Cross-Domain Learning for Video Anomaly Detection with Limited Supervision

Yashika Jain
University of Delhi
yashikajain201@gmail.com

Ali Dabouei *
Carnegie Mellon University
ali.dabouei@gmail.com

Min Xu *
Carnegie Mellon University
mxu1@cs.cmu.edu

Abstract

Video Anomaly Detection (VAD) automates the identification of unusual events, such as security threats in surveillance videos. In real-world applications, VAD models must effectively operate in cross-domain settings, identifying rare anomalies and scenarios not well-represented in the training data. However, existing cross-domain VAD methods focus on unsupervised learning, resulting in performance that falls short of real-world expectations. Since acquiring weak supervision, i.e.*, video-level labels, for the source domain is cost-effective, we conjecture that combining it with external unlabeled data has notable potential to enhance cross-domain performance. To this end, we introduce a novel weakly-supervised framework for Cross-Domain Learning (CDL) in VAD that incorporates external data during training by estimating its prediction bias and adaptively minimizing that using the predicted uncertainty. We demonstrate the effectiveness of the proposed CDL framework through comprehensive experiments conducted in various configurations on two large-scale VAD datasets: UCF-Crime and XD-Violence. Our method significantly surpasses the state-of-the-art works in cross-domain evaluations, achieving an average absolute improvement of 19.6% on UCF-Crime and 12.87% on XD-Violence.*

1. Introduction

Video anomaly detection (VAD) aims to locate anomalous events in the videos [3, 10, 11, 15, 21, 25, 32, 33, 42, 47]. Unlike manual surveillance, which is costly and time-consuming, video anomaly detection eliminates the need for extensive human effort, saving resources and time. It holds significant potential for playing a vital role in video surveillance by identifying unusual behaviors and activities such as accidents, burglaries, explosions, and other events that signal security threats.

VAD has been extensively studied previously [11, 15, 21, 32, 33, 47]. Owing to the high costs and time associated with obtaining frame-level labels, most approaches formulate the problem as either an unsupervised [10, 15, 21] or weakly-supervised learning setup [11, 32, 33]. In the unsupervised or one-class classification-based) learning setup, only normal videos are used to model the underlying distribution of normal spatiotemporal patterns, and any deviations from the modeled distribution are regarded as anomalies. Despite the convenience of the unsupervised setup, the lack of anomalous videos during training limits the model's ability to learn the specific characteristics of anomalies. This results in limited performance which does not meet real-world expectations. To address this issue, weakly-supervised setup has attracted significant attention. In this setup, merely video-level labels indicating the presence of anomalies within the videos are incorporated as weak supervision to train models capable of making frame-level predictions at inference. Multiple Instance Learning (MIL) [32] is a prominent technique in this domain. By treating each video as a "bag" and each snippet as a "segment", MIL-based algorithms operate under the premise of a worst-case scenario where the segment with the highest predicted probability of being abnormal is considered as the candidate to represent the whole video.

In real-world applications, it is inevitable to encounter environments and scenarios not fully represented in the model's training set. However, it is essential that the model makes correct predictions in such novel situations. For in-

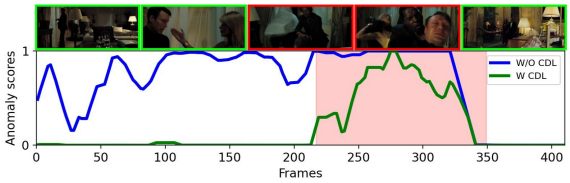

Figure 1. Anomaly score comparison on a video of XD-Violence dataset, with and without employing the proposed CDL framework. The model trained without CDL on UCF-Crime as the weakly labeled set consistently yields high anomaly scores. In contrast, the model trained with CDL, using UCF-Crime as the weakly labeled set and HACS as the unlabeled set, is better able to localize the anomalous frames.

*Corresponding authors.

stance, when the training data lacks samples of rare events like "riots" or accidents in novel scenes, the model should be able to characterize such occurrences as anomalous when they occur. Previous works study these novel situations under the cross-domain problem definition [3, 13, 23].

Existing cross-domain VAD methods [3, 13, 23, 25] rely on unsupervised techniques and consequently exhibit limited performance, as demonstrated later in our empirical evaluations in Tables 2 and 3. A solution to this could be the adoption of weakly-supervised techniques for cross-domain VAD. While weakly-supervised approaches have proven promising in single-domain scenarios [11, 32, 33], their effectiveness in cross-domain scenarios has not been extensively explored. Our evaluations in Tables 2 and 3 suggest that directly employing existing weakly-supervised methods to address the cross-domain challenges results in a significant performance drop when tested in scenarios of even similar nature, such as surveillance videos. We argue that this performance gap is due to the following reasons. First, anomalous events, by their very nature, lack a specific pattern or predefined structure. Hence, the definition of anomaly is context-dependent and a naive adaptation of the previous method cannot capture the context-dependencies in multiple domains. Second, anomalous events are relatively infrequent, making VAD a class imbalance problem. This issue becomes more severe when dealing with multiple domains. Third, because of the limited amount of weakly labeled training data, the model's learning capacity to detect novel (open-set) anomalies is also constrained. Due to these challenges, weakly-supervised methods cannot be readily applied to cross-domain or cross-dataset scenarios.

