $\mathbf{x}_i^{(t)}$ and $\mathbf{x}_j^{(t)}$ should be close to the mean of their assigned Gaussian component $\mathcal{N}(\mathbf{x}_t, \Sigma_t)$ to ensure a high likelihood optimized by HDP-HMM. Due to triangle inequality, $\mathbf{x}_i^{(t)}$ and $\mathbf{x}_j^{(t)}$ should be close to each other. As a result, we can assume that $\mathbf{x}_i^{(t)}$ is always chosen to evaluate the pairwise similarity $S_{i,p}$ with each segment $\mathbf{x}_p^{(t)}$ in component t . Next, we replace $\mathbf{x}_j^{(t)}$ with another segment $\mathbf{x}_j^{(t')}$ from component t' to construct another solution set $\overline{\mathcal{C}}^+$. Since $\mathbf{x}_j^{(t')}$ has positive similarity with each segment in t' and the pairwise similarity between $\mathbf{x}_j^{(t)}$ and all segments in t' is all zero, we have $F(\overline{\mathcal{C}}^+) > F(\widetilde{\mathcal{C}}^+)$, which contradicts the assumption that $\widetilde{\mathcal{C}}^+$ maximizes $F(\mathcal{C})$. \square

Since the representative set $\widehat{\mathcal{C}}^+$ is constructed by choosing one segment from each mixture component, it satisfies the necessary condition to be an optimizer of $F(\mathcal{C})$ specified in the above lemma. However, choosing a set of segments with the maximum diversity is not the primary goal and the overall objective function (3.28) includes both the MIL loss and the diversity, which are balanced through λ . Due to the lack of instance-level labels, choosing a λ that optimally balances the MIL loss and the set diversity is challenging. We argue that construction $\widehat{\mathcal{C}}^+$ essentially offers alternative way to set a specific λ to balance these two terms. First, since the constraint $|\mathcal{C}^+| \leq \kappa$ allows the set to contain less than κ segment, $\widehat{\mathcal{C}}^+$ excludes those segments with low prediction scores. This can be viewed as setting a λ to increase $-F(\mathcal{C}^+)$ while decreasing the MIL loss $L(\mathcal{C}^+)$. Similarly,

instead of choosing the instance with the largest pairwise similarity with all other segments in the same component, we choose a segment with the highest prediction score. Again, this can be viewed as further reducing the λ to give more preference to the MIL loss as such segments can further reduce the training MIL loss. Thus, instead of directly setting λ , which is highly challenging, $\widehat{\mathcal{C}}^+$ is constructed by leveraging both the mixture component assignments and the prediction scores of the segments. This is equivalent to implicitly setting a λ to balance the MIL loss and the diversity of the representative set $\widehat{\mathcal{C}}^+$, which completes the proof of the equivalence of these two objective functions.

Proof of the optimality of the greedy algorithm We first reformulate (3.28) as a minimization problem $\min_{\mathbf{w}} g(\mathbf{w})$ with $g(\mathbf{w})$ defined as

$$g(\mathbf{w}) \stackrel{\partial}{=} \min_{\mathcal{C}^+ \subseteq \mathcal{B}_{pos}, |\mathcal{C}^+| \leq \kappa} L(\mathcal{B}_{pos}, \mathcal{B}_{neg}) - \lambda F(\mathcal{C}^+) \quad (\text{A.19})$$

The above optimization involves finding a subset $\mathcal{C}^+ \subseteq \mathcal{B}_{pos}$ that maximizes $F(\mathcal{C}^+)$. This requires enumerating over all $\binom{n}{\kappa}$ possible subsets, which is expensive when there are large number of segments in a given video. Defining the discrete objective function $G_{\mathbf{w}}$ where

$$G_{\mathbf{w}}(\mathcal{C}^+) \stackrel{\partial}{=} L(\mathcal{B}_{pos}, \mathcal{B}_{neg}) - \lambda F(\mathcal{C}^+) \quad (\text{A.20})$$

Since $-G_{\mathbf{w}}(\mathcal{C}^+)$ is monotone non-decreasing submodular, a fast greedy procedure can be used to approximately optimize $G_{\mathbf{w}}(\mathcal{C}^+)$. A typical greedy procedure involves evaluating the similarity between each pair of segments in a video and then choose the segments with the largest overall similarity with the all other segments. We make two important adjustments of this standard greedy process. First, our non-parametric HDP-HMM process follows the clustering based heuristic (Lin and Bilmes 2018) by choosing one segment from each cluster, which avoids evaluating each candidate segment in the video. Different from (Lin and Bilmes 2018), which chooses the data point that is closest to the cluster centroid, we choose the one with the highest output score. Second, our similarity evaluation takes a linear complexity with respect to the bag size by leveraging the temporal neighborhood of the segments.

Appendix B

This section provides the additional information to support Chapter 5. First, we present the detailed description of the training process obtained through the bilevel optimization. Next, we provide the proofs of main theoretical results.

A. Training Through Bi-level Optimization: Our training involves a bi-level optimization, where we jointly optimize the network parameter Θ along with the MSF parameters \mathbf{W} . Algorithm 2 shows the overall training process based on the population based optimization. We randomly initialize the MSF parameters \mathbf{W}_p and network parameters Θ_p from the corresponding spaces \mathcal{H} and Θ_{param} respectively shown in Line 3. We perform this initialization for P different models. Next, in each epoch we independently optimize P models using the proposed objective function defined in Eq. (5.6). After s epochs, we evaluate the accuracy of each model by using ‘eval’ as the evaluation metric in the validation set. It should be noted that in our case, we used closed set classification performance (MAP) as ‘eval’ metric. We identify \hat{P} (with $\hat{P} < P$) worst performing models and replace their model parameters by the randomly selected model parameters from set of b highest accurate models. This process is known as exploitation and is demonstrated in Line 12. MSF parameters for those worst performing model can be obtained either through random selection from the original space \mathcal{H} or through small perturbation of the \mathbf{W} of the model whose parameter is copied. This process is called exploration as we are searching for the new MSF parameters and is shown in Line 13. The best performing model parameters and accuracy are stored in the Θ^* and acc^* respectively. Finally, the best model Θ^* is returned as the optimal model for the testing.

The optimization specified in (5.6) involves an inequality constraint, which incurs a higher computational overhead. Therefore, in our actual optimization process, we consider a regularized version

Algorithm 2: Multi-Scheduler Learning Process

Input: $\mathcal{H}, P, s, \text{eval}, \hat{P}, T$

```

1 Initialize: epoch = 0,  $\Theta^* = \text{None}$ ,  $acc^* = \text{None}$ 
2 for  $p \in [P]$  do
3    $\Theta_p, \mathbf{W}_p \leftarrow \text{initialize}(\Theta_{param}, \mathcal{H})$ 
4 while epoch  $\neq T$  do
5    $\Theta_p \leftarrow \text{optimize}(\Theta_p | \mathbf{W}_p), p \in [P]$ 
6   if epoch %  $s = 0$  then
7      $acc_p \rightarrow \text{eval}(\Theta_p, \mathbf{W}_p), p \in [P]$ 
8      $sorted\_idx \leftarrow \text{arg sortDesc}\{acc_p\}_{p=1}^P$ 
9      $bottom\_idx \leftarrow sorted\_idx[: -\hat{P}]$ 
10     $top\_idx \leftarrow sorted\_idx[: \hat{P}]$ 
11    for  $idx \in bottom\_idx$  do
12       $\Theta_{idx}, j \leftarrow \text{uniform}(\{\Theta_j\}_j^{top\_idx})$ 
13       $\mathbf{W}_{idx} \leftarrow \text{explore}(\mathcal{H}, \mathbf{W}_j)$ 
14     $best\_model\_idx \leftarrow top\_idx[0]$ 
15    if  $\Theta^*$  not  $\text{None}$  then
16      if  $acc^* < acc_{best\_model\_idx}$  then
17         $acc^* = acc_{best\_model\_idx}$ 
18         $\Theta^* = \Theta_{best\_model\_idx}$ 
19    else
20       $\Theta^* = \Theta_{best\_model\_idx}$ 
21    epoch  $\leftarrow$  epoch + 1
22 return  $\Theta^*$  with the highest acc

```

of the DREO loss as follows:

$$\mathcal{L}^{\text{DREL}} = \max_{\mathbf{p} \geq 0, \mathbf{p}^\top \mathbf{1} = 1} \sum_{n=1}^N p_n l_n^t - \lambda D_f \left(\mathbf{p} \parallel \frac{\mathbf{1}}{N} \right) \quad (\text{B.1})$$

Solving the above maximization problem leads to a closed form solution for \mathbf{p}^* as shown by the following lemma. It should be noted that the role of the λ is exactly opposite as that of the η . Specifically, we start from a high λ so that the model gives equal emphasis to all data samples. Next, in each step we decrease λ using the following Equation

$$\lambda_t = \lambda_{t-1} \text{MSF}(\mathbf{w}, \boldsymbol{\beta}, t, T) \quad (\text{B.2})$$

Decreasing λ helps the model focus on the difficult samples as training progresses.

Lemma B.1. *Assuming that D_f is the KL divergence, then solving (B.1) leads to the following solution*

$$\mathcal{L}^{\text{DREO}} = \sum_{n=1}^N p_n^* l_n^t \quad (\text{B.3})$$

where p_n^* is given by

$$p_n^* = \frac{\exp\left(\frac{l_n^t}{\lambda}\right)}{\sum_{j=1}^N \exp\left(\frac{l_j^t}{\lambda}\right)} \quad (\text{B.4})$$

Proof: The the Lagrangian of the regularized loss in (B.1) is

$$\mathcal{L}^{\text{DREL}}(\Theta, v, \lambda) = \sum_{n=1}^N p_n l_n^t - \lambda \left(\sum_{n=1}^N p_n \log p_n + \log N \right) + v \left[\left(\sum_{n=1}^N p_n \right) - 1 \right] \quad (\text{B.5})$$

where v is the Lagrangian multiplier. Taking the derivative with respect to p_n and setting it to 0:

$$l_n^t - \lambda \log p_n - \lambda + v = 0 \quad (\text{B.6})$$

Simplifying above equation, we get p_n as

$$p_n = \exp\left(\frac{l_n^t + v}{\lambda} - 1\right) \quad (\text{B.7})$$

Using the summation constraint over p_n i.e., $\sum_{n=1}^N p_n = 1$, it leads to following

$$\sum_{n=1}^N \exp\left(\frac{l_n^t + v}{\lambda} - 1\right) = 1 \quad (\text{B.8})$$

Solving the above equation we get the expression for v as follows

$$v = \lambda \log \left(\frac{1}{\sum_{n=1}^N \exp \left(\frac{l_n^t}{\lambda} - 1 \right)} \right) \quad (\text{B.9})$$

Substituting the v value into (B.7) gives

$$p_n = \frac{\exp \left(\frac{l_n^t}{\lambda} \right)}{\sum_{n=1}^N \exp \left(\frac{l_n^t}{\lambda} \right)} \quad (\text{B.10})$$

The concludes the proof.

B. Proofs of Theoretical Results:

In this section, we present the detailed proofs for Lemmas 5.1, 5.2, and Theorem 5.3.

Proof of Lemma 5.1: By setting $\eta \rightarrow 0$, we have $D_f(\mathbf{p} \parallel \frac{\mathbf{1}}{N}) \rightarrow 0$. This implies that \mathbf{p} is uniform with each element as $\frac{1}{N}$. As a result, the optimization problem becomes

$$\mathcal{L}^{DREL}(\Theta) = \frac{1}{N} \sum_{n=1}^N l_n^{EL}(\Theta) \quad (\text{B.11})$$

Proof of Lemma 5.2: With $\eta \rightarrow \infty$, the uncertainty set defined in (5.5) reduces to the following

$$\mathcal{P}^{DRO} := \left\{ \mathbf{p} \in \mathbb{R}^N : \mathbf{p}^\top \mathbf{1} = 1, \mathbf{p} \geq 0 \right\} \quad (\text{B.12})$$

Now, the corresponding Lagrangian form of (5.6) becomes

$$\mathcal{L}^{DREL}(\Theta, \mathbf{u}, \lambda) = \sum_{n=1}^N (p_n l_n^{EL}(\Theta) + u_n p_n) + \mu \left(\sum_{n=1}^N p_n - 1 \right) \quad (\text{B.13})$$

where u_n and μ are Lagrangian multipliers. Taking gradient with respect to p_n and setting it zero, we get

$$l_n^{EL}(\Theta) + u_n + \mu = 0 \quad (\text{B.14})$$

Let $k = \arg \max_n l_n^{EL}(\Theta)$ be the index of data sample with the maximum loss (and assuming it is unique). Then, the following holds true

$$u_k < u_n; \quad \forall n \in [1, N], n \neq k \quad (\text{B.15})$$

This consequently leads to $u_n > 0, \forall n \in [1, N], n \neq k$. Due to the KKT conditions,

$$u_n p_n = 0; \quad \forall n \in [1, N] \quad (\text{B.16})$$

we have $p_n = 0, \forall n \in [1, N], n \neq k$. By using the following constraint

$$\sum_{n=1}^N p_n = 1 \quad (\text{B.17})$$

we have the following conclusion

$$p_n = \begin{cases} 1, & \text{if } n = k \\ 0, & \text{otherwise} \end{cases} \quad (\text{B.18})$$

This means our optimization reduces to the following

$$\mathcal{L}^{DREL}(\Theta) = \max_n l_n^{EL}(\Theta) \quad (\text{B.19})$$

which proves the lemma.

Proof of Theorem 5.3. AdaBoost can be achieved through alternative optimization between a classification function f and the worst case probability solution [38]. To show equivalence with the proposed DREO, our proof includes three steps: (i) a specially designed deep neural network (DNN) architecture and a loss function adapted to match the learning process of AdaBoost, (ii) projected functional sub-gradient descent to optimize the classification function f , and (iii) optimizing the worst case probability solution.

