Towards Open Set Video Anomaly Detection Supplementary Material

Yuansheng Zhu, Wentao Bao, and Qi Yu

Rochester Institute of Technology
{yz7008, wb6219 and qi.yu}@rit.edu

The Appendix is organized as follows. In Section A, we summarize the major notions used in the paper. In Section B, we provide the implementation details. In Section C, we provide additional results. In Section D, we show some qualitative examples.

A Notations

The main notations are divided into four major types: Data, Model, Loss, and Hyperparameters, and summarized in Table 1.

Table 1: Summary of notations

Type	Notation
Data	Bag feature X, Bag label Y, instance feature $\mathbf{x}$, instance label y
	Adjacent matrix A
Model	GCNs $\mathcal{H}(\cdot)$
	EDL $\Phi(\cdot)$
	NFs $f(\cdot)$
Loss	Triplet loss $\mathcal{L}_{triplet}$
	MIL loss $\mathcal{L}_{\text{MIL}}$
	NFs loss $\mathcal{L}_{NF}$
Hyperparameters	Loss weight β
	Triplet loss margin m
	thresholds τ_u, τ_p, ϵ for constructing Ω

B Implementation Details

The hyperparameters are chosen as follows: m is set as 0.3 across three datasets, and β is set as 0.001, 0.0001, 0.0001 for XD-Violence, UCF-Crime, and ShanghaiTech, respectively. To be adaptive during the training process, τ_p, τ_u, and ϵ are chosen based on the i-th largest value in a candidate pool during every iteration. Generally, τ_p and τ_u are set to make Ω retain a moderate portion

of instances in every bag, and ϵ is set to make the pseudo anomalies with low probability density. In practice, on XD-Violence, ShanghaiTech, UCF-Crime, τ_p is set as the 50-th, 30-th, 3-rd largest p_+, and τ_u is set as the the 150-th, 150-th, and 24-th largest α_+ in a bag. ϵ is set as the 4750-th largest $p(\tilde{\mathbf{x}}|y=0)$ in a pseudo anomaly pool of size 5000. We gradually perform sample selection, *i.e.,* increasing τ_p from smallest to the assigned value during a warmup stage (Ω evolves from all instances to the most confident clean subset). We perform early stopping to avoid overfitting whenever needed. We optimize the model via the Adam optimizer equipped with cosine annealing learning rate scheduling. We use Python 3.9.7 and PyTorch 1.10.0 to build the test platform, running it on NVIDIA RTX A6000 GPUs. Whenever public results are available, we directly use them for comparison.

C Additional Results

In this section, we present more experimental results along with an additional ablation study to further justify the key components of the proposed framework.

C.1 AUC-ROC on XD-Violence

We show the AUC-PR scores on the XD-Violence in the main paper because it is used in previous works [51,46] for this dataset. In combination with the AUC-PR, we provide the AUC-ROC scores in Table 2, which are collected under the same setting. It can be seen that our method achieves the highest AUC-ROC scores among the weakly supervised methods under all settings, and the conclusion using two metrics are consistent.

Table 2: AUC-ROC (%) results on XD-Violence for anomaly frame detection with various number of **seen anomaly classes**.

No. seen anomaly	1	2	3	4
Wu *et al.* [51](off-line)	67.05	71.88	73.06	85.32
Wu *et al.* [51](on-line)	66.13	72.32	72.49	83.49
RTFM [46]	66.54	70.78	76.70	82.41
Ours	**72.50**	**77.51**	**84.57**	**88.25**

C.2 Additional Ablation Study

In Table 3, we provide additional ablation study results on the XD-Violence under the close set setting. For the ablation study, the NFs and NFs (w/o Triplet) denote using the NFs to score a sample during testing, and NFs (w/o Triplet) mean that we remove the Triplet loss. To explore the impact of feature encoder,

Table 3: Ablation study results for anomaly frame detection on XD-Violence in close-world Setting (NUM ANOMALY=ALL).

