Supplementary Material: Cross-Domain Learning for Video Anomaly Detection with Limited Supervision

Yashika Jain[1], Ali Dabouei[2⋆], and Min Xu[2⋆]

[1] University of Delhi
[2] Carnegie Mellon University
yashikajain201@gmail.com, ali.dabouei@gmail.com, mxu1@cs.cmu.edu

6 Revisiting Multiple Instance Learning

Since acquiring frame-level labels requires significant time and effort, following Sultani *et al.* [6], we use Multiple Instance Learning (MIL) to train the classifiers using weakly-supervised video-level labels. By dividing a video (bag) into multiple temporal non-overlapping segments (instances) and encouraging anomalous video segments to have higher anomaly scores as compared to the normal segments, they formulate anomaly detection as a regression problem.

The multiple instance ranking objective function is given by:

$$\max_{\substack{X^i \in \mathcal{D}_l^a \\ 1 \le j \le n_s}} F(X^{i,j}|\theta) > \max_{\substack{X^i \in \mathcal{D}_l^n \\ 1 \le j \le n_s}} F(X^{i,j}|\theta), \tag{1}$$

where $\mathcal{D}_l^a = \{(X, Y) \in \mathcal{D}_l : Y = 1\}$ and $\mathcal{D}_l^n = \{(X, Y) \in \mathcal{D}_l : Y = 0\}$ are the set of abnormal and normal videos, respectively and max is taken over all video segments in a bag.

Instead of ranking every segment of the positive and negative bags, ranking is enforced on one segment from each bag, having the highest anomaly score. The overall loss function, $\mathcal{L}_{\text{rank}}$, for a pair of abnormal and normal videos, is given by:

$$\begin{aligned} \mathcal{L}_{\text{rank}} = \max(0, 1 - \max_{\substack{X^i \in \mathcal{D}_l^a \\ 1 \le j \le n_s}} F(X^{i,j}|\theta) + \max_{\substack{X^i \in \mathcal{D}_l^n \\ 1 \le j \le n_s}} F(X^{i,j}|\theta)) \\ + \lambda_1 \mathcal{L}_{\text{Ts}} + \lambda_2 \mathcal{L}_{\text{Sp}}, \end{aligned} \tag{2}$$

where $\mathcal{L}_{\text{Ts}}$ is the temporal smoothness constraint, and $\mathcal{L}_{\text{Sp}}$ is the sparsity constraint.

7 Datasets

UCF-Crime [6]: This is a large-scale VAD dataset having a total duration of 128 hours. It contains long and untrimmed real-world surveillance videos across 13 realistic anomaly categories that are specifically chosen due to their significant impact on public safety. The dataset comprises 1610 weakly-labeled training videos and 290 test

⋆ Corresponding authors.

videos annotated at the frame level.
XD-Violence (XDV) [10]: This is a large-scale and multi-scene audio-visual dataset for violence detection, having a total duration of 217 hours. Its long and untrimmed videos are collected from movies, games, and in-the-wild scenarios, with anomalies spread over 6 categories. It comprises 3954 weakly-labeled training videos and 800 test videos annotated at the frame level.
HACS [12]: This is a large-scale dataset for human action recognition, sourced from YouTube. It features 200 action classes across 140K segments on 50K videos. Due to its diverse range of actions, larger size, and longer video durations compared to other video datasets such as UCF-101, Kinetics, and ActivityNet, we use a subset of 11K videos from HACS Segments as external, unlabeled data.