Step 1: A specially designed DNN. Let $\phi(\mathbf{x}) \in \mathcal{R}^M$ denote a M -dimensional feature vector learned using a DNN. By applying a fully connected linear layer with a weight matrix $W \in \mathbb{R}^{K \times M}$ on top of the feature vector, we obtain a set of K (discriminative) functions: $\mathbf{f} = (f_1, \dots, f_K)^\top = W\phi(\mathbf{x})$. Then, the final output of the DNN is obtained by aggregating these K functions, leading to $f = \boldsymbol{\sigma}^\top \mathbf{f}$, where $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_K)^\top$. As a result of this design, the final function output by the DNN can be regarded as lying in the linear span of a set of functions $\mathcal{F} = \{f_1, \dots, f_K\}$, given by

$$LS(\mathcal{F}) = \left\{ f : f = \sum_{k=1}^K \sigma_k f_k, 1 \leq k \leq K, \sigma_k \in (-\infty, \infty) \right\} \quad (\text{B.20})$$

Training of DREO involves alternating between re-weighting using the worst case probability distribution and updating the prediction function f . Next, we prove that given the specially designed DNN, we can exactly optimize the classification function f by keeping the worst case probability fixed and vice versa.

Step 2: Optimizing the classification function f under the worst case probability. We first formulate the distributional robust evidential loss, which is given by

$$\mathcal{L}^{DREL} = \max_{\mathbf{p} \in \mathcal{P}^{DRO}} \sum_{n=1}^N p_n \mathcal{L}_n(f) \quad (\text{B.21})$$

where $\mathcal{L}_n(f)$ is the loss associated with the datasample \mathbf{x}_n . Then, the optimal f^* can be obtained by minimizing the distributional robust loss:

$$f^* = \min_{f \in LS(\mathcal{F})} \mathcal{L}^{DREL} \quad (\text{B.22})$$

This optimization involves a nonconvex loss \mathcal{L}^{DREL} . To ensure the convergence of f to a stationary point, we adapt the Probabilistic Gradient Estimator (PAGE) technique [42] to the DRO setting (shown in Algorithm 3) which ensures the convergence in $\mathcal{O}(b + \frac{b}{\epsilon^2})$ steps with b being the batch size. Please refer to Theorem B.3 further details.

To show that an optimal f^* can be achieved, we first verify that the specially designed DNN and the loss function as described above meet a number key conditions as specified by [9]: (i) the loss functional \mathcal{L}^{DREL} is L -smooth, (ii) for two different functions $f^1, f^2 \in LS(\mathcal{F})$, $f^1(\phi(\mathbf{x}_n)) \neq f^2(\phi(\mathbf{x}_n))$, and (iii) $LS(\mathcal{F})$ has a finite dimensional basis. First, (i) is true because \mathcal{L}^{DREL} is the convex combination of the losses $\mathcal{L}_n(f)$. As each individual loss involves the ReLU term with ReLU added in the output of DNN (to ensure non-negativity of the evidence), the resulting convex combination may not be smooth. Therefore, we use the SoftPlus which is smooth function to approximate the ReLU. The the convex combination of SoftPlus results in the function \mathcal{L}^{DREL} to be L -smooth. Second, the rich and high dimensional input data (*i.e.*, diverse images) and the feature encoding through the deep architecture of the DNN ensures (ii) is true. Last, since the dimensionality of the weight matrix W is $K \times M$, it implies that the dimensionality of the basis of $LS(\mathcal{F})$ is bounded by K , so (iii) holds true.

The smoothness of \mathcal{L}^{DREL} ensures that a stationary solution is achieved within the $\mathcal{O}(b + \frac{b}{\epsilon^2})$ gradient steps. This allows us to have a guaranteed stationary solution with $\mathbb{E}[\|\nabla \mathcal{L}^{DREL}\|] \leq \epsilon$ in a non-convex optimization setting. Furthermore, since \mathcal{L}^{DREL} is a functional on f , the next two conditions ensure that the functional gradient exists and can be evaluated [9]. During the optimization process, we need to make sure that the trajectory of the functional gradient lies in the space $LS(\mathcal{F})$, which can be achieved through functional gradient projection.

Step 3: Optimizing the worst case probability solution. Let f_t denote the optimal classification function for the current iteration t . Next, we continue to optimize the worst case probability solution. The following lemma shows that such an optimal solution exists.

Lemma B.2. *Assuming that $\mathcal{L}_n(f_t)$ has a finite exponential moment with $\alpha \geq 0$ being sufficiently large and*

$$\eta_t = \beta^* \psi'(\beta^*) - \psi(\beta^*) \quad (\text{B.23})$$

the worst case probability is given as

$$p_n^* = \frac{\exp\left(\frac{\mathcal{L}_n(f_t)}{\alpha}\right)}{\sum_{j=1}^N \exp\left(\frac{\mathcal{L}_j(f_t)}{\alpha}\right)} \quad (\text{B.24})$$

where $\beta^* = \frac{1}{\alpha^*}$, $\alpha^* \geq 0$ be the optimal α , and $\psi(\beta) = \log\left[\frac{\sum_{n=1}^N \exp(\beta \mathcal{L}_n(f_t))}{N}\right]$.

Proof. Taking the derivative of the Lagrangian for the optimization problem given in (5.13) with respect to p_n leads to

$$\mathcal{L}_n(f_t) - \alpha \log p_n - \alpha + u_n = 0 \quad (\text{B.25})$$

where u_n is the Lagrangian multiplier for the constraint $\mathbf{p} \geq 0$ and α is the Lagrange multiplier for the DRO constraint with the size defined by η_t . Simplification of the above expression yields

$$\log p_n = \frac{\mathcal{L}_n(f_t)}{\alpha} + \frac{u_n - \alpha}{\alpha} \quad (\text{B.26})$$

For some λ' with $p_n = \lambda' \exp\left(\frac{\mathcal{L}_n(f_t)}{\alpha}\right)$, a candidate solution is

$$p_n^* = \frac{\exp\left(\frac{\mathcal{L}_n(f_t)}{\alpha}\right)}{\sum_{j=1}^N \exp\left(\frac{\mathcal{L}_j(f_t)}{\alpha}\right)} \quad (\text{B.27})$$

The above equation involves the expression in terms of the Lagrangian multiplier. By leveraging the sufficiency result presented in Chapter 8 Theorem 1 of [88], we can find the relationship between the multiplier and our constraint parameter η_t . As such, our optimal solution can be expressed in terms of original constraint. Suppose that we can find $\alpha^* \geq 0$ and $\mathbf{p}^* \in \mathcal{P}^{DRO}$ such that \mathbf{p}^* maximizes (5.13) for $\alpha = \alpha^*$ and $\sum_{n=1}^N p_n^* \log p_n^* = \eta_t$ with the optimal solution defined in (B.27). Considering this, we have the following

$$\eta_t = \sum_{n=1}^N p_n^* \log p_n^* = \sum_{n=1}^N p_n^* \frac{\mathcal{L}_n(f_t)}{\alpha^*} - \log\left(\sum_{j=1}^N \exp\left(\frac{\mathcal{L}_j(f_t)}{\alpha^*}\right)\right) = \beta^* \psi'(\beta^*) - \psi(\beta^*) \quad (\text{B.28})$$

where we define $\beta^* = \frac{1}{\alpha^*}$ and $\psi(\beta) = \log \sum_{n=1}^N \exp(\beta \mathcal{L}_n(f_t))$. This allows us to express the Lagrangian multiplier using η_t . Next, we verify that there exists a unique solution defined in (B.27)

by leveraging the convexity of the exponential function. Specifically, substituting (B.27) in (5.13), we get the following

$$\sum_{n=1}^N p_n^* \mathcal{L}_n(f_t) - \alpha \left(\sum_{n=1}^N p_n^* \log p_n^* \right) = \alpha \log \sum_{n=1}^N \exp \left(\frac{\mathcal{L}_n(f_t)}{\alpha} \right) \quad (\text{B.29})$$

If we could show the following inequality holds true

$$\alpha \log \sum_{n=1}^N \exp \left(\frac{\mathcal{L}_n(f_t)}{\alpha} \right) \geq \sum_{n=1}^N p_n \mathcal{L}_n(f_t) - \alpha \sum_{n=1}^N p_n \log p_n \quad (\text{B.30})$$

then we can claim that the above candidate solution is the optimal solution. Rearranging the terms, we get the following

$$\sum_{n=1}^N \exp \left(\frac{\mathcal{L}_n(f_t)}{\alpha} \right) \geq \exp \sum_{n=1}^N \left(\frac{p_n \mathcal{L}_n(f_t)}{\alpha} - p_n \log p_n \right) \quad (\text{B.31})$$

This can be shown as

$$\sum_{n=1}^N \exp \left(\frac{\mathcal{L}_n(f_t)}{\alpha} \right) = \sum_{n=1}^N p_n p_n^{-1} \exp \left(\frac{\mathcal{L}_n(f_t)}{\alpha} \right) = \sum_{n=1}^N p_n \exp \left(\frac{\mathcal{L}_n(f_t)}{\alpha} - \log p_n \right)$$

Now applying Jensen inequality to the exponential function $\psi \left(\frac{\sum x_i}{n} \right) \leq \frac{\sum \psi(x_i)}{n}$, we have the following

$$\sum_{n=1}^N p_n \exp \left(\frac{\mathcal{L}_n(f_t)}{\alpha} - \log p_n \right) \geq \exp \left(\sum_{n=1}^N \frac{p_n \mathcal{L}_n(f_t)}{\alpha} - p_n \log p_n \right) \quad (\text{B.32})$$

This completes the proof of the lemma. \square

Theorem B.3. *Suppose that \mathcal{L}^{DREL} holds the L -smoothness criteria with following inequality*

$$\|\nabla \mathcal{L}^{DREL}(f^1) - \nabla \mathcal{L}^{DREL}(f^2)\| \leq L \|f^1 - f^2\| \quad (\text{B.33})$$

Then choosing a learning rate $\gamma \leq \frac{1}{L(1+\sqrt{\frac{1-p}{pb'}})}$ with minibatch size $b = n$, secondary minibatch size $b' < b$, the number of iterations required to be performed by our algorithm for finding ϵ -approximate solution i.e., $\mathbb{E}[\|\nabla \mathcal{L}^{DREL}(\hat{f}_T)\| \leq \epsilon]$ can be bounded by the following:

$$T = \frac{2\Delta_0 L}{\epsilon^2} \left(1 + \sqrt{\frac{1-p}{pb'}} \right) \quad (\text{B.34})$$

Further the gradient complexity in terms of number of gradient steps is given as

$$N_{grad} = b + \frac{2\Delta_0 L}{\epsilon^2} \left(1 + \sqrt{\frac{1-p}{pb'}} \right) (pb + (1-p)b') \quad (\text{B.35})$$

Before giving the formal proof, we first show two lemmas that are used during the proof.

Lemma B.4. *The L -smoothness condition given by Eq. (B.33), leads to the following inequality*

$$\mathcal{L}^{DREL}(f^2) \leq \mathcal{L}^{DREL}(f^1) + \langle \nabla \mathcal{L}^{DREL}(f^1), f^2 - f^1 \rangle + \frac{L}{2} \|f^2 - f^1\|^2, \forall f^1, f^2 \in \mathcal{R}^m. \quad (\text{B.36})$$

where $\langle a, b \rangle = a^T b$, and $\|\cdot\|$ is the Euclidean norm.

Proof of Lemma B.4. For the completeness the proof of the above Lemma is as follow.

$$\begin{aligned} & \mathcal{L}^{DREL}(f^2) \\ & \leq \mathcal{L}^{DREL}(f^1) + \int_0^1 \langle \nabla \mathcal{L}^{DREL}(f^1) + \tau(f^2 - f^1), f^2 - f^1 \rangle d\tau \\ & = \mathcal{L}^{DREL}(f^1) + \langle \nabla \mathcal{L}^{DREL}(f^1), f^2 - f^1 \rangle \\ & + \int_0^1 \langle \nabla \mathcal{L}^{DREL}(f^1 + \tau(f^2 - f^1)) - \nabla \mathcal{L}^{DREL}(f^1), f^2 - f^1 \rangle d\tau \end{aligned}$$

Cauchy-Schwarz inequality $\langle u, v \rangle \leq \|u\| \|v\|$ leads to the following

$$\begin{aligned} & \mathcal{L}^{DREL}(f^2) \\ & \leq \mathcal{L}^{DREL}(f^1) + \langle \nabla \mathcal{L}^{DREL}(f^1), f^2 - f^1 \rangle \\ & + \int_0^1 \|\nabla \mathcal{L}^{DREL}(f^1 + \tau(f^2 - f^1)) - \nabla \mathcal{L}^{DREL}(f^1)\| \|f^2 - f^1\| d\tau \end{aligned}$$

Now lets use the L -smoothness assumption from Eq. (B.33), we have

$$\begin{aligned} & \mathcal{L}^{DREL}(f^2) \\ & \leq \mathcal{L}^{DREL}(f^1) + \langle \nabla \mathcal{L}^{DREL}(f^1), f^2 - f^1 \rangle + \int_0^1 L\tau \|f^2 - f^1\|^2 d\tau \\ & = \mathcal{L}^{DREL}(f^1) + \langle \nabla \mathcal{L}^{DREL}(f^1), f^2 - f^1 \rangle + \frac{L}{2} \|f^2 - f^1\|^2 \end{aligned}$$

Now, we provide another important Lemma required to prove the above Theorem based on Lemma B.4

Lemma B.5. *Considering L -smoothness assumption in Eq. (B.33), and let $f_{t+1} := f_t - \gamma g_t$. Then for any $g_t \in \mathcal{R}^M$ and $\gamma > 0$ we have the following*

$$\begin{aligned} & \mathcal{L}^{DREL}(f_{t+1}) \\ & \leq \mathcal{L}^{DREL}(f_t) - \frac{\gamma}{2} \|\nabla \mathcal{L}^{DREL}(f_t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2} \right) \|f_{t+1} - f_t\|^2 + \frac{\gamma}{2} \|g_t - \nabla \mathcal{L}^{DREL}(f_t)\|^2 \quad (\text{B.37}) \end{aligned}$$

