Appendix of 'Unbiased Multiple Instance Learning for Weakly Supervised Video Anomaly Detection'

This appendix is organized as follows:

- Section A.1 provides more details about our training objectives. We detail the implementation of the self-training used in our experiments: FixMatch [7].
- Section A.2 explains about how the feature clustering boosts the UMIL during training.
- Section A.3 shows more comparisons and standard deviations on UCF-crime [8], and TAD [6]. In particular, we first discuss the statics of anomaly events in UCF-crime in Section A.3.1, and then provide more experimental results of the proposed UMIL.
- Section A.4 gives the full version of ROC curves on various benchmarks. This is a supplement to Figure. 6 in the manuscript.
- Codes are also provided, which include the training and testing scripts on the two classic datasets. The setup instructions and commands used in our experiments are included in the `README.md` file.

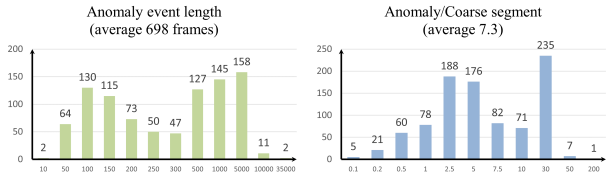
A.1. Loss Objectives

In this section, we give the details of the self-training objective $\mathcal{L}_{st}$ used in the MIL pre-training and UMIL training, as in Eq.(1) and Eq.(4) in the manuscript, then the overall MIL pre-training objective is derived as the following:

$$\mathcal{L}_{mil} = \text{BCE}(\mathcal{C}) + \lambda \mathcal{L}_{st}, \tag{A1}$$

where λ stands for the balance weight. The overall objective of UMIL is derived correspondingly. Note that self-training strategy is an important approach popular in domain adaption [1–3, 5, 10]. In this work, we introduce self-training to boost feature learning in WSVAD by incorporating data augmentation of **FixMatch** [7]. Specifically, along with the training of MIL, we generate pseudo labels with original video snippet data and seek to minimize the entropy between the predictions of augmented data as well as original data. Given the pair of feature $\mathbf{x}$ and $\mathbf{x}'$ from the original data and random augmentation data, respectively, the **FixMatch**-driven self-training loss derives as:

$$\mathcal{L}_{st} = \mathbb{1}(\text{argmax} f(\mathbf{x}) > \delta)\text{BCE}(\mathbf{x}', \text{argmax} f(\mathbf{x})), \tag{A2}$$



(a) (b)

Figure A1. (a) The statistics of anomaly event length and (b) the ratio of anomaly event length to coarse snippet length.

here, $\mathbb{1}$ represents the indicator function that returns 1 if the condition is met, δ is the confident threshold, and BCE corresponds to the binary cross entropy mentioned in the manuscript. In the experiments, we used grid searching for finding a proper δ. The results are listed in Section A.3.

A.2. Discussion on clustering

During the training of UMIL, wrong clustering can bring risks. However, the modern pre-trained backbones (*e.g.*, CLIP) capture rich prior knowledge, such that the intrinsic difference between normal and abnormal snippets is sufficiently expressed in the feature space. This ensures that 1) during clustering, the normal/abnormal snippets can be separated into different clusters; and 2) combining $\mathcal{C}$- and $\mathcal{A}$-supervision in Eq. 4 leads to a classifier using true anomaly features instead of context for prediction (*e.g.*, vertical black line in Figure 2). In future work, we will explore other prior knowledge or inductive bias to further separate normal/abnormal snippets.

A.3. Additional Experiments

A.3.1. Pre-processing Analysis

In this section, we first analyze the rationality in the pre-processing step of previous WSVAD approaches. As mentioned in Section 4.2 of the manuscript, existing works follow the average feature pipeline. They first divide video sequences into multiple coarse snippets, *e.g.* 1 video 32 snippets, then take the *snippet-level average features* as inputs into anomaly detectors. However, real-world anomalies are extremely rare and short in time. The subtle anomaly events

are easily diluted or even covered by normal patterns after the spatio-temporal pooling operation.

To better analyze the problem, we annotate the large training set of UCF-crime [8]. In detail, 5 trained annotators are involved in the process and the final labels are generated by averaging the results. In Figure A1, we depict the statistics of the anomaly events' length. The average length of anomaly events is about 698 frames (extracted from videos with 30FPS), compared with an average coarse snippets' length of 200 frames. Note that coarse snippets' length is obtained by dividing each video into 32 snippets, which is widely used as the default in existing works [6, 8, 9]. We also depict the ratio of anomaly event length to coarse snippet length in Figure A1b. As is shown, there are 164 out of 925 anomaly events whose length ratios are less than 1. It means that these anomaly events are short than the coarse snippet. More importantly, the length ratios of 86 anomaly events are less than 0.5. Considering the anomalies only take place in a small part of whole frames, the anomaly information is inevitably concealed in the spatio-temporal feature pooling process, which is hurtful for video anomaly detection.

A.3.2. Feature Fine-tuning Analysis

When the backbone is loaded with pre-trained weights on kinetics 400 and frozen during UMIL training, the performance will drop from 86.75% (with fine-tuning) to 83.44% (frozen) on UCF, and 92.93% to 90.71% on TAD. This validates that fine-tuning in UMIL enables learning a representation tailored for WSVAD, which is beneficial for anomaly detection.