8 Implementation Details

To ensure consistency and gradient stability, while training on $\mathcal{D}_l \cup \mathcal{D}_u$, each mini-batch consists of an equal number of samples from $\mathcal{D}_l$ and $\mathcal{D}_u$. Since the computation of $\mathcal{L}_{\text{rank}}$ necessitates pairs of abnormal and normal videos, each labeled sample within the mini-batch comprises a pair of anomalous and normal videos. All the experiments were conducted on an NVIDIA RTX A5000 24 GB GPU. For the experiments using UCF-Crime as the weakly-labeled data, we set the batch size to 64, and for the experiments using XD-Violence as the weakly-labeled data, we set the batch size to 32. In all our experiments except the open-set, we set n_s to 64, τ to 1.25, λ_1 to 5e-3, λ_2 to 1e-3, λ_3 to 1e-3. We set λ_4 to 2000 for UCF+HACS and UCF+XDV, 1250 for XDV+HACS, and 700 for XDV+UCF. For all our experiments, we use the Adam optimizer with a weight decay of 1e-3. For the fully connected layers, we use a learning rate of 5e-4 when UCF-Crime is used as the weakly-labeled dataset and a learning rate of 1e-4 when XDV is used as the weakly-labeled dataset. For the transformer encoder layers, we use a learning rate of 3e-5 when UCF-Crime is used as the weakly-labeled dataset and a learning rate of 5e-5 when XDV is used as the weakly-labeled dataset. In all our experiments, we explicitly encode positional information in the segments using sinusoidal positional encodings [8]. We train on the weakly-labeled source dataset for 200 epochs, followed by training on the union of weakly-labeled and external datasets for 40 CDL steps, each CDL step comprising 4 epochs. Due to the finer granularity and semantic richness inherent in CLIP features, we choose to use CLIP features during inference.

9 Comparison with Unsupervised Baselines in Open-Set Settings

Table 6 depicts that the proposed method outperforms all the baselines in open-set settings on the UCF-Crime dataset by a large margin. As expected, all the weakly-supervised methods outperform the unsupervised methods, even when a small subset of the data is used for weakly-supervised training. This highlights the necessity of incorporating weak labels during training. Since a direct comparison of the proposed weakly-supervised framework with unsupervised methods is not fair, we did not include unsupervised baselines in Table 4.

Table 6: Comparison with prior works in open-set setting on UCF-Crime dataset; c denotes the number of anomalous classes included for weakly-supervised training. The values represent AUC (%).

	c	**0**	**1**	**3**	**6**	**9**
Unsup.	Conv-AE [2]	50.60	-	-	-	-
	Sohrab *et al*. [5]	58.50	-	-	-	-
	Lu *et al*. [4]	65.51	-	-	-	-
	BODS [9]	68.26	-	-	-	-
	GODS [9]	70.46	-	-	-	-
Weakly-Sup.	Wu *et al*. [10] (offline)	-	73.22	75.15	78.46	79.96
	Wu *et al*. [10] (online)	-	73.78	74.64	77.84	79.11
	RTFM [7]	-	75.91	76.98	77.68	79.55
	Zhu *et al*. [13]	-	76.73	77.78	78.82	80.14
	Ours (w/o CDL)	-	75.17	81.51	82.97	83.02
	Ours	-	77.45	82.57	83.44	83.37

10 Comparison of the Original and Proposed Annotations for UCF-Crime Dataset

Figure 4 illustrates a subset of instances from the UCF-Crime's test set where the original annotations do not label frames as anomalous, despite their actual anomalous nature. We also provide a comparison of the proposed and original annotations superimposed on the videos at this link: `https://rb.gy/4vkr1r`.

11 Limitations

Similar to some recent weakly-supervised VAD works [1, 3, 11], the training process of the proposed CDL framework involves two stages. Consequently, the training does not operate in an end-to-end manner. This incurs additional complexity and challenges for training the model in real-world applications. However, since the generalization obtained using this multi-stage training is significant, the complex training setup of the multi-stage framework is reasonable. Nonetheless, developing end-to-end training frameworks would be an important direction for future research. This can facilitate the advancement of anomaly detection approaches for real-world applications, particularly the ones with limited training budgets.

Additionally, the cross-domain performance in case of drastic distribution shifts between the source and target domains may be hindered. For instance, a model primarily trained on videos from stationary surveillance cameras may not effectively work on videos with rapidly evolving scenes from car dashcams. This is mainly because the uncertainty-based reweighing approach in our framework aims to select samples from the external set that are similar to the source domain. In case of drastic shifts between the two domains, finding informative samples from the target domain would not be trivial.