Proof: Let $\bar{f}_{t+1} := f_t - \gamma \nabla \mathcal{L}^{DREL}(f_t)$. Then using L -smoothness of \mathcal{L}^{DREL} , we have the following

$$\begin{aligned}
& \mathcal{L}^{DREL}(f_{t+1}) \\
& \leq \mathcal{L}^{DREL}(f_t) + \langle \nabla \mathcal{L}^{DREL}(f_t), f_{t+1} - f_t \rangle + \frac{L}{2} \|f_{t+1} - f_t\|^2 \\
& = \mathcal{L}^{DREL}(f_t) + \langle \nabla \mathcal{L}^{DREL}(f_t) - g_t, f_{t+1} - f_t \rangle + \langle g_t, f_{t+1} - f_t \rangle + \frac{L}{2} \|f_{t+1} - f_t\|^2 \\
& = \mathcal{L}^{DREL}(f_t) + \langle \nabla \mathcal{L}^{DREL}(f_t) - g_t, -\gamma g_t \rangle - \left(\frac{1}{\gamma} - \frac{L}{2} \right) \|f_{t+1} - f_t\|^2 \\
& = \mathcal{L}^{DREL}(f_t) + \gamma \|\nabla \mathcal{L}^{DREL}(f_t) - g_t\|^2 - \gamma \langle \nabla \mathcal{L}^{DREL}(f_t) - g_t, \nabla \mathcal{L}^{DREL}(f_t) \rangle \\
& \quad - \left(\frac{1}{\gamma} - \frac{L}{2} \right) \|f_{t+1} - f_t\|^2 \\
& = \mathcal{L}^{DREL}(f_t) + \gamma \|\nabla \mathcal{L}^{DREL}(f_t) - g_t\|^2 - \frac{1}{\gamma} \langle f_{t+1} - \bar{f}_{t+1}, f_t - \bar{f}_{t+1} \rangle \\
& \quad - \left(\frac{1}{\gamma} - \frac{L}{2} \right) \|f_{t+1} - f_t\|^2 \\
& = \mathcal{L}^{DREL}(f_t) + \gamma \|\nabla \mathcal{L}^{DREL}(f_t) - g_t\|^2 - \left(\frac{1}{\gamma} - \frac{L}{2} \right) \|f_{t+1} - f_t\|^2 \\
& \quad - \frac{1}{2\gamma} (\|f_{t+1} - \bar{f}_{t+1}\|^2 + \|f_t - \bar{f}_{t+1}\|^2 - \|f_{t+1} - f_t\|^2) \\
& = \mathcal{L}^{DREL}(f_t) + \gamma \|\nabla \mathcal{L}^{DREL}(f_t) - g_t\|^2 - \left(\frac{1}{\gamma} - \frac{L}{2} \right) \|f_{t+1} - f_t\|^2 \\
& \quad - \frac{1}{2\gamma} (\|\gamma^2 \|\nabla \mathcal{L}^{DREL}(f_t) - g_t\|^2 + \gamma^2 \|\nabla \mathcal{L}^{DREL}(f_t)\|^2 - \|f_{t+1} - f_t\|^2) \\
& = \mathcal{L}^{DREL}(f_t) - \frac{\gamma}{2} \|\nabla \mathcal{L}^{DREL}(f_t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2} \right) \|f_{t+1} - f_t\|^2 + \frac{\gamma}{2} \|g_t - \nabla \mathcal{L}^{DREL}(f_t)\|^2
\end{aligned}$$

This completes the proof. The last term is the variance and it can be bounded using the following lemma.

Lemma B.6. *Suppose that the smoothness assumption in Eq. (B.33) holds. If the gradient estimator g_{t+1} is defined in Algorithm 3 Line 13, then we have the following*

$$\mathbb{E}[\|g_{t+1} - \nabla \mathcal{L}^{DREL}(f_{t+1})\|^2] \leq (1 - p_t) \|g_t - \nabla \mathcal{L}^{DREL}(f_t)\|^2 + \frac{(1 - p_t)L^2}{b'} \|f_{t+1} - f_t\|^2 \quad (\text{B.38})$$

Proof: According to Algorithm 3, we have the following

$$g_{t+1} = \begin{cases} \frac{1}{b} \sum_{n \in B} a_n(f_{t+1}) \nabla \mathcal{L}_n(f_{t+1}) & \text{with probability } p_t \\ g_t + \frac{1}{b'} \sum_{n \in B'} (a_n(f_{t+1}) \nabla \mathcal{L}_n(f_{t+1}) - a_n(f_t) \nabla \mathcal{L}_n(f_t)), & \text{with probability } 1 - p_t \end{cases} \quad (\text{B.39})$$

Using this the left hand side of the above lemma can be written as

$$\begin{aligned}
& \mathbb{E}[\|g_{t+1} - \nabla \mathcal{L}^{DREL}(f_{t+1})\|^2] \\
&= p_t \mathbb{E} \left[\left\| \frac{1}{b} \sum_{n \in B} a_n(f_{t+1}) \nabla \mathcal{L}_n(f_{t+1}) - \nabla \mathcal{L}^{DREL}(f_{t+1}) \right\|^2 \right] \\
&\quad + (1 - p_t) \mathbb{E} \left[\left\| g_t + \frac{1}{b'} \sum_{n \in B'} (a_n(f_{t+1}) \nabla \mathcal{L}_n(f_{t+1}) - a_n(f_t) \nabla \mathcal{L}_n(f_t)) - \nabla \mathcal{L}^{DREL}(f_{t+1}) \right\|^2 \right] \\
&= (1 - p_t) \mathbb{E} \left[\left\| g_t + \frac{1}{b'} \sum_{n \in B'} (a_n(f_{t+1}) \nabla \mathcal{L}_n(f_{t+1}) - a_n(f_t) \nabla \mathcal{L}_n(f_t)) - \nabla \mathcal{L}^{DREL}(f_{t+1}) \right\|^2 \right] \\
&= (1 - p_t) \mathbb{E} \left[\left\| g_t - \nabla \mathcal{L}^{DREL}(f_t) + \frac{1}{b'} \sum_{n \in B'} (a_n(f_{t+1}) \nabla \mathcal{L}_n(f_{t+1}) - a_n(f_t) \nabla \mathcal{L}_n(f_t)) \right\|^2 \right] \\
&\quad + (1 - p_t) \mathbb{E} [\|-\nabla \mathcal{L}^{DREL}(f_{t+1}) + \nabla \mathcal{L}^{DREL}(f_t)\|^2] \\
&= (1 - p_t) \mathbb{E} \left[\left\| \frac{1}{b'} \sum_{n \in B'} (a_n(f_{t+1}) \nabla \mathcal{L}_n(f_{t+1}) - a_n(f_t) \nabla \mathcal{L}_n(f_t)) - \nabla \mathcal{L}^{DREL}(f_{t+1}) + \nabla \mathcal{L}^{DREL}(f_t) \right\|^2 \right] \\
&\quad + (1 - p_t) \|g_t - \nabla \mathcal{L}^{DREL}(f_t)\|^2 \\
&= \frac{1 - p_t}{b'^2} \mathbb{E} \left[\sum_{n \in B'} \|(a_n(f_{t+1}) \nabla \mathcal{L}_n(f_{t+1}) - a_n(f_t) \nabla \mathcal{L}_n(f_t))\|^2 \right] \\
&\quad - \frac{1 - p_t}{b'^2} \mathbb{E} [\|(\nabla \mathcal{L}^{DREL}(f_{t+1}) - \nabla \mathcal{L}^{DREL}(f_t))\|^2] + (1 - p_t) \|g_t - \nabla \mathcal{L}^{DREL}(f_t)\|^2 \\
&\leq \frac{(1 - p_t)L^2}{b'} \|\mathcal{L}^{DREL}(f_{t+1}) - \mathcal{L}^{DREL}(f_t)\|^2 + (1 - p_t) \|g_t - \nabla \mathcal{L}^{DREL}(f_t)\|^2
\end{aligned}$$

Using the L -smoothness assumption in Eq. (B.33), we have

$$\mathbb{E}[\|g_{t+1} - \nabla \mathcal{L}^{DREL}(f_{t+1})\|^2] \leq \frac{(1 - p_t)L^2}{b'} \|f_{t+1} - f_t\|^2 + (1 - p_t) \|g_t - \nabla \mathcal{L}^{DREL}(f_t)\|^2$$

Proof of Theorem B.3. We leverage the above lemmas to prove the Theorem. Adding Eq. (B.37)

with $\frac{\gamma}{2p} \times \text{Eq. (B.38)}$ and taking expectation results in the following:

$$\begin{aligned}
& \mathbb{E} \left[\mathcal{L}^{DREL}(f_{t+1}) - L_*^{DREL} + \frac{\gamma}{2p} \|g_{t+1} - \nabla \mathcal{L}^{DREL}(f_{t+1})\|^2 \right] \\
& \leq \mathbb{E} \left[\mathcal{L}^{DREL}(f_t) - \mathcal{L}_*^{DREL} - \frac{\gamma}{2} \|\nabla \mathcal{L}^{DREL}(f_t)\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2} \right) \|f_{t+1} - f_t\|^2 \right] \\
& \quad + \frac{\gamma}{2} \mathbb{E} [\|g_t - \nabla \mathcal{L}^{DREL}(f_t)\|^2] + \frac{\gamma}{2p} \mathbb{E} [(1-p) \|g_t - \nabla \mathcal{L}^{DREL}(f_t)\|^2] \\
& \quad + \frac{\gamma}{2p} \mathbb{E} \left[\frac{(1-p)L^2}{b'} \|f_{t+1} - f_t\|^2 \right] \\
& = \mathbb{E} \left[\mathcal{L}^{DREL}(f_t) - \mathcal{L}_*^{DREL} + \frac{\gamma}{2p} \|g_t - \nabla \mathcal{L}^{DREL}(f_t)\|^2 \right] \\
& \quad - \mathbb{E} \left[\frac{1}{2\gamma} - \frac{L}{2} - \frac{(1-p)\gamma L^2}{2pb'} \|f_{t+1} - f_t\|^2 \right]
\end{aligned}$$

where L_*^{DREL} is the loss at the optimal f^* . Using the inequality of $\frac{1}{2\gamma} - \frac{L}{2} - \frac{(1-p)\eta L^2}{2pb'} \geq 0$, *i.e.*,

$$\gamma \leq \frac{1}{L \left(1 + \sqrt{\frac{1-p}{pb'}} \right)} \quad (\text{B.40})$$

we can write the following

$$\begin{aligned}
& \mathbb{E} [\|g_{t+1} - \nabla \mathcal{L}^{DREL}(f_{t+1})\|^2] \\
& \leq \mathbb{E} \left[\mathcal{L}^{DREL}(f_t) - \mathcal{L}_*^{DREL} + \frac{\gamma}{2p} \|g_t - \nabla \mathcal{L}^{DREL}(f_t)\|^2 - \frac{\gamma}{2} \|\nabla \mathcal{L}^{DREL}(f_t)\|^2 \right]
\end{aligned}$$

Now let us define $\phi_t := \mathcal{L}^{DREL}(f_t) - \mathcal{L}_*^{DREL} + \frac{\gamma}{2p} \|g_t - \nabla \mathcal{L}^{DREL}(f_t)\|^2$ then we can write the following

$$\mathbb{E}[\phi_{t+1}] \leq \mathbb{E}[\phi_t] - \frac{\gamma}{2} \mathbb{E}[\|\nabla \mathcal{L}^{DREL}(f_t)\|^2] \quad (\text{B.41})$$

Now summing from $t = 0$ to $T - 1$ results in the following

$$\mathbb{E}[\phi_T] \leq \mathbb{E}[\phi_0] - \frac{\gamma}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla \mathcal{L}^{DREL}(f_t)\|^2] \quad (\text{B.42})$$

According to the Algorithm 3, \hat{f}_T is chosen from $\{f_t\}_{t \in [T]}$ and $\phi_0 = \mathcal{L}^{DREL}(f_0) - \mathcal{L}_*^{DREL} + \frac{\gamma}{2p} \|g_0 - \nabla \mathcal{L}^{DREL}(f_0)\|^2 = \mathcal{L}^{DREL}(f_0) - \mathcal{L}_*^{DREL} = \Delta_0$, we have

$$\mathbb{E}[\|\nabla \mathcal{L}^{DREL}(\hat{f}_T)\|^2] \leq \frac{2\Delta_0}{\gamma T} \quad (\text{B.43})$$

Setting $T = \frac{2\Delta_0}{\epsilon^2 \gamma}$ and using Jensen's inequality results in the following

$$\mathbb{E}[\|\nabla \mathcal{L}^{DREL}(\hat{f}_T)\|] \leq \mathbb{E}[\|\nabla \mathcal{L}^{DREL}(\hat{f}_T)\|^2] \leq \sqrt{\frac{2\Delta_0}{\gamma T}} = \epsilon \quad (\text{B.44})$$

With the following total number of iterations

$$T = \frac{2\Delta}{\epsilon^2\gamma} = \frac{2\Delta_0L}{\epsilon^2} \left(1 + \sqrt{\frac{1-p}{pb'}}\right) \quad (\text{B.45})$$

we can obtain ϵ -approximate stationary point solution. The number of gradient steps required in the Algorithm 3 is given as

$$N_{grad} = b + T(pb + (1-p)b') \quad (\text{B.46})$$

Replacing T by Equation (B.45), we have the following

$$N_{grad} = b + \frac{2\Delta_0L}{\epsilon^2} \left(1 + \sqrt{\frac{1-p}{pb'}}\right) (pb + (1-p)b') \quad (\text{B.47})$$

This proves Theorem B.3.