To overcome these challenges and develop a generalized VAD model, substantial amounts of weakly-labeled data are required. However, acquiring even video-level labels for a large number of videos is inefficient and labor-intensive. On the other hand, vast streams of unlabeled videos are generally available. Utilizing the limited weakly-labeled data alongside this abundant unlabeled data provides a notable opportunity to address the aforementioned challenges in cross-domain VAD. Prudent utilization of the unlabeled data can provide valuable insights into the underlying data distribution, leading to improved decision-making and identification of anomalous events.

To this end, we propose a weakly-supervised *Cross-Domain Learning (CDL) framework* for VAD that integrates external, unlabeled data, from the wild with limited weakly-labeled data to provide competitive generalization across the domains. This is achieved by adaptively minimizing the prediction bias over the external data using the estimated prediction variance, which serves as an uncertainty regularization score. In the proposed framework, we first train fine-grained pseudo-label generation models on the weakly-labeled data to obtain sets of segment-level predictions for the external dataset. Second, we compute the variance of the predictions across multiple predictors as a proxy to represent uncertainty associated with the segments in the external data. Third, during the optimization process, involving training on both labeled and external data, we adaptively reweigh the bias on each external data using the uncertainty regularization scores. This dynamic reweighing ensures that segments from the external dataset closer to the source dataset are emphasized during the training, while those with higher uncertainty are down-weighted. Finally, we iteratively regenerate pseudo-labels using the models trained on labeled and pseudo-labeled data, re-estimate the uncertainties, and re-train the model on the union of labeled and external datasets. This iterative process helps refine the pseudo-labels as the training progresses. With this training process, the model learns to generalize to both source and external data, given only supervision on the source data. Figure 1 illustrates the effectiveness of the CDL framework.

Method(s)	Sup. on $\mathcal{D}$	Target
Acsintoae *et al.* [1]	unsupervised	$\mathcal{D}$
rGAN [23], MPN [25]	unsupervised	$\mathcal{D}'$
zxVAD [3]	unsupervised	$\mathcal{D} \cup \mathcal{D}'$
Ours	weakly-supervised	$\mathcal{D} \cup \mathcal{D}'$

Table 1. Brief overview of the taxonomy of current works for VAD using a source domain dataset ($\mathcal{D}$) and a secondary domain dataset ($\mathcal{D}'$). All these methods do not utilize any labels for training on ($\mathcal{D}'$) and assume distinct distributions for $\mathcal{D}$ and $\mathcal{D}'$.

To summarize, we make the following contributions:

- We present a practical CDL framework for weakly-supervised VAD, in which unlabeled external videos are employed to enhance the cross-domain generalization of the model.
- We design a novel uncertainty quantification method that enables the adaptive uncertainty-driven integration of external videos into the training set.
- Through extensive experiments and ablation studies on benchmark datasets, we validate the proposed approach, demonstrating state-of-the-art performance in cross-domain settings while retaining a competitive performance on the in-domain data.

2. Related Works

Video Anomaly Detection (VAD). Video Anomaly Detection (VAD). VAD is a well-established problem, with most works formulating it either as unsupervised learning [15, 21, 22, 41, 44] or weakly-supervised learning [29, 32, 33, 43, 48] problem. In unsupervised setups, the training data consists solely of normal videos, with the majority of works encoding normal patterns through techniques like frame reconstruction [15, 39], future frame pre-

diction [21], dictionary learning [22, 44], and one-class classification [17, 24]. Any deviation from the encoded patterns is considered anomalous. Since the model categorizes anything beyond its learned representations as anomalous, it can label novel video actions and scenarios encountered during training but in altered environments as anomalous. Weakly-supervised VAD methods help mitigate these issues by incorporating video-level labels as weak supervision for the model, with the majority of methods utilizing the Multiple Instance Ranking Loss [11, 32, 35, 47]. Given that a VAD model is expected to encounter previously unseen scenarios during deployment, it is of paramount importance for the model to have a high generalization across domains. Previous works refer this as cross-domain [3] or cross-dataset generalization [9]. We provide an overview of the existing works employing external data in VAD in Table 1. Previous works on cross-domain generalization focus on unsupervised methods based on few-shot target-domain scene adaptation. [23, 25] employ data from the target domain via meta-learning to adapt to that specific domain. Aich *et al.* [3] proposed a zero-shot target domain adaptation method that incorporates external data to generate pseudo-abnormal frames. Despite the intriguing setup, these unsupervised cross-domain generalization methods lack explicit knowledge about what constitutes an anomaly, hindering the model's ability to learn the specific characteristics of anomalies. To this end, we propose the use of weakly-supervised learning for cross-domain generalization. We integrate external datasets from diverse domains to enable the cross-domain generalization of a model trained in a weakly-supervised fashion.

Pseudo-Labeling and Self-training. Pseudo-labeling [4, 28] is a common technique where the model trained on labeled data assigns labels to unlabeled data. Subsequently, the model is trained on both the initially labeled data and the pseudo-labeled data. This self-training strategy [26, 40] operates iteratively, allowing the model to progressively enhance its generalization. In VAD, several works leverage pseudo-labeling and self-training for generating fine-grained pseudo-labels [11, 20, 42]. However, in contrast to the previous methods, instead of generating pseudo-labels for the weakly labeled data, we leverage pseudo-labels for incorporating the external data.