A.3.3. Self-training In UMIL

In this work, we use the learned anomaly classifier to generate pseudo-labels on samples in the ambiguous set $\mathcal{A}$. Consequently, the ambiguous samples, which are largely neglected in existing MIL, can further participate in our UMIL with pseudo-labels. When the self-training loss is removed, the results of UMIL are 83.66% ($\downarrow$ 3.09%) on UCF and 91.74% ($\downarrow$ 1.19%) on TAD. This validates the effectiveness of the self-training loss. Further experimental analysis of the Self-training can be found in the following.

A.3.4. Confident Threshold in Self-training

In Table A1, we list the results of adding self-training to MIL baseline model with varying confident threshold δ. By increasing the confident threshold, less and highly confident samples are involved in the objective of self-training. As is shown in the table, 0.8 is a suitable threshold that the self-training tool obtains good results on both datasets. When the threshold is up to 1, few samples will be selected leading to the ineffectiveness of self-training.

Threshold(%)	0.3	0.5	0.7	**0.8**	0.9	1.0
AUC_O (%) - UCF	80.9	81.2	81.9	**82.0**	81.5	80.7
AUC_O (%) - TAD	89.0	90.1	90.5	**90.8**	90.1	89.1

Table A1. Ablation on the Confident threshold in self-training based on MIL model on UCF-Crime and TAD.

Threshold(%)	0.5	0.6	0.7	**0.8**	0.9
AUC_O (%) - UCF	86.4	86.6	86.8	**86.8**	86.6
AUC_O (%) - TAD	92.5	92.7	92.8	**93.0**	92.9

Table A2. Ablation on the similarity threshold in clustering on UCF-Crime and TAD.

A.3.5. Similarity Threshold in Clustering

The cluster component alone is for separating normal/abnormal snippets in the ambiguous set $\mathcal{A}$ as two clusters. It doesn't directly benefit the learning of anomaly classifier f. If the $\mathcal{A}$-supervision is removed, f will be trained only on confident normal/abnormal snippets, and our approach will basically reduce to the existing MIL with similar performance. We also conducted experiments to analyze the effect of varying similarity thresholds in clustering. The experimental results are listed in Table A2. As we can see, the performance is insensitive to the change of the threshold in cosine similarity in Eq. (2) of the manuscript. Because the clustering property is acquired along with the feature fine-tuning of the backbone. Then 0.8 is chosen as the default similarity threshold in clustering.

A.3.6. Confident Sample Selection Strategy

In this section, we also conducted experiments to compare *Historical Variance* with *Max Confidence* in the confident sample selection strategy. Specifically, we select the top k (%) abnormal and normal snippets with maximum confidence in abnormal videos as the confident set $\mathcal{C}$. As we can see, the best AUC performances of *Historical Variance* (86.8% for UCF-crime and 93.0% for TAD) are superior to those of *Max Confidence* (85.9% for UCF-crime and 92.2% for TAD). As mentioned in the manuscript (Section 3.2 Step 1), the predictions of the 'easy' normal or abnormal snippets tend to quickly converge to confident normal or anomaly with small variance over time. As a result, collecting historical information of score variance is a better choice for distinguishing confident and ambiguous samples.

A.3.7. Comparison on ShanghaiTech

We also added experiment on ShanghaiTech benchmark [4]. We failed to implement our method on Ped2, due to the absence of data splits. We achieved comparative per-

Threshold(%)	10	30	50	70	90
AUC_O (%) - UCF	**85.9**	85.7	85.3	84.6	83.8
AUC_O (%) - TAD	92.1	**92.2**	92.0	91.4	90.8

Table A3. Ablation on the max confident threshold to divide the confident/ambiguous snippet set on UCF-Crime and TAD.

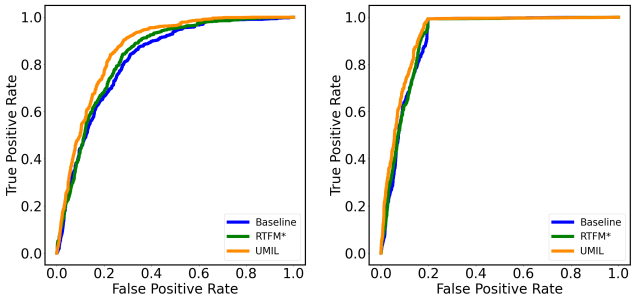


Figure A2. Full version of the ROC curves on UCF (left) and TAD (Right).

formance with existing SOTAs as shown below, and believed that there is potential for further improvements given more time. Additionally, the high accuracy on this dataset indicates that it may contain mainly apparent anomalies, which explains why MIL-based methods already perform well (Section 3.1).

Method	GCN	RTFM	Baseline	Ours
AUC(%)	84.44	97.21	95.20	96.78

A.4. Visualization of ROC Curves

References

[1] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. *ICLR*, 2018. 1

[2] Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, 2020. 1

[3] Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. In *NeurIPS*, 2021. 1

[4] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *CVPR*, 2018. 2

[5] Mingsheng Long, ZHANGJIE CAO, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, 2018. 1

[6] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *TIP*, 2021. 1, 2

[7] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeuIPS*, 2020. 1

[8] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, 2018. 1, 2

[9] Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *BMVC*, 2019. 2

[10] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018. 1