Algorithm 3: Alternative Optimization between f and p using Probabilistic SGD

- 1 **Initialize:** f_0 , stepsize γ , minibatch sizes $b, b' < b$, $p_t \in [0, 1]$, $t = 0$, $p_n(f_0) = 1 \forall n \in [1, b]$
 - 2 **compute** $g_0 = \frac{1}{b} \sum_{n \in B} a_n(f_0) \nabla \mathcal{L}_n^{DRO}(f_0)$ $a_n(f_0) = b * p_n(f_0)$ with B, B' being random minibatch samples with $|B| = b$ and $|B'| = b'$
 - 3 **while** $t < T$ **do**
 - 4 $f_{t+1} \leftarrow f_t - \gamma g_t$
 - 5 $prev_use \sim Ber(p_t)$
 - 6 **if** $prev_use = 1$ **then**
 - 7 Find loss $\mathcal{L}_n(f)$ associated with datasample \mathbf{x}_n , $\forall n \in B$
 - 8 Find $a_n(f^{t+1}) = b * \frac{\exp\left(\frac{\mathcal{L}_n(f_t)}{\alpha}\right)}{\sum_{j=1}^b \exp\left(\frac{\mathcal{L}_j(f_t)}{\alpha}\right)}$
 - 9 Find $g_{t+1} = \frac{1}{b} \sum_{n \in B} a_n(f_{t+1}) \nabla \mathcal{L}_n(f^{t+1})$
 - 10 **else**
 - 11 Find loss $\mathcal{L}_n(f)$ associated with data sample \mathbf{x}_n , $\forall n \in B'$
 - 12 Find $a_n(f_{t+1}) = b' * \frac{\exp\left(\frac{\mathcal{L}_n(f_t)}{\alpha}\right)}{\sum_{j=1}^{b'} \exp\left(\frac{\mathcal{L}_j(f_t)}{\alpha}\right)}$
 - 13 Find $g_{t+1} = g_t + \frac{1}{b'} \sum_{n \in B'} (a_n(f_{t+1}) \nabla \mathcal{L}_n(f_{t+1}) - a_n(f_t) \nabla \mathcal{L}_n(f_t))$
 - 14 $t \leftarrow t + 1$
 - 15 **return** \hat{f}_T chosen from $\{f_t\}_{t \in [T]}$
-

Appendix C

In this section we provide additional information for Chapter 6. First, we summarize all notations used in this work. After that, we discuss other related works in addition to those reviewed in the related work section of the Chapter 6. Next, we provide the theoretical proof for Theorem 6.2. Then, we provide experimental details along with additional results. Finally, we provide a link to the source code.

C.1 Summary of Notations

Table C.1 organizes all the major notations into three groups and describes their meanings.

C.2 Additional Related Work

In this section, we review some additional related works, including few-shot learning and open-set recognition.

Few-shot Learning. Few-shot learning is becoming a popular method due to its ability to quickly generalize to new tasks containing only a few examples. These methods are grouped into three categories: *model-based*, *optimization-based*, and *metric-based*. Model-based methods largely depend on a model design for the fast adaptation [96,115], which are less frequently used in recent years. Optimization-based methods back-propagate the gradients to deal with generalization problems. Ravi et al. [111] model a meta-learner as an LSTM so that knowing historical gradients can benefit current gradient updates. MAML [34] and its variants [35,43,113] learn meta parameters with outer

Table C.1: Notations with Descriptions

Symbol Group	Notation	Description
Meta Learning	N^{tr}	Number of training tasks
	S_i^{tr}	Support set for i^{th} task in Meta-train
	Q_i^{tr}	Query set for i^{th} task in Meta-train
	S_i^{te}	Support set for i^{th} task in Meta-test
	Q_i^{te}	Query set for i^{th} task in Meta-test
	K	Number of examples in support set
	N	Number of Classes in support set
	C^s	Set of Closed-set Classes
	C^u	Set of Open-set Classes
Evidential Loss	θ	Neural Network Parameter
	h	Hidden dimensionality of feature extractor
	e_k	Evidence belonging to class k
	S	Total Dirichlet Strength
	u	Uncertainty (vacuity) mass associated with a given data point
	α_{ik}	Dirichlet parameter for the i^{th} data point in the k^{th} class
	$KL(P Q)$	Kullback–Leibler divergence between two distributions P and Q
Transformer	A	Square $N \times N$ matrix with attention weights
	F	Backbone feature extractor
	T	Transformer
	\mathcal{P}	Prototype obtained from backbone
	\mathcal{P}'	Transformed prototype representation
	\mathbf{p}_n	Prototype corresponding to the n^{th} class
	\mathbf{p}'_n	Transformed prototype representation corresponding to class n
	\mathcal{P}_a	Altered prototype by replacing nearest class prototype by query instance
	\mathcal{P}'_a	Transformed prototype representation of altered prototype
	\mathcal{P}'_ϵ	Transformed prototype representation of original prototype using cross-attention

updates utilizing query samples and task-specific parameters via support samples. Metric-based methods learn a good distance function to compare feature similarity between support and query set samples. Cosine distance is learned in [141] with a recurrent network to measure similarities between samples. Prototypical network [127] represents each class as a prototype utilizing support set samples and then computes its similarity with the query set ones. Relation network [131] predicts a relation score between a pair of support and query set samples rather utilizing metrics directly on the feature space. FEAT [152] transforms each class prototype via transformer functions and results in a richer representation. Since feature-based metrics are also useful for open-set recognition, we largely focus on those approaches. While those methods show promising results in closed-set settings, few attempts have explored whether they can be effectively adapted for open-set recognition.

Open-set Recognition. Various support vector machines (SVMs) and reconstruction-based approaches have been proposed to tackle the OSR problem in existing literature [54, 119]. For instance, Scheirer et al. [119] propose a Weibull-calibrated SVM (W-SVM) technique by leveraging the Extreme Value Theory (EVT). Zhang & Patel [157] propose a reconstruction-based approach, where a threshold defined over the reconstruction error is used to distinguish known-class samples from unknown classes. Additionally, various traditional models such as nearest neighbor [59] and quasi-linear function [15], have also been used in the open-set detection tasks. More recently, deep learning models have been adopted for open-set detection and multiple approaches have been proposed. For instance, Scheirer et al. propose OpenMax [119], in which the probability output from a softmax function is redistributed in order to produce the probability of being unknown. VAE-based approaches have also been proposed for the open-set detection [155], [130]. For example, Yoshihashi et al. propose a reconstruction-based approach that performs open-set detection similar to OpenMax by leveraging the effective latent representation trained using VAE [155].

C.3 Theoretical Proof

In this section, we provide the Proof of Lemma 6.1 and Theorem 6.2.

C.3.1 Proof of Lemma 6.1

Proof. Based on **P1**, we can write the following

$$\left\{ \max_{n \in N} [e_{jn}] \right\}_{\text{easy-closed}} \geq \left\{ \max_{n \in N} [e_{jn}] \right\}_{\text{ch-open}} \quad (\text{C.1})$$

where easy-closed indicates the easy closed-set sample whereas ch-open indicates a challenging open-set sample. According to this equation, for the easy-closed set sample as $\max_{n \in N} [e_{jn}]$ is high, the EVR_i will be high *i.e.*, evidence dominates EVR_i to make it high. Based on **P2**, we can write the following

$$\{\text{var}_{n \in N} [e_{jn}]\}_{\text{easy-open}} \leq \{\text{var}_{n \in N} [e_{jn}]\}_{\text{ch-open}} \quad (\text{C.2})$$

where easy-open indicates an easy open-set sample. In this case, for the easy open-set sample the output evidence will remain low (closed to 0) with respect to all closed-set classes making $\text{var}_{n \in N} [e_{jn}]$ low. This low variance will dominate EVR_i to make it high. In case of a challenging open-set sample, the maximum evidence $\max_{n \in N} [e_{jn}]$ is relatively low while variance $\text{var}_{n \in N} [e_{jn}]$ is high, making EVR_i low. Therefore, we can say that with a being a challenging task and b being a regular task *i.e.*, easy closed-set or easy open-set, we have $\text{EVR}_a < \text{EVR}_b$. This completes the proof of Lemma 6.1. \square

C.3.2 Proof of Theorem 6.2

Proof. Specifically, we need to show the following:

$$d(\mathcal{P}'_a, \mathcal{P}') \leq d(\mathcal{P}'_a, \mathcal{P}'_c) \quad (\text{C.3})$$

In the above equation, the transformed representation for \mathcal{P}' on j^{th} prototype on the l^{th} dimension can be represented as

$$\{\mathcal{P}'\}_{jl} = \sum_{n=1}^N a_{jn} f_{nl}; \forall j \in [N], \forall l \in [h] \quad (\text{C.4})$$

where a_{jn} is the attention for j^{th} row and n^{th} column and f_{nl} be the associated feature obtained with value *value* (\mathcal{V}) in the transformer network, h is the feature dimensionality. Let c be the closest class for a given query sample then each element of the altered transformed prototype *i.e.*, \mathcal{P}'_a for this sample can be represented as

$$\{\mathcal{P}'_a\}_{jl} = \sum_{n=1}^N a'_{jn} f'_{nl}; \forall j \in [N], \forall l \in [h] \quad (\text{C.5})$$

where a'_{jn} is the attention weight for altered transformed prototype for j^{th} row and n^{th} column and f'_{jl} being associated feature. It should be noted that $a'_{jn} = a_{jn}$ if $j \neq c$ or $n \neq c$ and $f'_{nl} = f_{nl}$ if $n \neq c$. The transformed prototype obtained using our proposed cross-attention mechanism can be represented as

$$\{\mathcal{P}'_{\epsilon}\}_{jl} = \left\{ \begin{array}{ll} a_{cc}f_{cl} + \frac{\epsilon}{EVR} \sum_{n=1, n \neq c}^N a_{jn}f_{nl} & \text{if } j == c \\ \frac{\epsilon}{EVR} a_{jc}f_{cl} + \sum_{n=1, n \neq c}^N a_{jn}f_{nl}, & \text{otherwise} \end{array} \right\} \quad (\text{C.6})$$

Considering d being the Euclidean distance, we compute $d(\mathcal{P}'_a, \mathcal{P}')$ as:

$$d(\mathcal{P}'_a, \mathcal{P}') = \frac{1}{N} \sum_{j=1}^N \sqrt{\sum_{l=1}^h (\{\mathcal{P}'_a\}_{jl} - \{\mathcal{P}'\}_{jl})^2} \quad (\text{C.7})$$

Similarity, we can compute $d(\mathcal{P}'_a, \mathcal{P}'_{\epsilon})$ term as:

$$d(\mathcal{P}'_a, \mathcal{P}'_{\epsilon}) = \frac{1}{N} \sum_{j=1}^N \sqrt{\sum_{l=1}^h (\{\mathcal{P}'_a\}_{jl} - \{\mathcal{P}'_{\epsilon}\}_{jl})^2} \quad (\text{C.8})$$

It is noted that $\forall j \in [N], \forall l \in [h]$, if we show $(\{\mathcal{P}'_a\}_{jl} - \{\mathcal{P}'\}_{jl})^2 \leq (\{\mathcal{P}'_a\}_{jl} - \{\mathcal{P}'_{\epsilon}\}_{jl})^2$ then the inequality in Eq (C.4) becomes valid. For simplicity, let us assume that $\mathbf{U}_{jl} = \{\mathcal{P}'_a\}_{jl}$, $\mathbf{V}_{jl} = \{\mathcal{P}'\}_{jl}$, $\mathbf{W}_{jl} = \{\mathcal{P}'_{\epsilon}\}_{jl}$. Then, $\forall j \in [N], l \in [h]$, we need to prove:

$$(\mathbf{U}_{jl} - \mathbf{V}_{jl})^2 \leq (\mathbf{U}_{jl} - \mathbf{W}_{jl})^2 \quad (\text{C.9})$$

Let us write both sides where we seek to find the inequality relation between them

$$(\mathbf{U}_{jl} - \mathbf{V}_{jl})^2 \stackrel{?}{\leq} (\mathbf{U}_{jl} - \mathbf{W}_{jl})^2 \quad (\text{C.10})$$

Expanding both sides and canceling common terms we have the following

$$\mathbf{V}_{jl}^2 - 2\mathbf{U}_{jl}\mathbf{V}_{jl} \stackrel{?}{\leq} \mathbf{W}_{jl}^2 - 2\mathbf{U}_{jl}\mathbf{W}_{jl} \quad (\text{C.11})$$

Further simplification leads to the following

$$2\mathbf{U}_{jl}(\mathbf{W}_{jl} - \mathbf{V}_{jl}) \stackrel{?}{\leq} (\mathbf{W}_{jl} - \mathbf{V}_{jl})(\mathbf{W}_{jl} + \mathbf{V}_{jl}) \quad (\text{C.12})$$

As $\frac{\epsilon}{EVR} > 1$, $\mathbf{W}_{jl} > \mathbf{V}_{jl}$. As such $(\mathbf{W}_{jl} - \mathbf{V}_{jl})$ is non-negative and therefore, we can cancel $(\mathbf{W}_{jl} - \mathbf{V}_{jl})$ on both sides without changing their inequality sign. This leads to the following:

$$2\mathbf{U}_{jl} \stackrel{?}{\leq} (\mathbf{W}_{jl} + \mathbf{V}_{jl}) \quad (\text{C.13})$$

In the inequality, since $\frac{\epsilon}{EVR} > 1$, there exists a constant $k > 1$ that makes $\mathbf{W}_{jl} = k\mathbf{V}_{jl}$. Substituting this in the above equation, we have the following:

$$2\mathbf{U}_{jl} \textcircled{?} (1+k)\mathbf{V}_{jl} \tag{C.14}$$

It is noted that attention weights of the altered prototype in \mathbf{U}_{jl} are likely to be similar to \mathbf{V}_{jl} in case of a challenging query sample. This is because the challenging sample may be very similar to one of the prototypes making the output representation almost identical. This makes \mathcal{U}_{jl} to be similar to \mathcal{V}_{jl} . However, on the right-hand side, we have $(1+k) > 2$, which therefore makes the right term bigger than the left term. It should be noted that k is the term dependent on the ratio $\frac{\epsilon}{EVR}$ where the higher the ratio, the higher the k term would be. Therefore, under the non-negativity assumption of the feature representation (which can be achieved simply using a non-negative transformation function in the output), with high probability following holds for the challenging samples.

$$2\mathbf{U}_{jl} \leq (1+k)\mathbf{V}_{jl} \tag{C.15}$$

This completes the proof of Theorem 6.2. □

C.4 Experimental Details and Additional Results

In this section, we first provide the dataset distribution of all datasets used in the experimentation. After that, we provide the implementation-specific details along with the data split strategy. Next, we provide the closed-set performance of those datasets with respect to competitive baselines. Next, we explain the ROC curves generated for the same set of datasets. After that, we conduct an additional ablation study. Finally, we conduct an in-depth qualitative analysis to show the effectiveness of our proposed technique.

C.4.1 Dataset Distribution

Table C.2 shows the dataset splits for four datasets: MiniImageNet, TieredImageNet, Cifar100, and Caltech101. It should be noted that to serve our purpose we have considered the whole data distribution and divided it into closed-set and open-set.

C.4.2 Experimentation Details

In this section, we describe the way we split the data along with the implementation details.

Table C.2: Train/Evaluation/Test partition on different datasets.