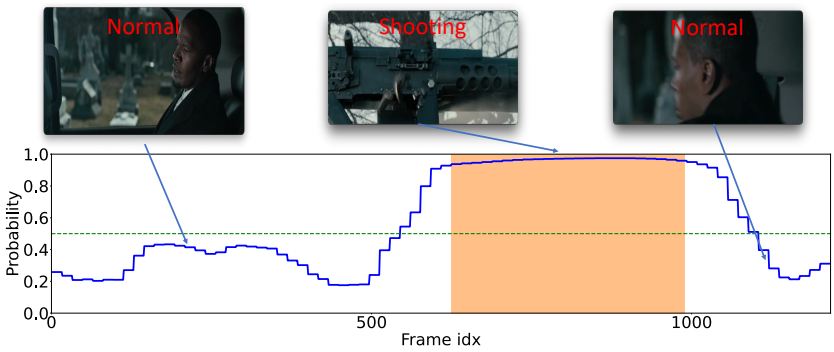
	AUC-PR	AUC-ROC
Wu *et al.* [51](off-line)	75.80	93.07
Wu *et al.* [51](on-line)	72.92	92.02
RTFM [46]	69.40	88.09
NFs(w/o Trip)	52.10	77.40
NFs	73.13	89.77
Ours(w/o GCNs)	69.39	89.14
Ours(w top-k)	77.43	92.66
Ours	77.91	93.23

we replace the GCNs with two FC layers, denoted as the Ours (w/o GCNs). Finally, we provide results of the top-k selection by setting $\Omega = \{\mathbf{x}_i | p_i > \tau_p\}$.

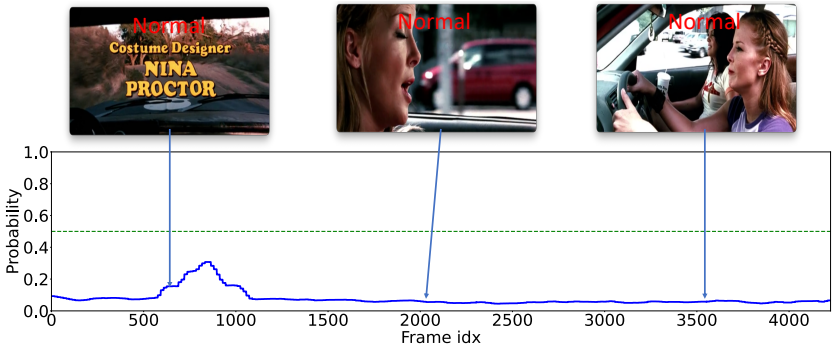
Table 3 shows the results of two weakly supervised baselines, NFs, and ours, under a close set setting. To use the NFs for anomaly detection, we leverage its density estimation capability to score a sample, *i.e.,* a sample with low density is considered to be likely to be an anomaly, similar to the usage of NFs with Cho *et al.* [9]. Results show that the triplet loss contributes a lot to the performance of NFs, proving its important role in facilitating the learning process of NFs (See the NFs vs NFs w/o Trip). Besides, when the GCNs is equipped with triplet loss for representation learning, NFs can achieve comparable performance with the Wu *et al.* [51] and RTFM [46]. Nevertheless, our approach outperforms the NFs by a large margin, justifying the advantage of our usage of NFs over the previous use (pseudo anomaly generation vs density estimation).

Results also show that the choice of feature encoder significantly impacts the anomaly detector; the performance drops a lot when replacing the GCNs with FC layers (See ours vs ours w/o GCNs). We also compare our evidence-based instance selection with the top-k strategy. Based upon top-k, which solely uses the predicted probability p_+ to perform selection, our instance selection method adds the evidence α_+ to improve its robustness. The relation between p_+ and α_+ is determined by $\mathbb{E}[p_+] = \frac{\alpha_+}{\alpha_+ + \alpha_-}$, where p_+ and α_+ denote the probability of being positive and evidence of supporting a positive prediction, respectively. After acquiring the evidence, we use τ_α to filter out samples that are likely the false anomaly. Comparison between ours and top-k shows that adding the evidence could improve the robustness of the latter. We remark that existing literature also uses u to estimate the predictive uncertainty rather than using α_+. However, using α_+ achieves similar, but superior, effect compared with u because u is upper bounded by $\frac{2}{\tau_\alpha}$: $u = \frac{2}{\alpha_+ + \alpha_-} < \frac{2}{\alpha_+} \leq \frac{2}{\tau_\alpha}$. Among the samples with low u, using α_+ would prefer the desired confident anomaly ones.