Uncertainty Estimation. To address pseudo-label noise, prior research in different contexts has explored uncertainty estimation using various approaches, such as data augmentation [5, 30], inference augmentation [12], and model augmentation [46]. While data augmentation is effective for images, it can disrupt temporal relationships in video frames and is not efficient for training on high-cardinality data like videos. On the other hand, inference augmentation methods, such as MC Dropout [12, 42], introduce perturbations during model inference to obtain slightly different predictions, but that is inefficient for training with fixed backbones. In contrast, model augmentation uses different models. Since different models may have varying biases and receptive fields, this would result in diverse predictions. This prediction discrepancy can help quantify uncertainty, making model augmentation well-aligned with our problem. To avoid any manual thresholding for learning from pseudo-labels during training, following [16, 46] we use adaptive reweighing of loss with uncertainty values. In [46], Zheng *et al.* quantify uncertainty by estimating discrepancies between predictions made by two classifiers using Kullback–Leibler (KL) divergence. However, given that VAD is a binary classification task, the divergence based on only two outcomes for the posterior probability is not optimally informative. Hence, we propose a method to quantify uncertainty in the high-dimensional feature space instead of the probability space.

3. Method

3.1. Problem Definition

In this work, we address a real-world VAD problem, where a weakly-labeled dataset $\mathcal{D}_l = \{(X_l^i, Y_l^i)\}_{i=1}^{n_l}$ and an external unlabeled dataset $\mathcal{D}_u = \{X_u^i\}_{i=1}^{n_u}$ are available for training. Here, n_l and n_u indicate the number of videos in the two datasets, respectively, with $n_u \gg n_l$ due to the convenience of gathering unlabeled video data. The video-level labels of X_l are denoted by $Y_l \in \{0, 1\}$. We do not make any assumption about distributions of $\mathcal{D}_l$ and $\mathcal{D}_u$, and therefore, they can be drawn from different distributions. We aim to find the model $F(\cdot|\theta)$, parameterized by θ, that provides accurate predictions on weakly-labeled data while adaptively minimizing the prediction bias on the external data using the uncertainty regularization scores. We illustrate the proposed framework in Figure 2.

3.2. Feature Extraction and Temporal Processing

The proposed uncertainty quantification method (Section 3.4) compares two diverse representations of each sample to estimate the uncertainty associated with the segment-level predictions on external data. To this aim, we employ two different backbones for feature extraction from videos, which are widely used for anomaly detection tasks. The first one is the conventional I3D backbone [6], which extracts segment-level features using 3D convolution, and the other is the CLIP backbone [27], which extracts frame-level features using the frozen CLIP Model's ViT encoder. The contrasting inductive biases of the 3D convolution-based I3D and the transformer-based CLIP help to effectively capture the prediction variance. It is to be noted that only the CLIP backbone is used during inference. We develop two prediction heads, namely the main model, P_m, built on top of the CLIP backbone, and the auxiliary model, P_a, built on top

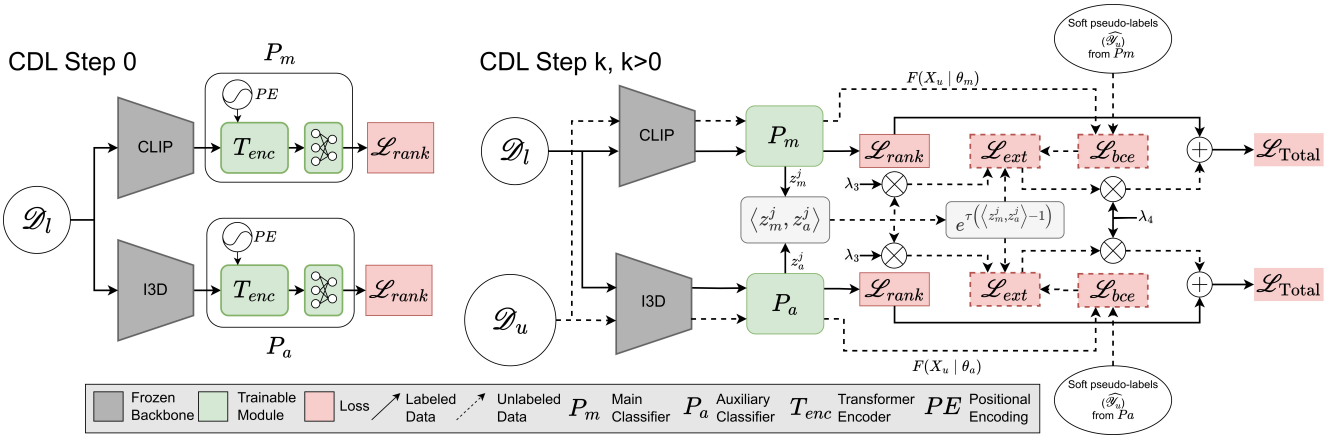

Figure 2. Overview of the proposed CDL Framework. **CDL Step 0:** The Ranking Loss, $\mathcal{L}_{\text{rank}}$ (Supp Mat. §6), is employed to train two pseudo-label generation models, P_m and P_a, §3.2, on weakly-labeled data, $\mathcal{D}_l$. **CDL Step $k, k > 0$**: P_m and P_a are trained iteratively on $\mathcal{D}_l \cup \mathcal{D}_u$, incorporating pseudo-labels for $\mathcal{D}_u$ generated at the end of the previous CDL step. To deal with noise in pseudo-labels, uncertainty regularization scores are estimated using the divergence between the predictions of the two models, §3.4. When optimizing on $\mathcal{D}_u$, the prediction bias, $\mathcal{L}_{\text{bce}}$ (§3.3), for external data is reweighed using the computed uncertainty regularization scores, §3.5.

of the I3D backbone.