Split	MiniImageNet			TieredImageNet			Cifar100			Caltech101		
	Train	Eval	Test	Train	Eval	Test	Train	Eval	Test	Train	Eval	Test
<i>Open-set</i>	21	6	8	116	24	41	27	5	8	20	10	10
<i>Closed-set</i>	46	10	9	259	73	95	35	10	15	40	10	10

Dataset Split. According to Table C.2, we first partition the entire dataset into training, validation, and testing. Within the training set, we perform semantic analysis at the class level to identify groups of semantically relevant classes (*e.g.*, different categories of dogs). For datasets with a relatively small number of classes (*e.g.*, MiniImageNet), this introduces minimal overhead. For larger datasets, we can usually benefit from some existing hierarchical structure among the classes. For instance, in MiniImageNet, training classes Ferrets (88) and Malamute dog (83) are semantically similar. Instead of using both of them as closed-set samples during training like all existing approaches, we assign Malamute dog (83) as one of the opponent classes, which are used as part of the evidential open-set loss. As demonstrated in our experiments, this arrangement clearly improves the detection of some similar open-set classes, such as Golden Retriever Golden Retriever (82) and African Hunting dog (85).

Implementation Details. For the experimentation on both datasets, we used the ResNet-12 as a backbone architecture for the feature extractor followed by the transformer network. For good initialization, the feature extractor is connected to a fully connected layer (with output nodes equal to a number of classes present in the training set) and trained using the cross entropy (CE) classification loss by treating it as a multi-class classification problem. Once the model is trained, the last layer is removed and the transformer network is connected. Finally, the model is trained in the FSL open-set detection setting using the training loss defined in (6.10). For the training, stochastic gradient descent (SGD) is used with a total of 200 epochs. The initial learning rate of 0.002 is set and is decreased by 10% at an interval of every 20 epochs. The weight decay is set to 0.005 and λ is set to 1 throughout the experimentation.

Table C.3: Closed set performance (ACC) on different datasets.

Approach	MiniImageNet		TieredImageNet	
	1-Shot	5-Shot	1-Shot	5-Shot
PEELER	61.24 ± 0.46	76.46 ± 0.50	45.64 ± 0.44	60.22 ± 0.23
SnaTCHer	63.91 ± 0.63	79.93 ± 0.43	47.75 ± 0.72	64.04 ± 0.65
MET	63.13 ± 0.62	79.00 ± 0.43	47.91 ± 0.73	63.33 ± 0.48

C.4.3 Closed-Set Performance

Table C.3 shows the closed-set performance of MET with respect to the competitive baselines. As shown our approach generates comparable closed-set performance while having a much better OSR performance as demonstrated in Table 6.1.

C.4.4 ROC Curves

To provide a detailed view of AUROC, we further show the ROC curves for the 1-shot and 5-shot scenarios in the miniImageNet and TieredImageNet datasets as shown in Figure C.1. The ROC plot has a similar pattern in the other two datasets. It is worth mentioning from the ROC curves that the proposed technique stays on the top, especially for the lower false positive rate (FPR) region. For example, in the case of Figure C.1 (a), we can achieve True Positive Rate (TPR) around 70% while maintaining FPR below 30% which is more than 20% higher than the second best competitive model. This concludes that the proposed approach can correctly identify far more open-set samples compared to other baselines while being able to maintain a low FPR.

C.4.5 Ablation Study

In this section, we conduct an ablation study with regard to the hyperparameters λ and ϵ . Next, we explain the effectiveness of our proposed technique using different backbones. After that, we report the performance in the original data split. Finally, we also conduct additional experimentation using Cifar100 and Caltech datasets.

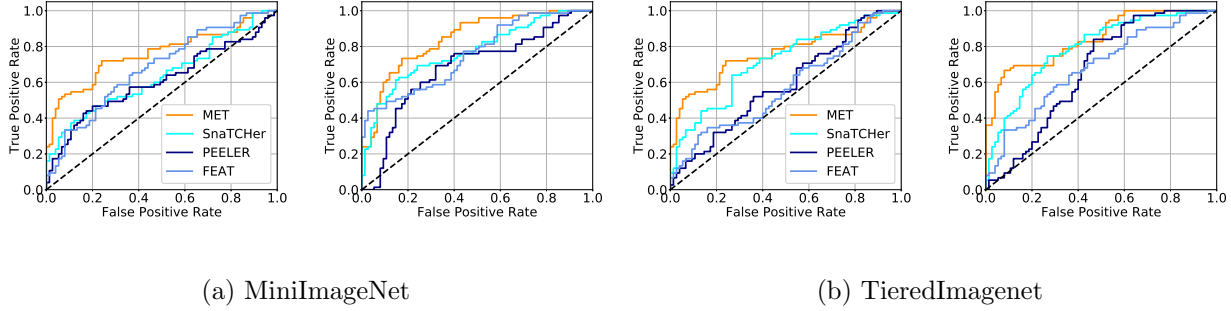
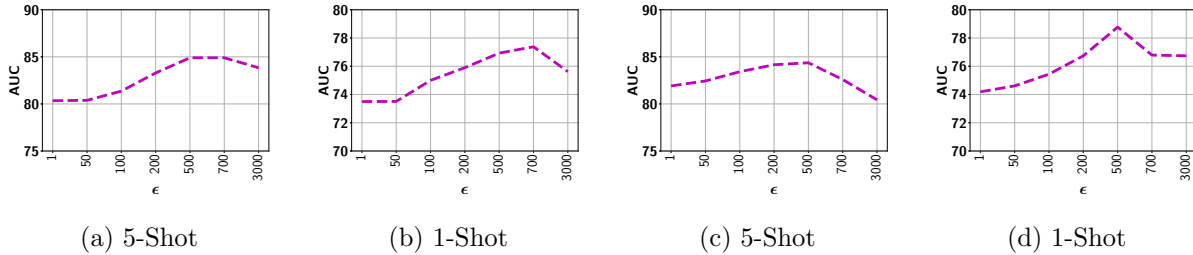


Figure C.1: ROC curves on both 5-way-1-shot and 5-way-5-shot tasks

Figure C.2: OSR performance with respect to hyperparameter ϵ : (a-b) MiniImageNet, (c-d) TieredImageNet.

Sensitivity to λ and ϵ : Figure C.2 shows the impact of hyperparameter ϵ on the model performance. As can be seen, a very small ϵ value is not beneficial because the model fails to shift the attention by assigning a larger weight to the predicted class. Having a higher ϵ value helps the model to change the attention weight according to EVR, *i.e.*, a higher EVR leads to a lower change. But, having a very high ϵ leads to degradation in performance as it dramatically changes the representation irrespective of the EVR value. In general, the model performance is quite stable as long as ϵ is not set to very high or very low values. With the middle range of ϵ as shown by Figure C.2, the model automatically calibrates the change in accordance with EVR leading to better performance.

Figure C.3 shows the impact of the open-set weight λ on the performance. As shown, having a low λ value, the model puts less emphasis on opponent open-set classes, leading to a less compact representation of closed-set classes that can benefit open-set detection. On the other hand, having a very high λ value may be problematic as the model puts too much emphasis on the opponent open-set classes without paying much attention on learning from the closed-set classes resulting in performance degradation as well. In general, the λ value in the middle range (*e.g.*, $\lambda = 1$) gives a good balance between open-set and closed-set losses resulting in the best performance.

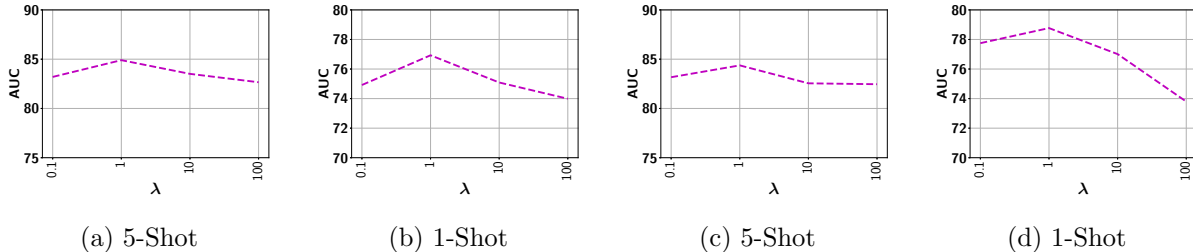


Figure C.3: OSR performance with respect to hyperparameter λ : (a-b) MiniImageNet, (c-d) Tiered-ImageNet.

Table C.4: MiniImageNet performance with: (a) Different backbones, (b) Original data split.

Approach	ResNet12	ResNet18
<i>PEELER</i>	60.36 \pm 0.72	61.52 \pm 0.64
<i>SnaTCHer</i>	67.37 \pm 0.91	68.11 \pm 0.94
<i>MET</i>	76.93 \pm 0.59	77.11 \pm 0.62

(a) Performance on Different Backbones.

Approach	1-shot	5-shot
<i>PEELER</i>	60.12 \pm 0.72	68.23 \pm 0.66
<i>SnaTCHer</i>	72.98 \pm 0.61	79.57 \pm 0.49
<i>MET</i>	73.20 \pm 0.45	81.19 \pm 0.47

(b) Original Data Split Performance.

Different Backbones. Table C.4 (a) demonstrates the performance comparison between different backbones for the MiniImageNet dataset (1-Shot setting). As shown, for multiple backbones, our technique has a superior performance compared to the competitive baselines.

Original Data Split. Our technique requires similar classes in open-set as well as closed-set to demonstrate the effectiveness of our technique. Therefore, in the main evaluation, we altered the data split. In this section, we show that even for the original data split, we are being able to outperform the previous baselines. Table C.4 (b) shows the performance for different baselines for the MiniImageNet dataset in the original data split. As shown, the proposed technique has superior performance compared to the other baselines. It should be noted that because of our novel technique paradigm along with the novel cross-attention technique, we are being able to outperform other baselines. Different from another evaluation strategy like SnaTCHer, we fixed open-set and closed-set samples in both training as well as testing processes while keeping the original split i.e., training, validation, and testing identical. Also, it is worth mentioning that for a fair comparison, we consider the identical setting (e.g., backbone, transformer) for PEELER and

Table C.5: OSD (AUROC) performance on additional datasets.

Approaches	Cifar100 5-way		Caltch101 5-way	
	1-shot	5-shot	1-shot	5-shot
<i>ProtoNet</i>	48.12 ± 0.23	50.63 ± 0.35	47.34 ± 0.26	51.35 ± 0.54
<i>RelationNet</i>	48.76 ± 0.65	51.54 ± 0.56	47.95 ± 0.46	52.62 ± 0.35
<i>OpenMAX</i>	51.42 ± 0.54	54.45 ± 0.60	49.18 ± 0.52	49.77 ± 0.52
<i>FEAT (Probability)</i>	49.25 ± 0.60	52.30 ± 0.59	48.99 ± 0.53	51.08 ± 0.52
<i>Feat (Distance)</i>	54.69 ± 0.46	59.86 ± 0.46	59.19 ± 0.52	65.30 ± 0.49
<i>PEELER</i>	52.46 ± 0.43	56.11 ± 0.15	51.10 ± 0.72	55.96 ± 0.22
<i>SnaTCHer</i>	57.60 ± 0.57	62.06 ± 0.52	62.37 ± 0.63	67.35 ± 0.53
<i>TANE</i>	55.14 ± 0.64	63.08 ± 0.58	52.71 ± 0.19	56.65 ± 0.18
MET	61.76 ± 0.60	66.17 ± 0.54	64.85 ± 0.55	72.12 ± 0.47

SnaTCHer and rerun them.

Experimental Results on Additional Datasets. In this section, we conduct experimentation on additional datasets (Cifar100, Caltech101) to further justify the effectiveness of our proposed technique. Table C.5 demonstrates the OSD performance on those datasets. As shown, in Cifar 100 and Caltech101 datasets, MET achieves around 5 – 8% improvement over most competitive baseline SnaTCHer for both 5-shot and 1-shot setups.

C.4.6 Qualitative Analysis

In this section, we perform an in-depth quantitative analysis to justify the effectiveness of our proposed technique. Figure C.4 (a) demonstrates some difficult examples from closed-set class Ferrets (88) and open-set class African Hunting Dog (85). As shown in the first image *i.e.*, leftmost of Figure C.4 (a), the image looks different from many others from class Ferrets (88) because of the different color and camera angle. As a result, SnaTCHer incorrectly classifies it as an open-set sample and assigns the highest distance among all samples in the same task. Although MET (w/o EVR) helps to decrease the distance by leveraging the opponent open-set classes in training, it is not sufficient to correctly identify it as open-set. With the help of the novel EVR detection, MET (w/ EVR) is able to correctly identify it as a closed-set sample. In the case of the second image *i.e.*, middle image in Figure C.4 (a), because of its similarity with the first image, due to color (and possible other low-level image features), both SnaTCHer and MET (w/o EVR) incorrectly

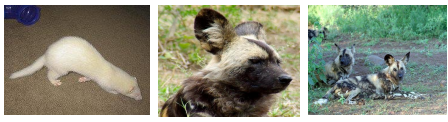


Image	SnaTCHer	MET (w/o EVR)	MET (w/ EVR)
Image (a) 88	32	32	15
Image (b) 85	1	2	17
Image (c) 85	7	15	26

(a) Images: Ferrets (88), African Hunting dog (85)

(b) Ranking

Figure C.4: Examples of difficult images with the corresponding ranking

classify it as a closed-set sample. In contrast, MET (w/ EVR) is able to correctly identify it as an open-set sample. In the case of the third image *i.e.*, the rightmost image in Figure C.4 (a), it shares some similarities (in terms of color and body pattern) with closed-set samples, SnaTCHer has trouble identifying it as an open-set sample. MET (w/o EVR) helps to increase the distance (*i.e.*, uncertainty) but it is not sufficient to classify confidently as an open-set sample. With further help from EVR-based detection, we are able to correctly identify it as an open-set sample.