D Qualitative Results



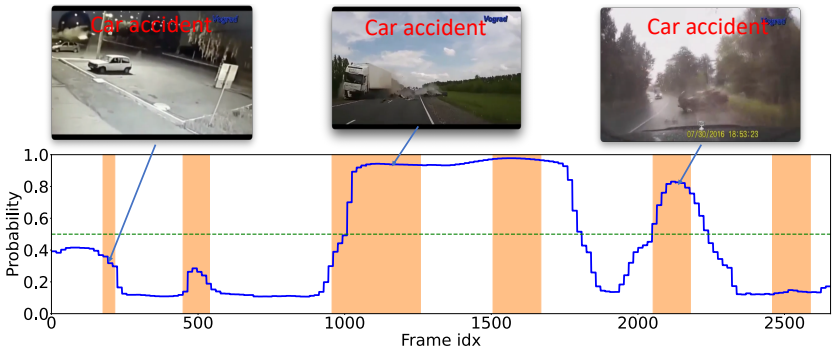
(a) Shooting



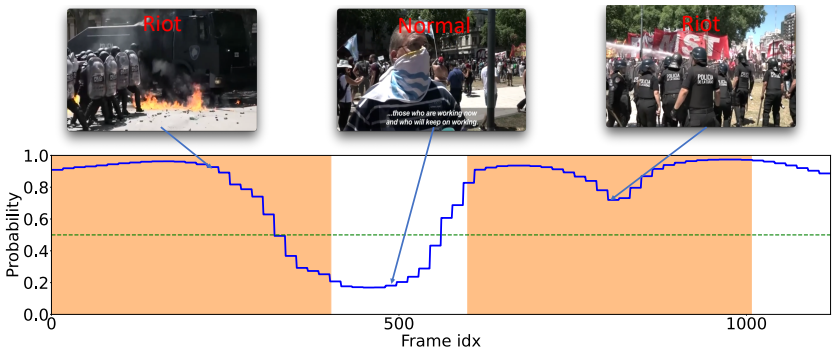
(b) Normal

Fig. 1: Visualized results on XD-Violence for **seen** anomaly frame detection in (a) a *Shooting* video and (b) a normal video. The top row in each example shows raw frames from the video, and the bottom row shows the predicted anomaly score (blue curve) with ground-truth anomaly regions (orange window). Model is trained with *Fighting, Shooting, Abuse, Explosion* and normal videos. *Riot* and *Car accident* are set aside as unseen anomalies.

We plot the results of a model trained with 4 types of anomalies on the XD-Violence dataset. Figure 1 shows that our model fully captures the anomaly region (*i.e.,Shooting*) as they have been seen during training. For the unseen anomaly frames, which are more challenging, Figure 2 shows that our model performs well on detecting them, especially the *Riot*. Our model misses in detecting some *Car accident* events as they last briefly. We also note that our model gives relatively high anomaly scores to some normal frames in anomaly videos, but the margin between anomaly and normal ones is still noticeable. This can



(a) Car accident



(b) Riot

Fig. 2: Visualized results on XD-Violence for **unseen** anomaly frame detection in (a) a *Car Accident* video and (b) a *Riot* video. The top row in each example shows raw frames from the video, and the bottom row shows the predicted anomaly score (blue curve) with ground-truth anomaly regions (orange window). Model is trained with *Fighting, shooting, Abuse, Explosion* and Normal videos. *Riot* and *Car Accident* are set aside as unseen anomalies.

be explained that these frames show a sign of violence and are ambiguous, while they are labelled as normal by the human annotator. These observations validate the effectiveness of our model for the proposed OpenVAD task, *i.e.,* detecting arbitrary anomalies.