Video frames are highly correlated in the temporal dimension. To reduce the redundancy in frame-level features extracted by the CLIP backbone, we pool the representations by bilinearly interpolating them to a fixed, empirically determined length, n_s. Each of the n_s interpolated features represents one segment. To ensure consistency, we also fix the length of representations extracted by the I3D backbone. Evaluation in Section 4.6 analyzes the role of n_s on the model's performance. To capture long-range temporal information over the sequence, we employ a lightweight temporal network, *i.e.*, transformer encoder, to implement P_m and P_a.

3.3. Bias Estimation for External Data

Similar to [46], we formulate the prediction bias on external data as:

$$Bias(\mathcal{D}_u) = \mathbb{E}_{X_u}[F(X_u|\theta) - \mathcal{Y}_u], \tag{1}$$

where $F(X_u|\theta)$ represents a set of predicted probability distributions, each one corresponding to a distinct segment of X_u, and $\mathcal{Y}_u$ denotes the set of unknown segment-level labels of X_u. $Bias(\mathcal{D}_u)$ can be re-written as:

$$Bias(\mathcal{D}_u) = \mathbb{E}_{X_u}[F(X_u|\theta) - \hat{\mathcal{Y}}_u] + \mathbb{E}_{X_u}[\hat{\mathcal{Y}}_u - \mathcal{Y}_u], \tag{2}$$

where $\hat{\mathcal{Y}}_u$ denotes the set of segment-level pseudo-labels for X_u. $\hat{\mathcal{Y}}_u$ can be generated by performing inference on the model trained on $\mathcal{D}_l$. The first term in Equation 2 denotes the difference between the predicted posterior probability and the pseudo-labels, while the second term denotes the error between the pseudo-labels and the ground-truth labels. While minimizing the prediction bias, due to the lack of ground truth supervision, we employ a self-training mechanism, considering $\hat{\mathcal{Y}}_u$ as the soft labels, thereby treating the second term as a constant and minimizing the first term. Specifically, we use the binary cross-entropy (BCE) loss, $\mathcal{L}_{\text{bce}}$, given by:

$$\mathcal{L}_{\text{bce}} = -\hat{\mathcal{Y}}_u \log(F(X_u|\theta)) - (1-\hat{\mathcal{Y}}_u)\log(1-F(X_u|\theta)), \tag{3}$$

to estimate the prediction bias associated with each video segment, for both P_m and P_a.

3.4. Uncertainty Estimation

Since $\mathcal{D}_u$ and $\mathcal{D}_l$ do not necessarily share the same distribution, the generated pseudo-labels are noisy. This noise can adversely affect the subsequent training process as it causes bias to further magnify and propagate within the model. This issue, known as Confirmation Bias [4], is often mitigated by quantifying the uncertainty associated with pseudo-labels and then incorporating this uncertainty into the training process to compensate for the noise. As discussed in Section 2, we opt to address the confirmation bias by computing uncertainty using model augmentation. To quantify uncertainty through model augmentation, following [46], we estimate prediction variance, which is formulated as:

$$Var(\mathcal{D}_u) = \mathbb{E}_{X_u}[(F(X_u|\theta) - \mathcal{Y}_u)^2]. \tag{4}$$

Due to the lack of ground-truth labels, Equation 4 can be approximated as:

$$Var(\mathcal{D}_u) \approx \mathbb{E}_{X_u}[(F(X_u|\theta) - \hat{\mathcal{Y}}_u)^2]. \tag{5}$$

When optimizing the prediction bias in Equation 2, the variance in Equation 5 will also be minimized, potentially re-

sulting in inaccurate quantification of the true prediction variance. To address this, we adopt an alternative approximation, expressed as:

$$Var(\mathcal{D}_u) \approx \mathbb{E}_{X_u}[\big(P_m(X_u|\theta_{P_m}) - P_a(X_u|\theta_{P_a})\big)^2]. \quad (6)$$

Since VAD is a binary classification task, the probability distributions corresponding to each segment have limited support. Consequently, estimating prediction variance using only the predicted anomaly scores, as in Equation 6, may not be robust. Hence, instead of measuring the divergence between the predicted posterior probabilities for the two classes, we propose quantifying pseudo-label uncertainty in the high-dimensional space. To this end, we compute the cosine similarity between the segments in each set of the representations, Z_m and Z_a, obtained from the penultimate layer of P_m and P_a, respectively. Here, $Z_m = \{\mathbf{z}_{\text{m}}^1, \mathbf{z}_{\text{m}}^2, \ldots, \mathbf{z}_{\text{m}}^{n_s}\}$ and $Z_a = \{\mathbf{z}_a^1, \mathbf{z}_a^2, \ldots, \mathbf{z}_a^{n_s}\}$.