Similarly, Table C.4 (b) shows the relative ranking of images based on the output prototype distance. It should be noted that a ranking of 1 *i.e.*, highest ranked (or lowest prototype distance) means the given sample is closest to the prototype among all samples. For Figure C.4 (a) leftmost image, we want the sample to be ranked close to the top so that it is closer to the prototype compared to most open-set samples. However, it ranks at 32 by both SnaTCHer and MET (w/o EVR), which will negatively impact the overall AUROC score. In contrast, MET (w/ EVR) ranks as 15, higher than $\frac{2}{3}$ of other samples, most of which are open-set one, leading to an improved AUROC score. In the case of the middle image in Figure C.4 (a), both SnaTCHer and MET (w/o EVR) rank it very high *i.e.*, better than most closed-set samples. In the case of the rightmost image in Figure C.4 (a), there is some improvement using our novel evidential loss by MET (w/o EVR). With the help of EVR, MET (w/ EVR) is able to further push this sample down to the rank of 26. We provide some additional qualitative analysis in the Appendix.

Appendix D

In this appendix, we provide additional supplementary information related to Chapter 7. We first present a table summarizing the major notations used by the Chapter 7. Next, we provide detailed information about the training process and hyperparameters setting. We provide the detailed proof of Lemma 7.1 and Theorem 7.2 in Section D.3. After that, we provide additional experimental details and results. Finally, we discuss the broader impacts, limitations, and future work of our DRE technique.

D.1 Summary of Notations

Table D.1 below shows the major notations used in the Chapter 7. We further assign each notation into one of four major categories: dataset, DRO formulation, sparse training, and theoretical results.

D.2 Robust Loss Optimization in DRO

In this section, we first provide a detailed description on how we optimize the robust loss function in (7.1). We then explain how to set the uncertainty set by choosing a proper hyperparameter.

D.2.1 Robust Loss Optimization

The optimization problem specified in (7.1) involves an inequality constraint so directly solving it may incur a higher computational overhead. Therefore, we consider a regularized version of the

Table D.1: Symbols with Descriptions.

Symbol Group	Notation	Description
Dataset	\mathbf{X}	Set of training images
	\mathbf{Y}	Set of training class labels
	C	Total classes
	\hat{y}	Predicted class label
	N	Total number of training samples
	D	Dimensionality of each data sample
DRO	D_f	f -divergence
	η	Parameter controlling size of uncertainty set in DRO framework
	z_n	Weight associated with n^{th} data sample
Sparse Training	M	Number of sparse sub-networks
	\mathcal{K}	Density of the given network
	Θ	Parameter associated with given neural network
	\hat{p}	Confidence associated with predicted class
	$l(\mathbf{x}_n, \Theta)$	Loss associated with n^{th} data sample
Theoretical Results	β	Learning rate of the given network
	P	Total number of patches in each data sample
	d	Dimensionality of each patch
	$\mathbf{v}_{c,l}$	Major l^{th} feature associated with class c
	L	Total number of features in each class class
	D_N^S	Collection of single-view data samples
	D_N^M	Collection of multi-view data samples
	\cup	Collection of features
	H	Number of convolution layers
	$F_c(\mathbf{x})$	Logistic output for the c^{th} class for the data sample \mathbf{x}
	$\mathcal{P}_{\mathbf{v}_{c,l}}$	Collection of patches containing feature $\mathbf{v}_{c,l}$ in sample \mathbf{x}_j
	SOFT_c	Softmax output for class c

robust loss to train each base learner by using the following loss:

$$\mathcal{L}^{Robust} = \max_{\mathbf{z} \geq \mathbf{0}, \mathbf{z}^\top \mathbf{1} = 1} \sum_{n=1}^N z_n l_n(\Theta) - \lambda D_f \left(\mathbf{z} \parallel \frac{\mathbf{1}}{N} \right) \quad (\text{D.1})$$

where $l_n(\Theta) = l(\mathbf{x}_n, \Theta)$. Solving the above maximization problem leads to a closed-form solution for \mathbf{z}^* as shown by the following lemma:

Lemma D.1. *Assuming that D_f is the KL divergence, then solving (D.1) leads to the following solution*

$$\mathcal{L}^{Robust} = \sum_{n=1}^N z_n^* l_n(\Theta) \quad (\text{D.2})$$

where z_n^* is given by

$$z_n^* = \frac{\exp\left(\frac{l_n(\Theta)}{\lambda}\right)}{\sum_{j=1}^N \exp\left(\frac{l_j(\Theta)}{\lambda}\right)} \quad (\text{D.3})$$

It can be verified that there is a one-to-one correspondence between η in (7.2) and λ in (D.1). Given their roles in the corresponding equations, a large η implies a small λ and a small η implies a large λ .

D.2.2 Hyperparameter settings

The hyperparameter in the regularization term is chosen based on the difficulty of a dataset. Specifically, for DRE, we always consider the $\lambda \rightarrow \infty$ for the first sparse sub-network which is equivalent to Expected Risk Minimization (ERM). For the second and third sub-networks, we choose this hyperparameter based on the difficulty of data samples. It should be noted that we need to set higher λ values for more difficult datasets as difficult samples are more common on those datasets. Using this notion, for Cifar10, we choose small λ values so that the model can focus on the difficult samples that are few. For this, we choose $\lambda = 10$ for the second sparse sub-network and $\lambda = 500$ for the third sparse sub-network. Considering Cifar100 is more difficult, we would have more difficult samples and therefore higher λ value is preferred. For this, we choose $\lambda = 50$ for the second sparse sub-network and $\lambda = 500$ for the third one. In the case of TinyImageNet, we have many difficult samples and therefore we choose relatively large λ values. Specifically, we choose $\lambda = 100$ for the second sparse sub-network and $\lambda = 1,000,000$ for the third sparse sub-network.

D.3 Theoretical Proof

In this section, we provide detailed proofs of the theoretical results presented in the Chapter 7.

D.3.1 Proof of Lemma 7.1

Proof. For $y_n = c$, with respect to data sample $\{\mathbf{x}_n, y_n\}$, the gradient can be evaluated as

$$-\nabla_{\Theta_{c,h}} l(\Theta; \mathbf{x}_n, y_n) = [1 - \text{SOFT}_c(F(\mathbf{x}_n))] \sum_{p \in [P]} \text{ReLU}[\langle \Theta_{c,h}, \mathbf{x}_n^p \rangle] \mathbf{x}_n^p \quad (\text{D.4})$$

Assume that the given sample has a major feature $\mathbf{v}_{c,l}$, taking dot product with respect to $\mathbf{v}_{c,l}$ on both side of (D.4) leads

$$\langle -\nabla_{\Theta_{c,h}} l(\Theta; \mathbf{x}_n, y_n), \mathbf{v}_{c,l} \rangle = [1 - \text{SOFT}_c(F(\mathbf{x}_n))] \sum_{p \in [P]} \langle \text{ReLU}[\langle \Theta_{c,h}, \mathbf{x}_n^p \rangle] \mathbf{x}_n^p, \mathbf{v}_{c,l} \rangle \quad (\text{D.5})$$

Let's further assume that the feature set is orthonormal: $\forall c, c', \forall l \in [L], \|\mathbf{v}_{c,l}\|_2 = 1$ and $\mathbf{v}_{c,l} \perp \mathbf{v}_{c',l'}$ when $(c, l) \neq (c', l')$. Using $\mathbf{x}^p = a^p \mathbf{v}_{c,l} + \sum_{\mathbf{v}' \in \mathcal{U} \setminus \mathbf{v}_c} \alpha^{p, \mathbf{v}'} \mathbf{v}' + \epsilon^p$ given in (7.4), we have

$$\langle -\nabla_{\Theta_{c,h}} l(\Theta; \mathbf{x}_n, y_n), \mathbf{v}_{c,l} \rangle = [1 - \text{SOFT}_c(F(\mathbf{x}_n))] \left(\sum_{p \in \mathcal{P}_{v,l}(\mathbf{x}_n)} \text{ReLU}[\langle \Theta_{c,h}, \mathbf{x}_n^p \rangle] a^p + \sum_{p \in [P]} \langle \epsilon^p, \mathbf{v}_{c,l} \rangle \right) \quad (\text{D.6})$$

It should be noted that the term *i.e.*, $\sum_{\mathbf{v}' \in \mathcal{U} \setminus \mathbf{v}_c} \alpha^{p, \mathbf{v}'} \langle \mathbf{v}', \mathbf{v}_{c,l} \rangle$ becomes zero due to the orthogonal properties of the feature set. Let us represent the second term by κ : $\sum_{p \in [P]} \langle \epsilon^p, \mathbf{v}_{c,l} \rangle = \kappa$. Then, we have

$$\langle -\nabla_{\Theta_{c,h}} l(\Theta; \mathbf{x}_n, y_n), \mathbf{v}_{c,l} \rangle = (1 - \text{SOFT}_c(F(\mathbf{x}_n))) \left(\sum_{p \in \mathcal{P}_{v,l}(\mathbf{x}_n)} \text{ReLU}[\langle \Theta_{c,h}, \mathbf{x}_n^p \rangle] a^p + \kappa \right) \quad (\text{D.7})$$

Furthermore, let us define $V_{c,h,l}(\mathbf{x}_j) = \sum_{p \in \mathcal{P}_{v,l}(\mathbf{x}_j)} \text{ReLU}[\langle \Theta_{c,h}, \mathbf{x}_j^p \rangle] a^p$ then above equation further reduces to following

$$\langle -\nabla_{\Theta_{c,h}} l(\Theta; \mathbf{x}_n, y_n), \mathbf{v}_{c,l} \rangle = (1 - \text{SOFT}_c(F(\mathbf{x}_n))) (V_{c,h,l}(\mathbf{x}_n) + \kappa) \quad (\text{D.8})$$

Recall the above equation is the gradient with respect to the n^{th} data sample. Considering the gradient with respect to all data samples with $y_n = c$, and let us consider the total loss, where the

weight z_n of each loss is assigned according to a distribution specified by the uncertainty set \mathcal{U} . Then, the total gradient is

$$\langle -\nabla_{\Theta_{c,h}} l(\Theta; \mathbf{X}, \mathbf{Y}), \mathbf{v}_{c,l} \rangle = \max_{\mathbf{z} \in \mathcal{U}} \sum_{n=1}^N z_n [\mathbb{1}_{y_j=c}(V_{c,h,l}(\mathbf{x}_n) + \kappa)(1 - \text{SOFT}_c(F(\mathbf{x}_n)))] \quad (\text{D.9})$$

Now using the standard gradient update rule with β being the learning rate, we have

$$\langle \Theta_{c,h}^{t+1}, \mathbf{v}_{c,l} \rangle = \langle \Theta_{c,h}^t, \mathbf{v}_{c,l} \rangle + \beta \max_{\mathbf{z} \in \mathcal{U}} \sum_{n=1}^N z_n [\mathbb{1}_{y_j=c}(V_{c,h,l}(\mathbf{x}_n) + \kappa)(1 - \text{SOFT}_c(F(\mathbf{x}_n)))] \quad (\text{D.10})$$

Let $\mathbf{x}_k \in \mathcal{D}_N^S$ be the most difficult sample having $\mathbf{v}_{c,l}$ as the main feature. Also, consider $\mathbf{x}_n \in \mathcal{D}_N^M$ to be the easy sample with $y_n = c, y_k = c$. Then, we have

$$[1 - \text{SOFT}_c(F(\mathbf{x}_k))] \geq [(1 - \text{SOFT}_c(F(\mathbf{x}_n)))] , \forall n \in [1, N], n \neq k, y_n = c \quad (\text{D.11})$$

Using above property, we can write the following using (D.10)

$$\begin{aligned} & \langle \Theta_{c,h}^t, \mathbf{v}_{c,l} \rangle + \beta \max_{\mathbf{z} \in \mathcal{U}} \sum_{n=1}^N z_n [\mathbb{1}_{y_j=c}(V_{c,h,l}(\mathbf{x}_n) + \kappa)(1 - \text{SOFT}_c(F(\mathbf{x}_n)))] \\ & \leq \langle \Theta_{c,h}^t, \mathbf{v}_{c,l} \rangle + \beta N z_k (1 - \text{SOFT}_c(F(\mathbf{x}_k))) \end{aligned} \quad (\text{D.12})$$

On the r.h.s., we have $z_n = \frac{1}{N}$ for ERM, which assigns equal weights to all samples. Under the assumption of $N_{\mathbf{v}_{c,l}} \ll N_{\cup \mathbf{v}_{c,l}}$, the contribution of the $N_{\mathbf{v}_{c,l}}$ on overall gradient will be negligible. In contrast, for the DRO framework, using (D.3), we have

$$z_k = \frac{1}{\sum_{j=1, j \neq k}^N \exp\left(\frac{l_j(\Theta) - l_k(\Theta)}{\lambda}\right) + 1} \quad (\text{D.13})$$

Since $l_k(\Theta) > l_j(\Theta), \forall \lambda > 0, \lambda \neq \infty$, we have $z_k > \frac{1}{N}$. Using r.h.s. of (D.12) and incorporating $z_k = \frac{1}{N}$ for ERM and $z_k > \frac{1}{N}$, we have

$$\{\langle \Theta_{c,h}^t, \mathbf{v}_{c,l} \rangle + \beta(1 - \text{SOFT}_c(F(\mathbf{x}_k)))\}_{ERM} \leq \{\langle \Theta_{c,h}^t, \mathbf{v}_{c,l} \rangle + \beta(1 - \text{SOFT}_c(F(\mathbf{x}_k)))\}_{Robust} \quad (\text{D.14})$$

This subsequently leads to the following:

$$\{\langle \Theta_{c,h}^t, \mathbf{v}_{c,l} \rangle\}_{Robust} > \{\langle \Theta_{c,h}^t, \mathbf{v}_{c,l} \rangle\}_{ERM}; \forall t > 0 \quad (\text{D.15})$$

which completes the proof of Lemma 7.1. \square

D.3.2 Proof of Theorem 7.2

Let $\mathbf{x} \in \mathcal{D}_S^N$ from class c with $\mathbf{v}_{c,l}$ as the main feature and \mathbf{v}' as the dominant feature learned through the memorization. Also consider \mathbf{v}' to be the main feature characterizing class k . Then for any class c' , we can define the following

$$\text{SOFT}_{c'}(\mathbf{x}) = \frac{\exp(F_{c'}(\mathbf{x}))}{\sum_{j \in [C]} \exp(F_j(\mathbf{x}))} \quad (\text{D.16})$$

In the above equation, $F_{c'}(\mathbf{x})$ can be written as

$$F_{c'}(\mathbf{x}) = \sum_{h \in [H]} \sum_{p \in [P]} \text{ReLU}[\langle \Theta_{c',h}, \mathbf{x}^p \rangle] \quad (\text{D.17})$$