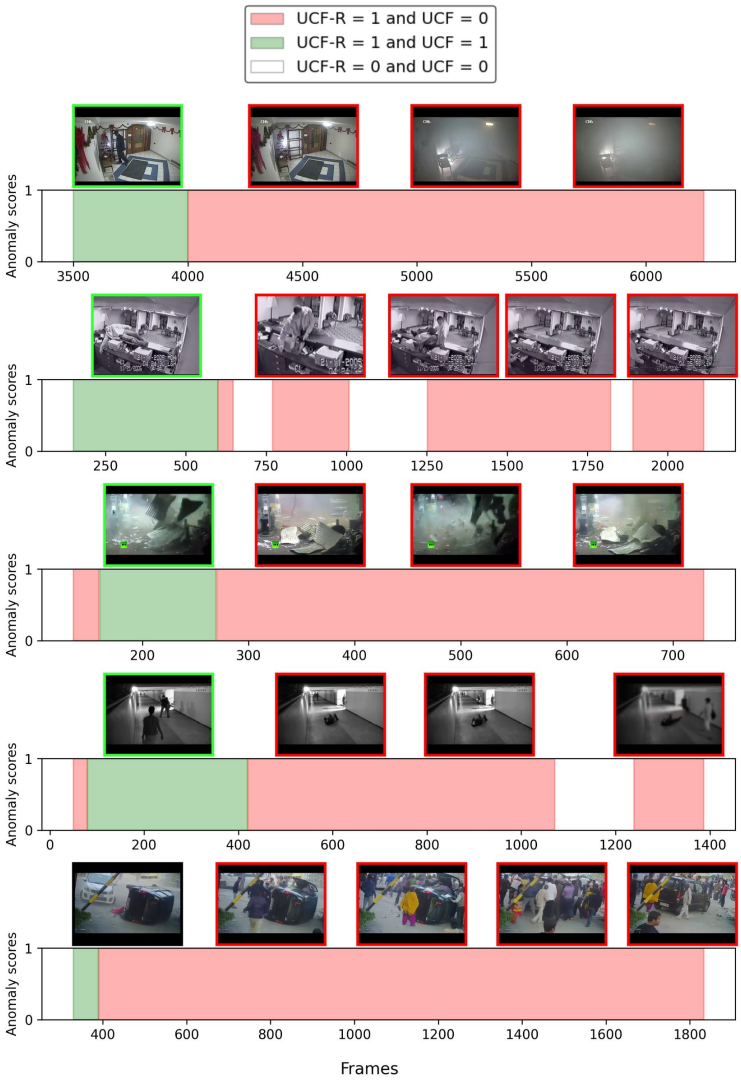

Fig. 4: A comparison between the original annotations (UCF) and the proposed annotations (UCF-R). The green region represents frames labeled as anomalous by both the original and proposed annotations. The red region indicates frames labeled as anomalous by the proposed annotations but not by the original annotations. The unshaded (white) region denotes normal frames. For instance, in the first row, while the original annotations just label frames depicting arson (a person setting the Christmas tree on fire) as anomalous, UCF-R also labels the frames depicting the fire and smoke following arson as anomalous.

References

1. Feng, J.C., Hong, F.T., Zheng, W.S.: MIST: Multiple instance self-training framework for video anomaly detection. In: CVPR. pp. 14009–14018 (2021)

2. Hasan, M., Choi, J., Neumann, j., Roy-Chowdhury, A.K., Davis, L.: Learning temporal regularity in video sequences. In: Proceedings of IEEE Computer Vision and Pattern Recognition (2016)
3. Li, S., Liu, F., Jiao, L.: Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. AAAI pp. 1395–1403 (2022)
4. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in matlab. In: ICCV. pp. 2720–2727 (2013)
5. Sohrab, F., Raitoharju, J., Gabbouj, M., Iosifidis, A.: Subspace support vector data description. In: ICPR. pp. 722–727 (2018)
6. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: CVPR. pp. 6479–6488 (2018)
7. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: ICCV. pp. 4975–4986 (2021)
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS. p. 6000–6010 (2017)
9. Wang, J., Cherian, A.: Gods: Generalized one-class discriminative subspaces for anomaly detection. In: ICCV. pp. 8200–8210 (2019)
10. Wu, P., Liu, j., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z.: Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: ECCV (2020)
11. Zhang, C., Li, G., Qi, Y., Wang, S., Qing, L., Huang, Q., Yang, M.H.: Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In: CVPR. pp. 16271–16280 (2023)
12. Zhao, H., Torralba, A., Torresani, L., Yan, Z.: Hacs: Human action clips and segments dataset. In: ICCV. pp. 8668–8678 (2019)
13. Zhu, Y., Bao, W., Yu, Q.: Towards open set video anomaly detection. In: ECCV. pp. 395–412 (2022)