To obtain a set of stabilized, segment-level uncertainty regularization scores within a bounded range from the computed cosine similarity, we introduce the following function. Let $S = \{\mathbf{s}^1, \mathbf{s}^2, \ldots, \mathbf{s}^{n_s}\}$ be the set of surrogate variances that we use as proxies for the uncertainty of segments. The surrogate variance is computed as:

$$\mathbf{s}^j = e^{\tau(\langle \mathbf{z}_{\text{m}}^j, \mathbf{z}_a^j \rangle - 1)}, \quad (7)$$

where $\mathbf{s}^j$ indicates the uncertainty regularization score for the j^{th} segment, $\langle \mathbf{z}_{\text{m}}^j, \mathbf{z}_a^j \rangle$ indicates the cosine similarity, and τ denotes the temperature parameter.

Higher uncertainty regularization scores indicate the similar encoding of data between the models, implying less uncertainty in the predicted labels, while, lower scores imply high uncertainty in the predicted labels. Empirical evidence in Section 4.4 demonstrates a significant negative correlation between uncertainty regularization scores and Binary Cross-Entropy (BCE) loss between the predicted labels and ground truths. This affirms that the proposed uncertainty regularization score effectively serves as a proxy for the quality of pseudo-labels.

3.5. Training Process

CDL Step 0. We initially train P_m and P_a separately on the labeled set, optimizing both of them using the Ranking Loss, $\mathcal{L}_{\text{rank}}$, discussed in Supp. Mat. Sec. 6. We then perform inference on the trained models to generate the sets of soft segment-level pseudo-labels for training on $\mathcal{D}_u$.

CDL Step $>$ 0. Following the generation of the sets of pseudo-labels for $\mathcal{D}_u$, we enter an iterative pseudo-label refinement phase, where we train P_m and P_a on $\mathcal{D}_l \cup \mathcal{D}_u$ for multiple CDL steps. Each CDL step comprises a fixed number of epochs. In each epoch, we regenerate the sets of segment-level uncertainty regularization scores. To enable the uncertainty-driven learning from external data, similar to [46], we use the estimated uncertainty regularization scores, S, as automatic thresholds as this dynamically adjusts learning from noisy labels by scaling the prediction bias associated with external data based on S. This helps filter out unreliable predictions while prioritizing highly confident predictions. To encourage lower prediction variance, which would in turn lead to increased pseudo-label quality, we explicitly add the prediction variance to the optimization objective corresponding to the external data, $\mathcal{L}_{\text{ext}}$, as:

$$\mathcal{L}_{\text{ext}} = \mathbb{E}_{X_u}[\frac{1}{Var(\mathcal{D}_u)} \cdot Bias(\mathcal{D}_u) + Var(\mathcal{D}_u)]. \quad (8)$$

Equation 8 is rewritten with the approximated terms as:

$$\mathcal{L}_{\text{ext}} = \mathbb{E}_{X_u}[S \cdot \mathcal{L}_{\text{bce}} - \lambda_3 \cdot \langle Z_m, Z_a \rangle]. \quad (9)$$

Alternatively, Equation 9 can be rewritten as:

$$\mathcal{L}_{\text{ext}} = \frac{1}{n_s \cdot n_u} \sum_{i=1}^{n_u} \sum_{j=1}^{n_s} \left(S^{i,j} \cdot \mathcal{L}_{\text{bce}}^{i,j} - \lambda_3 \cdot \langle \mathbf{Z}_{\text{m}}^{i,j}, \mathbf{Z}_a^{i,j} \rangle \right), \quad (10)$$

where λ_3 is a hyper-parameter to balance the losses. Similar to CDL step 0, to optimize the training on $\mathcal{D}_l$, we use $\mathcal{L}_{\text{rank}}$. The total optimization objective for training on $\mathcal{D}_l \cup \mathcal{D}_u$ can be expressed as:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{rank}} + \lambda_4 \cdot \mathcal{L}_{\text{ext}}, \quad (11)$$

where λ_4 is a trade-off parameter for $\mathcal{L}_{\text{ext}}$. We employ the optimization objective defined in Equation 11 during training on $\mathcal{D}_l \cup \mathcal{D}_u$ for each epoch within every CDL step. After each CDL step is completed, we re-generate the set of soft segment-level pseudo-labels using the models trained on $\mathcal{D}_l \cup \mathcal{D}_u$. This iterative refinement process repeats k times, where k is a hyper-parameter determining the number of CDL steps. With each CDL step, the models' performance gets further refined as the pseudo-labels get iteratively improved.

3.6. Inference - Extending Segment-level Scores to Frame-level Scores

During inference, we compute segment-level anomaly scores for the videos using P_m. Since we encounter long-untrimmed videos with varying numbers of frames, for extending the segment-level anomaly score to the frame level, for each video, we divide the total number of frames n_f by the number of segments n_s to obtain the number of frames per segment, n_{fs}. We assign the anomaly score of each segment to its consecutive frames. The first segment corresponds to the first n_{fs} frames, and so forth until the $(n_s - 1)^{th}$ segment. For the last segment, its anomaly score is assigned to any remaining frames, potentially exceeding n_{fs}, if there is a remainder.