Substituting \mathbf{x}^p from (7.4), we have

$$F_{c'}(\mathbf{x}) = \sum_{h \in [H]} \sum_{p \in [P]} \text{ReLU} \left[a^p \langle \Theta_{c',h}, \mathbf{v}_{c,l} \rangle + \sum_{\mathbf{v}' \in \mathcal{U} \setminus \mathbf{v}_c} \alpha^{p,\mathbf{v}'} \langle \Theta_{c',h}, \mathbf{v}' \rangle + \langle \Theta_{c',h}, \epsilon^p \rangle \right] \quad (\text{D.18})$$

Substituting c' by k , we have

$$F_k(\mathbf{x}) = \sum_{h \in [H]} \sum_{p \in [P]} \text{ReLU} \left[a^p \langle \Theta_{k,h}, \mathbf{v}_{c,l} \rangle + \sum_{\mathbf{v}' \in \mathcal{U} \setminus \mathbf{v}_c} \alpha^{p,\mathbf{v}'} \langle \Theta_{k,h}, \mathbf{v}' \rangle + \langle \Theta_{k,h}, \epsilon^p \rangle \right] \quad (\text{D.19})$$

In case of ERM, the $\mathbf{v}_{c,l}$ signal is fairly weak during the training process due to $N_{\mathbf{v}_{c,l}} \ll N_{\mathcal{U} \setminus \mathbf{v}_{c,l}}$. Therefore, the term $\langle \Theta_{k,h}, \mathbf{v}_{c,l} \rangle$ is negligible. Also, the last term $\langle \Theta_{k,h}, \epsilon^p \rangle$ is also small as this corresponds to the Gaussian noise. For the second term $\exists \mathbf{v}'$ for which $\langle \Theta_{k,h}, \mathbf{v}' \rangle$ is very high because of the spurious correlation. In contrast, for the robust loss, using Lemma 7.1, the model learns a stronger correlation with the true class parameter and therefore $\langle \Theta_{c,h}, \mathbf{v}_{c,l} \rangle$ is high. As such, both terms $\langle \Theta_{k,h}, \mathbf{v}_{c,l} \rangle$ as well as $\langle \Theta_{k,h}, \mathbf{v}' \rangle, \forall \mathbf{v}'$ becomes low. As a result, we have

$$\{F_k(\mathbf{x})\}_{ERM} > \{F_k(\mathbf{x})\}_{Robust} \quad (\text{D.20})$$

Substituting this inequality to (D.16), we have

$$\{\text{SOFT}_k(\mathbf{x})\}_{Robust} < \{\text{SOFT}_k(\mathbf{x})\}_{ERM} \quad (\text{D.21})$$

This completes the proof of Theorem 7.2.

D.4 Experimental Details and Additional Results

In this section, we first provide a detailed description of datasets used in our experimentation followed by hardware description of our experimentation. Consequently, we provide examples of single-view and multi-view data samples. Next, we provide additional experimental results on Cifar10 and Cifar100 datasets with a 15% density. After that, we provide additional baselines results on TinyImageNet. We also compare our model performance with different calibration techniques commonly used in dense networks. Then, we perform an in-depth ablation study. Parameter size and inference speed are discussed in the subsequent subsection. We also further investigate the diversity of the sparse subnetworks. Finally, we provide detailed qualitative analysis to support our proposed claim.

D.4.1 Detailed Dataset Description

For general classification setting, we consider Cifar10, Cifar100 [67], and TinyImageNet [72] datasets. For the out of distribution setting, we consider corrupted version of Cifar10 and Cifar100, which are named as Cifar10-C and Cifar100-C [49], respectively. Finally, for open-set detection, we leverage SVHN [100] as the open-set dataset. The detailed description of each dataset is given below:

- *Cifar10*. This dataset consists of total 10 classes, each consisting of 5,000 training samples and 1,000 testing (evaluation) samples. Each image is a colored image with size 32×32 .
- *Cifar100*. This dataset consists of 20 super classes where each super-class consists of 5 classes resulting into total 100 classes. Each class consists of 500 training samples and 100 testing samples. Each image is a colored image with size 32×32 .
- *TinyImageNet*. The original dataset consists of 200 classes with 1,000,000 samples where each class has 500 training images, 50 validation images, and 50 test images. Each image is a colored image with size 64×64 .
- *Cifar10-C*. Fifteen different types of corruptions are applied on the Cifar10 clean testing dataset where each corruption has 5 severity levels, ranging from 1 to 5 with 1 being least severe and 5 being most severe. The corruptions include Gaussian noise, shot noise, impulse noise, defocus blur, forsted glass blur, motion blur, zoom blur, snow, frost, fog, brightness, contrast, elastic, pixelate, and JPEG.

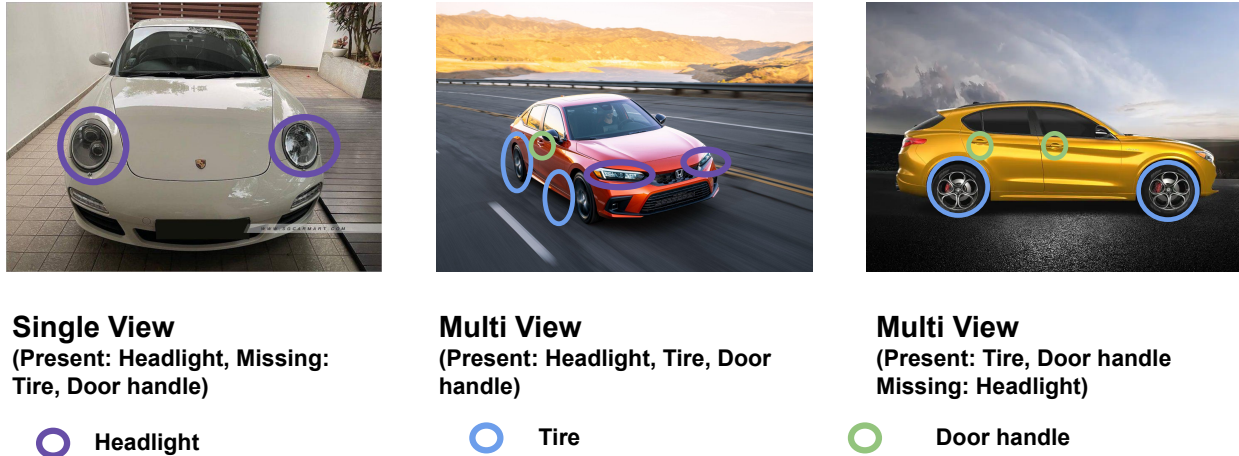


Figure D.1: Examples of single-view and multi-view samples.

- *Cifar100-C*. Similar to *Cifar10-C*, fifteen different corruptions are applied on the *Cifar100* clean testing dataset.
- *SVHN*. The Street View House Numbers (*SVHN*) dataset consists of 10 classes with digit 1 as class 1, digit 9 as class 9 and digit 0 as class 10. These are original, variable-resolution, colored house-number images with character level bounding boxes. We use this dataset as the open-set dataset in our experimentation.

D.4.2 Hardware Details for Experimentation

All experimentations are conducted using NVIDIA RTX A6000 GPU with 48GB memory requiring 300 Watt power. For GPU, CUDA Version: 11.6, Driver Version: 510.108.03, and NVIDIA-SMI: 510.108.03 is used. In terms of CPU, our experimentation uses an Intel(R) Xeon(R) Gold 6326 CPU @ 2.90GHz with a 64-bit system and an x86_64 architecture.

D.4.3 Single-view and Multi-view Examples

Figure D.1 show the three example images, where the first image is a representative single-view data sample whereas the last two are multi-view samples. In this example, we consider three major features for cars: *i.e.*, Tire, Headlight, and Door handle. As only headlight feature is present in the first image, it belongs to the single-view category. For the second and third images, multiple

features are presented and therefore we regard those images as multi-view data samples.

Table D.2: Accuracy and ECE performance with 15% density for Cifar10 and Cifar100 Dataset.

Training Type	Approach	Cifar10				Cifar100			
		ResNet50		ResNet101		ResNet101		ResNet152	
		<i>ACC</i>	<i>ECE</i>	<i>ACC</i>	<i>ECE</i>	<i>ACC</i>	<i>ECE</i>	<i>ACC</i>	<i>ECE</i>
	<i>Dense</i> [†]	94.82	5.87	95.12	5.99	76.40	16.89	77.97	16.73
Dense Training	<i>L1 Pruning</i>	93.88	5.69	94.23	5.88	75.53	15.52	75.83	15.78
	<i>LTH</i>	92.97	4.03	93.15	5.69	74.36	15.13	74.77	15.22
	<i>DLTH</i>	95.15	6.21	95.65	6.96	77.98	16.24	78.23	16.54
	<i>Mixup</i>	93.22	4.02	93.38	5.68	74.48	15.10	74.68	15.16
Sparse Training	<i>CigL</i>	92.25	4.67	93.34	4.59	77.88	10.16	77.27	10.62
	<i>DST Ensemble</i>	89.57	2.10	88.64	1.34	64.57	9.76	64.75	9.27
	<i>Sup-ticket</i>	94.65	3.20	94.95	3.09	78.68	10.16	78.95	10.32
Mask Training	<i>AdaBoost</i>	94.07	5.65	94.76	5.14	75.98	23.55	76.28	24.27
	<i>EP</i>	94.41	3.90	94.42	4.07	75.66	14.79	76.05	14.79
	<i>SNE</i>	94.85	3.05	94.96	3.18	76.82	11.12	77.23	11.63
	<i>DRE</i>	94.87	1.71	94.74	1.34	75.86	4.90	76.46	5.81

D.4.4 Additional Result on Cifar10 and Cifar100

Table D.2 shows the experimental result on Cifar10 and Cifar100 datasets with a 15% density. As shown, the proposed technique has a far superior performance in terms of the ECE score compared to the competitive baselines. This is consistent with the results with a 9% density as presented in the Chapter 7, which further justifies the effectiveness of our proposed technique.

D.4.5 Additional Baseline Results on TinyImageNet

As mentioned in the Chapter 7, the computational issue (*i.e.*, memory overflow) makes it impossible to run sparse learning techniques *i.e.*, CigL [73], DST Ensemble [81], and Sup-ticket [153] on the ResNet101 and WideResNet101 architectures to make a fair comparison. Therefore, in this section, we pick a lower capacity model (ResNet50) and compare the performance. Even for the ResNet50

architecture, CigL still runs into the memory overflow issue with a batch size of 128. Furthermore, lowering the batch size (*e.g.*, 16) makes the training process extremely slow even using a 48Gb GPU, where each training epoch takes more than half an hour, making model training extremely difficult. Therefore, we did not report the performance of CigL. It should be noted that CigL can be trained on Cifar10 and Cifar100 because of lower dimension of the input images and we have already reported its performance in the Chapter 7. Table D.3 shows the performance of DRE along with those from DST Ensemble and Sup-ticket on ResNet50. It is clear that DRE achieves better performance compared to these baselines.

D.4.6 Performance from Ensemble Members

We investigate how performance varies in different sparse sub-networks. We use Cifar100 as an example and Table D.4 report the individual sub-network performance on both accuracy and ECE. While each sparse sub-network is a relatively weaker learner (which is expected), they contribute to the final ensemble model in a complementary way, leading to a better ECE score as well as accuracy.

D.4.7 Comparison with Common Calibration Techniques

In this section, we investigate whether existing calibration techniques designed for training dense networks can be leveraged to further improve the calibration performance of sparse networks. However, most of these techniques (*e.g.*, temperature scaling and mix-n-match) are post hoc techniques, which require a separate validation set to fine-tune the parameters. This means we need to further divide the training data into training and validation sets, which may negatively impact the generalization capability of the trained model (due to less training data). To make a comparison, we pick Temperature Scaling (TS) [45], Label Smoothing (LS) [132], and

Table D.3: Additional baseline results on TinyImageNet using ResNet50 with $\mathcal{K} = 15\%$.

Training Type	Approach	ACC	ECE
Sparse Training	<i>DST Ensemble</i>	72.00	2.94
	<i>Sup-ticket</i>	68.68	10.96
Mask Training	<i>DRE</i>	71.57	1.51

Table D.4: Different subnetworks performance on Cifar100 Dataset.

Subnetworks	ResNet101		ResNet152	
	ACC	ECE	ACC	ECE
<i>Subnetwork 1 (3%)</i>	68.22	14.35	69.65	13.31
<i>Subnetwork 2 (3%)</i>	69.03	1.39	70.00	3.39
<i>Subnetwork 3 (3%)</i>	72.86	11.96	70.24	14.78
<i>DRE</i>	74.68	1.20	74.37	2.09

a few other techniques proposed in [158], including Ensemble Temperature Scaling (ETS) and Isotonic Regression One vs All combined with Temperature Scaling (IROvA-TS). We apply these calibration techniques on the top of the EP algorithm. Specifically, as LS does not require a separate validation set, we train it on the full training dataset using the LS loss (with $\epsilon = 0.1$). Other calibration techniques require a separate validation set and therefore we divide training data into training and validation with a 80:20 ratio. EP (No Validation) uses the full training dataset whereas EP (Validation) is trained using 80% of the training data. Once the model is trained with 80% of training data using EP, we further calibrate it using the aforementioned calibration techniques. Table D.5 shows the results. There are two key observations: (i) the classification accuracy decreases for all calibration techniques at the expense of improving calibration performance as they require a separate validation set, and (ii) DRE achieves the best ECE in all cases, which further justifies its strong calibration performance.

Table D.5: Different calibration techniques on the top of EP Algorithm with $\mathcal{K} = 9\%$.

Approach	Cifar10				Cifar100			
	ResNet50		ResNet101		ResNet101		ResNet152	
	<i>ACC</i>	<i>ECE</i>	<i>ACC</i>	<i>ECE</i>	<i>ACC</i>	<i>ECE</i>	<i>ACC</i>	<i>ECE</i>
TS	93.42	0.96	93.42	1.37	73.06	1.72	73.40	2.45
ETS	93.42	0.97	93.42	1.37	73.06	1.76	73.40	2.40
IROvA-TS	89.90	1.45	88.69	0.89	60.87	1.56	60.77	2.86
LS	94.06	7.56	94.21	7.41	75.96	9.36	76.40	7.71
EP (No Validation)	94.20	3.97	94.35	4.03	75.05	14.62	75.68	14.41
EP (Validation)	93.42	4.46	93.42	4.83	73.06	15.56	73.40	15.88
DRE	94.60	0.7	94.28	0.7	74.68	1.20	74.37	2.09

D.4.8 Ablation Study

In this section, we first show the impact of λ values on the prediction and calibration performance. We then investigate how the size of the ensemble affects its calibration performance. Finally, we show the effectiveness of the proposed technique as we vary the backbones. In addition to the backbones used in the main chapter, we will further evaluate two other commonly used backbones, including WideResNet28 and Vision Transformer (ViT) [28] as backbones.