4. Experiments

We evaluate the proposed method on the major video anomaly datasets, UCF-Crime (UCF) [32] and XD-Violence (XDV) [38]. Additionally, we use 11,000 videos from the HACS [45] dataset as a source of external data. We provide detailed information about the datasets in Supp. Mat. §7. In §4.1, we discuss the implementation details. In §4.2, we discuss the inherent noise in the test annotations of benchmark datasets. We proceed to compare the proposed framework with prior works in cross-domain scenarios (§4.3.1) and open-set scenarios (§4.3.2). Subsequently, in §4.4, we demonstrate a strong correlation between the quality of pseudo labels and the computed uncertainty scores. We then explore the evolution of these uncertainty scores through the training process in §4.5. Finally, in §4.6, we conduct ablation studies and hyper-parameter analysis to analyze the impact of individual components of the proposed framework.

4.1. Implementation Details

We implement the proposed method using PyTorch. We extract CLIP and I3D features at a fixed frame rate of 30 FPS. CLIP features are extracted from the frozen CLIP model's image encoder (ViT-B/32). For the hyper-parameters, in the open-set scenarios, we empirically set the value of n_s to 64, τ to 1.25, λ_1 and λ_2 to $5e-4$, λ_3 to $1e-3$, and λ_4 to 700. Ablation studies for selecting n_s and λ_3 are included in Section 4.6. We use the Adam optimizer with a weight decay of $1e-3$, and we set a learning rate of $3e-5$ for the transformer encoder and $5e-4$ for the fully connected layers. We use a batch size of 64. In both P_m and P_a, we explicitly encode positional information in the segments using sinusoidal positional encodings [34]. 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. Additional information regarding hyper-parameters is provided in Supp. Material Section 8.

Model Architecture. Both P_m and P_a consist of a transformer encoder layer with four heads, followed by four fully connected layers, each consisting of 4096, 512, 32, and 1 neurons, respectively. In both the models, for all the layers except the last, we use ReLU [2] activation while for the last layer, we use Sigmoid activation.

Evaluation Setup. To reduce bias, we perform each experiment three times with different seeds and average the results. In open-set experiments, we repeat each experiment three times, using different sets of anomaly classes each time.

Evaluation Metric. Following previous works on UCF-Crime [32], we adopt the frame-level area under the ROC curve (AUC) to evaluate on UCF-Crime. In line with previous works on XD-Violence [38], we use the frame-level area under the Precision-Recall curve (PRAUC), also known as Average Precision (AP), to evaluate on XDV.

	Methods	Features	UCF AUC(%)	UCF-R AUC(%)	XDV AP(%)
Cross-Domain (Unsup.)	rGAN [23]	-	64.35*	65.19*	37.74*
	MPN [25]	-	65.67*	67.98*	38.89*
	zxVAD [3]	-	68.74†	69.39†	40.68†
Non Cross-Domain	Sultani *et al.*[32]	I3D	80.70	84.63*	53.88*
	MIST [11]	I3D	82.30	86.17*	50.33*
	RTFM [33]	I3D	84.03	86.47*	37.30*
	S3R [37]	I3D	85.99	87.11*	49.84*
	CU-Net [42]	I3D	86.22	88.15*	37.98*
	MGFN [8]	I3D	86.98	87.33*	32.16*
	SSRL [19]	I3D	87.43	87.02*	51.60*
	CLIP-TSA [18]	CLIP	**87.58**	73.20*	44.33*
	Ours (No ext. data)	CLIP	84.49	89.96	58.13
Cross-Domain (Weakly-Sup.)	Ours (UCF + HACS)	CLIP	84.63	**90.53**	65.14
	Ours (UCF + XDV)	CLIP	84.73	90.26	**68.37**

Table 2. Comparison with prior works on XDV, considering UCF-Crime as the source data. Asterisk (∗) indicates that evaluations were conducted by us using the official code. Dagger (†) indicates that evaluations were conducted by our implementation due to the lack of an official implementation.

	Methods	Features	XDV AP(%)	UCF-R AUC(%)
Cross-Domain (Unsup.)	rGAN [23]	-	40.10*	59.82*
	MPN [25]	-	44.79*	60.35*
	zxVAD [3]	-	47.53†	63.61†
Non Cross-Domain	Sultani *et al.*[32]	I3D	73.20	71.23*
	RTFM [33]	I3D	77.81	70.46*
	MGFN [8]	I3D	80.11	69.12*
	S3R [37]	I3D	80.26	69.04*
	CLIP-TSA [18]	CLIP	**80.67**	67.58*
	Ours (No ext. data)	CLIP	75.13	76.39
Cross-Domain (Weakly-Sup.)	Ours (XDV + UCF)	CLIP	77.04	88.06
	Ours (XDV + HACS)	CLIP	78.61	**88.50**

Table 3. Comparison with prior works on UCF-Crime, considering XDV as the source data. Asterisk (∗) indicates that evaluations were conducted by us using the official code. Dagger (†) indicates that evaluations were conducted by our implementation due to the lack of an official implementation.

4.2. Noise in the Test Annotations of Benchmark Datasets

Our manual inspection reveals that the frame-level testing annotations of the UCF-Crime (UCF) [32] and XD-Violence (XDV) [38] datasets, which are commonly used for benchmarking VAD models, exhibit significant noise. This noise largely stems from the fact that the original annotations do not consistently label the frames leading up to the primary anomalous events and their subsequent consequences as anomalous. For instance, in a video assigned a label like "shooting", we assert that frames showing the person holding the gun and frames illustrating the injured victim should also be marked as anomalous. This perspective aligns with the fundamental goal of VAD, which is to

c	UCF (AUC%)					UCF-R (AUC%)	
	Wu *et al.*[38]	RTFM [33]	Zhu *et al.* [49]	Ours (w/o CDL)	Ours (CDL)	Ours (w/o CDL)	Ours (CDL)
1	73.22	75.91	76.73	75.17	**77.45**	84.32	**85.39**
3	75.15	76.98	77.78	81.51	**82.57**	86.84	**87.69**
6	78.46	77.68	78.82	82.97	**83.44**	87.85	**88.21**
9	79.96	79.55	80.14	83.02	**83.37**	89.22	**89.82**

Table 4. Comparison with other methods in Open-set setting on UCF-Crime dataset; c denotes the no. of anomalous classes included for weakly-supervised training.

identify all anomalous frames within a video, irrespective of the video's primary label. However, it should also be noted that in the original annotations, for some videos, certain frames related to the video's primary anomaly label are also not marked anomalous.

To address this, we re-annotate the test set of UCF-Crime by assigning each video to three independent annotators. We then combine their annotations to generate more accurate frame-level labels. Compared to the original annotations where 7.58% of the total frames are labeled as anomalous, the proposed annotations label 16.55% of the total frames as anomalous. The proposed annotations are available here[1]. We provide a comparison of the proposed and original annotations here[2]. For the remainder of this paper, we refer the re-annotated test set of the UCF-Crime dataset as UCF-R.

4.3. Comparison with Prior Works

4.3.1 Cross-Domain Scenarios

While the UCF-Crime [32] and XD-Violence [38] datasets share similar definitions of what constitutes anomalies, that definition differs from those of smaller datasets like ShanghaiTech [21], CUHK-Avenue [22], UCSD Pedestrian [7], UBnormal [1], where anomalies are more subtle. For instance, running is considered anomalous in UBnormal but not in XD-Violence. Due to these divergent notions of anomalies across datasets, we conduct cross-domain experiments by simultaneously evaluating on the UCF-Crime and XD-Violence datasets, given their more aligned anomaly definitions.

UCF-Crime as the Weakly-Labeled Source Set, XDV as the Cross-Domain Set. Table 2 summarizes the results for this scenario. First, we observe that the proposed method achieves state-of-the-art results on XDV and UCF-R even without utilizing any external data (without CDL). We believe this is due to the inductive bias of previous methods towards the noisy annotations of UCF-Crime. Next, we observe that the addition of external data, HACS and XDV, leads to a significant enhancement in the performance of the cross-domain dataset, XDV, by 11.26% and 14.49%, respectively, compared to the previous state-of-the-art baseline. Additionally, there is also a marginal improvement in the performance of the source set upon integration of external datasets.

XDV as the Weakly-Labeled Source Set, UCF-Crime as the Cross-Domain Set. Table 3 summarizes the results for this scenario. Notably, the proposed method achieves state-of-the-art performance on the cross-domain dataset, UCF-R, even without the utilization of any external data during training. This is attributed to the simplicity of the proposed architecture compared to other baselines. The proposed architecture prevents overfitting to the source dataset, thereby increasing its generalizability to the cross-domain dataset. Additionally, integrating external data further enhances performance on both the cross-domain and source sets. Specifically, leveraging the CDL framework with UCF-Crime and HACS as external datasets boosts UCF-R's AUC by 18.94% and 19.39% respectively, compared to previous state-of-the-art baselines. We also observe that the proposed method's performance is inferior on XDV. We attribute this to the noise in the annotations of XDV's test set.

These results highlight that the proposed CDL framework is capable of effectively exploiting external data with vast domain gaps to achieve a significant cross-domain generalization. It's noteworthy that the performance gain observed with the proposed CDL framework remains consistent across all tested datasets, suggesting that the performance improvement is not dependent on any specific source or external dataset.

4.3.2 Open-Set Scenarios

In Table 4, we evaluate the proposed framework's performance on the UCF-Crime dataset in a realistic open-set scenario, where the model is evaluated on both, previously seen and unseen anomaly classes. To simulate this scenario, we randomly include c anomalous classes in the weakly-labeled set, while the remaining anomalous classes are placed in the unlabeled set. In both the weakly-supervised source set and the unlabeled set, the number of normal videos equals the number of anomalous videos. We evaluate two model configurations; one trained solely on

[1] https://drive.google.com/drive/folders/1IVjQQFHXVcsaT63HUjpfk8C5KH6HsQ7t?usp=drive_link