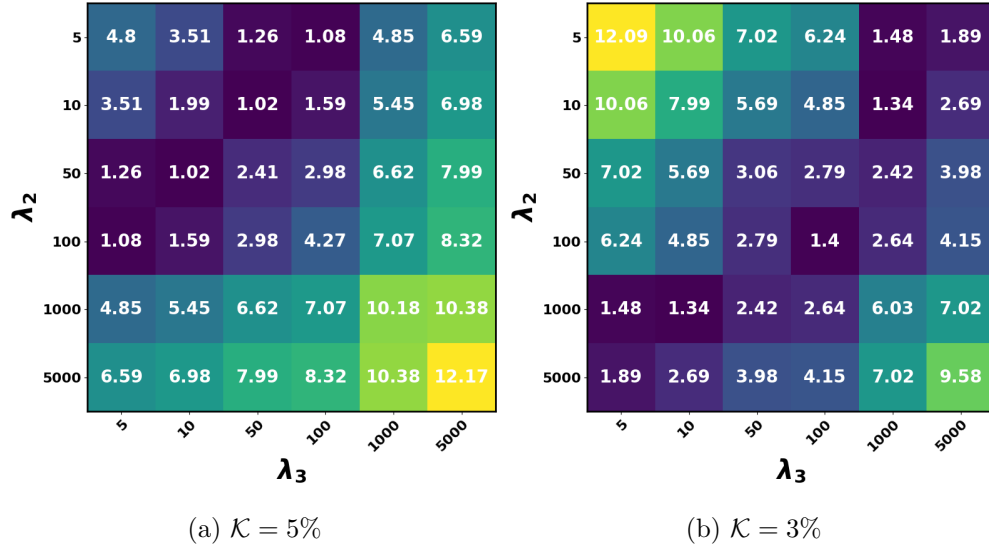
Figure D.2: (a-b) Impact of λ on ECE using ResNet101 architecture on Cifar100 dataset.

Table D.6: ACC and ECE with different: (a) backbones and (b) number of subnetworks.

Approach	WideResNet28-10		ViT	
	ACC	ECE	ACC	ECE
EP	94.12	4.53	86.16	10.01
DRE	93.98	1.93	85.53	4.18

(a) Different backbones on Cifar10 Dataset.

Approach	ResNet101		ResNet152	
	ACC	ECE	ACC	ECE
DRE ($M = 3$)	94.87	1.71	94.74	1.34
DRE ($M = 5$)	94.79	0.84	94.69	0.62

(b) Different M values on Cifar10 with $\mathcal{K} = 15\%$.

Impact of the uncertainty set size. For simplicity, we always keep one sparse sub-network in our framework to be with $\lambda_1 \rightarrow \infty$. The ECE performance with respect to different sets of λ value for the remaining sub-networks is shown using the heatmap given in Figure D.2 (a-b). As can be seen, it is important to choose λ_2 and λ_3 with very distinct values to achieve a low calibration error.

Performance analysis of different backbones. Table D.6 (a) reports the performance of Cifar10 from both DRE and EP using different backbone architectures. In case of WideResNet28-10, the calibration error is low without sacrificing the accuracy. It also demonstrates that the superior performance of DRE is not limited to a specific backbone. In case of ViT, DRE still achieves a much lower calibration error than EP. However, using ViT as a backbone, the accuracy

from both EP and DRE is lower and ECE is higher than other backbones. Existing studies show that without pretraining, the lack of useful inductive biases for ViT can cause performance drop [1]. Since no pretraining is conducted in both EP and DRE, it causes a lower accuracy (and a higher ECE).

Impact of number of sparse-sub-networks. In this analysis, we study the impact of number of sparse sub-networks. It should be noted that our work is not limited only for $M = 3$. We can instead increase the M value. For example, Table D.6 (b) shows the performance for ensemble model with $M = 5$, where each sub-network is trained with $\mathcal{K} = 3\%$ leading to a total $\mathcal{K} = 15\%$. We also show the performance with $M = 3$, where each sub-network is trained with $\mathcal{K} = 5\%$. As can be seen, if there is a sufficient learning capacity for each sub-network, the ECE score can further improve with the increase of M .

D.4.9 Parameter Size and Inference Speed

We compare parameter size and inference speed of different types of sparse networks. Table D.7 shows the FLOPS along with number of parameters associated with each technique. As can be seen, the proposed DRE has a comparable parameter size as that of the sparse network ensemble. In terms of computational times, our approach is comparable to the sparse network ensemble. Compared to a dense network, our technique has a much smaller parameter size with less FLOPS.

Table D.7: Parameter size and inference speed.

Approach	ResNet50		ResNet101	
	Params	Flops ($\times 10^9$)	Params	Flops ($\times 10^9$)
<i>Dense</i> [†]	23.6M	4.14	42.5M	7.88
<i>SNE</i>	3.5M	1.31	6.3M	2.53
<i>DRE</i>	3.5M	1.31	6.3M	2.53

D.4.10 Diversity on Sparse Sub-networks

To justify our claim that our technique ensures the diverse sparse sub-networks, we adapt the disagreement metric (d_{dist}) from [81]. This metric measures the disagreement among sub-networks in terms of class label prediction. Table D.8 below shows the results for Cifar10 and Cifar100 datasets. As shown, compared to Sparse Network Ensemble, DRE achieves higher disagreement which implies that the sparse sub-networks are more diverse.

Table D.8: Accuracy, ECE, and prediction disagreement performance with a $\mathcal{K} = 15\%$ density.

Approach	Cifar10						Cifar100					
	ResNet50			ResNet101			ResNet101			ResNet152		
	<i>ACC</i>	<i>ECE</i>	<i>d_{dist}</i>	<i>ACC</i>	<i>ECE</i>	<i>d_{dist}</i>	<i>ACC</i>	<i>ECE</i>	<i>d_{dist}</i>	<i>ACC</i>	<i>ECE</i>	<i>d_{dist}</i>
SNE	94.85	3.05	0.048	94.96	3.18	0.049	76.82	11.12	0.20	77.23	11.63	0.20
DRE (Ours)	94.87	1.71	0.088	94.74	1.34	0.069	75.86	4.90	0.24	76.46	5.81	0.24

D.4.11 Qualitative Analysis

In this section, we provide illustrative examples to further justify the proposed DRE is better calibrated compared to existing baselines. Figure D.3 (a)-(d) show the confidence values for the wrongly classified samples using different baselines. As can be seen, all of the baselines suffer from the overfitting issue, resulting into the incorrect predictions with high confidence. In contrast, as shown in Figure D.3 (e)-(f), the sparse sub-networks provide the confidence values in different ranges, where sub-network in (a) is learned from representative samples and (c) from the difficult ones. As these sub-networks are complementary with each other, the DRE has a much better confidence distribution for both the correct as well as incorrect samples. Figure D.4 shows the confidence score of correctly classified data samples from the CIFAR100 dataset with different techniques. As shown, our DRE technique remains confident on the correct data samples while being not confident on the incorrect data samples. This result shows our approach is well calibrated and trustworthy compared with the competitive baselines. In summary, our proposed technique remains uncertain for incorrect samples while being confident on the correct samples resulting in a much improved calibration.

D.5 Broader Impact, Limitations, and Future Work

In this section, we first describe the potential broader impacts of our work. We then discuss the limitations and identify some possible future directions.

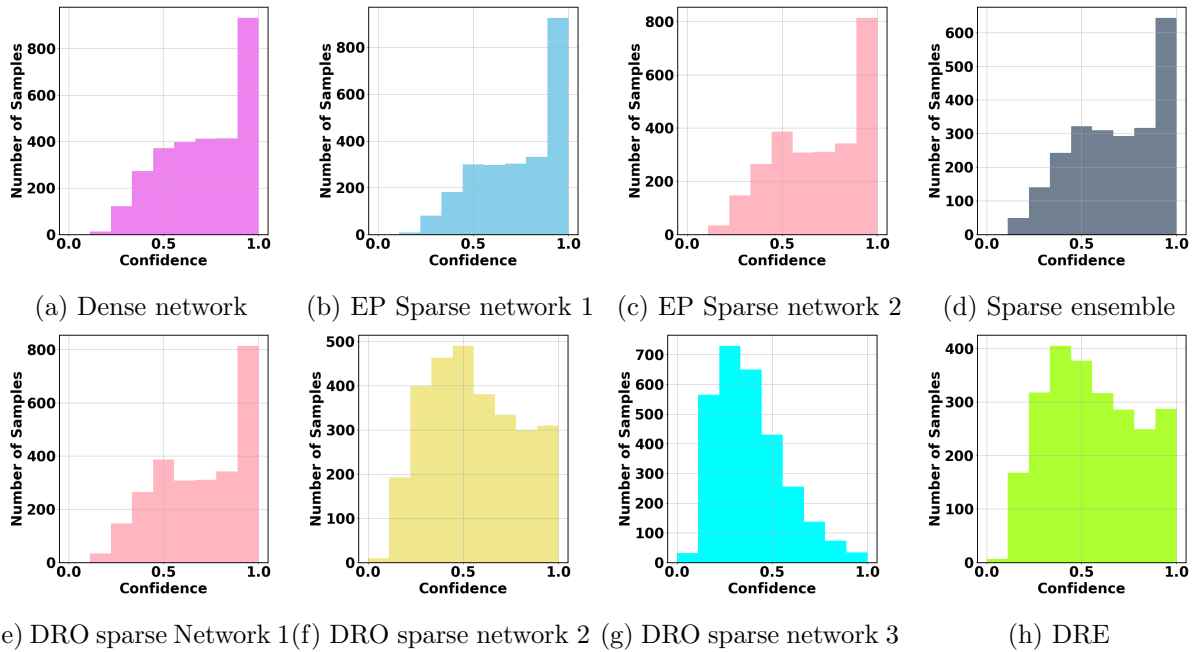


Figure D.3: Confidence scores of incorrectly classified samples in CIFAR100 with ResNet101

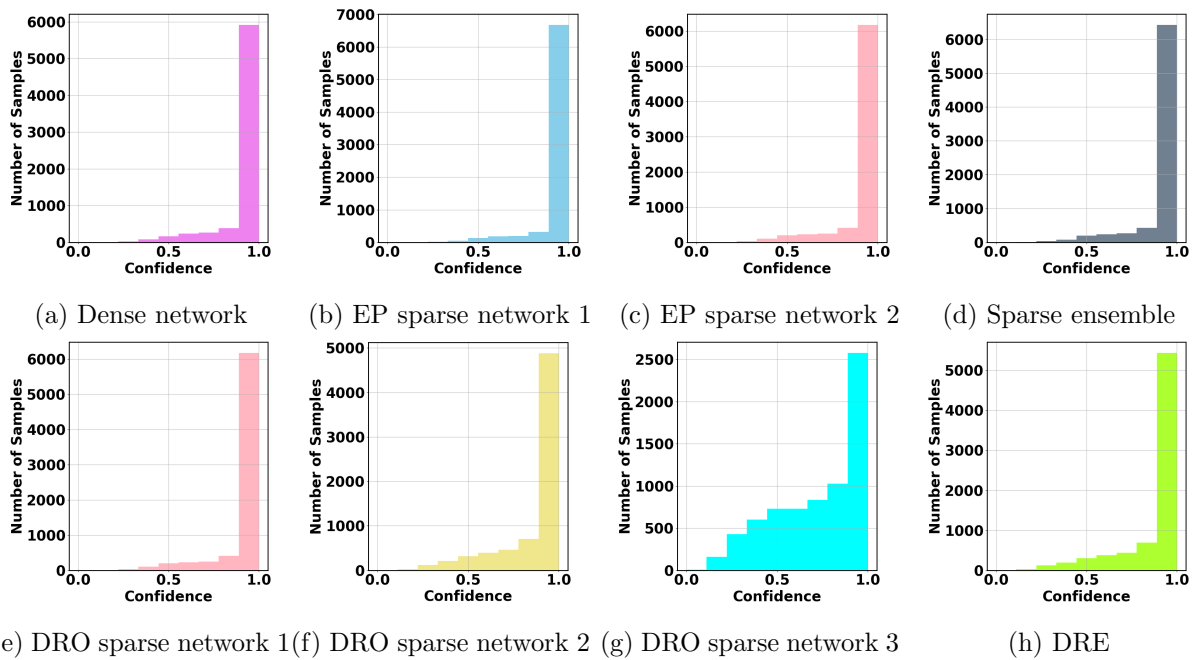


Figure D.4: Confidence scores of correctly classified samples in CIFAR100 with ResNet101

D.5.1 Broader Impact

Sparse network training provides a highly promising way to significantly reduce the computational cost for training large-scale deep neural networks without sacrificing their predictive power. Besides energy savings, it also opens the gate for deploying deep neural networks to lightweight computing or edge devices that can further broaden the applications of AI in more diverse and resource constrained settings. The proposed robust ensemble framework provides a general solution to achieve calibrated training of deep learning models. As a result, the trained model is expected to provide more reliable uncertainty predictions, which could be an important step towards using AI in safety-critical domains.

D.5.2 Limitations and Future Works

As an ensemble model, DRE involves multiple base learners (*i.e.*, sparse sub-networks). Consequently, it may lead to more computational overhead. This could create issues for real-time application as during the inference time, the input needs to be passed through all base learners to get the final output, which can slow down the prediction speed. A straightforward way to speed up the inference process is to execute all the base learners in parallel, which still incurs additional computational overhead. One interesting future direction is to investigate knowledge distillation and train a single sparse network from the ensemble model. Theoretical evidence [1] shows that knowledge distillation has the potential to largely maintain the ensemble performance while providing a promising way to train a single sparse network with an even higher sparsity level and improved inference speed.