[2] https://rb.gy/4vkr1r

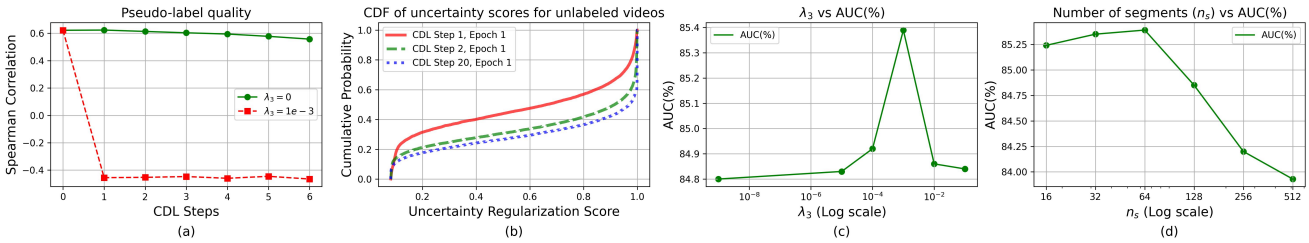

Figure 3. **(a)** Correlation between uncertainty scores and BCE loss computed between the estimated scores and ground truth. When $\lambda_3 = 1e-3$, as expected, a consistently high negative correlation emerges, demonstrating the effectiveness of the proposed uncertainty quantification method as a reliable proxy for pseudo-label quality. **(b)** Cumulative Distribution Function (CDF) plots illustrating the progression of average uncertainty regularization scores for each video during training. CDL step 20 has a higher concentration of scores around 1 compared to CDL step 2, while CDL step 2 has a higher concentration around 1 than CDL step 1. This suggests that, as training progresses, there is a higher tendency for scores to have elevated values, indicating more confident pseudo-label predictions. **(c)** Ablation study on the coefficient of the cosine similarity loss term, λ_3. **(d)** Ablation study on the number of segments, n_s.

the weakly-labeled set (without CDL) and the other on the union of weakly-labeled and unlabeled sets using the CDL Framework.

On UCF-Crime, the proposed model, without CDL, surpasses the state-of-the-art baselines for $c > 1$. This highlights its efficacy in open-set settings. While, with CDL, the model surpasses the baselines across all values of c by a considerable margin.

For both UCF-Crime and UCF-R, when unlabeled data is incorporated, we observe a consistent performance gain across all values of c, suggesting the effectiveness of the CDL framework across varying amounts of weakly-labeled and unlabeled data.

4.4. Correlation between Uncertainty Scores and BCE Loss (Proxy to Label Quality)

To assess the efficacy of the proposed uncertainty quantification method as a proxy for pseudo-label quality, we compute the non-parametric Spearman correlation between estimated uncertainty regularization scores and BCE loss between the predicted pseudo-labels and the corresponding ground truths. For this experiment, we consider UCF-Crime as the weakly-labeled source set and XDV as the external set. In Figure 3(a), with $\lambda_3 = 1e-3$, CDL step 1 onwards, a consistently high negative correlation (-0.46 in CDL step 6, with a p-value $<$ 1e-5) emerges, indicating the robustness of the proposed uncertainty quantification framework. Conversely, setting λ_3 to 0 results in a sustained positive correlation, signifying sub-optimal pseudo-labels in the absence of cosine similarity loss term.

4.5. Progression of Uncertainty Scores

To assess the evolution of uncertainty regularization scores through the training process, in Figure 3(b), we plot the Cumulative Distribution Function (CDF) of average uncertainty regularization scores for external videos across the first epoch of three different CDL steps. We conduct this experiment considering UCF-Crime as the weakly-labeled source set and XDV as the external set. We observe that in CDL step 1, 16.65% of the uncertainty scores fall within the range [0, 0.1]. As training progresses to CDL steps 2 and 20, this proportion decreases to 13.06% and 11.39%, respectively. Meanwhile, the proportion of uncertainty scores in the range [0.9, 1] increases from 35.11% in CDL step 1 to 56.70% in CDL step 2 and further to 57.68% in CDL step 20. This trend indicates a discernible shift towards higher uncertainty scores as training progresses, suggesting an improvement in model confidence due to increased pseudo-label quality.

4.6. Ablation Studies and Hyper-parameter Analysis

For the sake of consistency, we conduct all ablation studies on UCF-Crime in an open-set setting, with $c = 1$. However, it should be noted that for different training setups, hyper-parameters are tuned separately as well.

Impact of Various Components of the CDL Framework. We assess the effectiveness of each component of the CDL framework by adding them sequentially. The results are summarized in Table 5. We consider training on $c = 1$ anomaly class in a weakly-supervised fashion as our baseline. The remaining $c - 1$ anomalous classes are placed in the external set. We first observe that integrating external data into the source set without accounting for pseudo-label uncertainty ($S^{i,j} = 1, \forall i, j$) and without minimizing cosine similarity between representations ($\lambda_3 = 0$) yields a 0.35% gain in AUC, highlighting the effectiveness of external data in improving the model's performance. Next, we study the impact of uncertainty-aware integration of external data, *i.e.*, adaptively reweighing the prediction bias of external data with the computed uncertainty values and with λ_3 set to 0. This results in a gain of 0.13% in AUC,

External data	Uncertainty Coeff.	Cos. Similarity Loss	AUC(%)
✗	✗	✗	84.32
✓	✗	✗	84.67
✓	✓	✗	84.80
✓	✓	✓	85.39

Table 5. Ablation study of various components on the UCF-R dataset in an open-set setting ($c = 1$).

demonstrating the superiority of uncertainty-driven integration compared to the standard integration. Finally, we assess the impact of adding the cosine similarity loss term during uncertainty-aware training. This further leads to a significant boost of 0.59%, validating its effectiveness.

Impact of Cosine Similarity Loss. In Figure 3(c), we explore the impact of varying the coefficient of the cosine similarity loss on the model's performance. We observe a gradual increase in AUC as λ_3 increases from 1e-9 to 1e-3. This could be due to the effect of cosine similarity loss getting more pronounced with higher values of λ_3. However, beyond 1e-3, there is a rapid decline in AUC, likely due to the dominance of the cosine similarity loss over other losses when its coefficient is high. Therefore, we select 1e-3 as the optimal choice for λ_3.

Impact of Number of Segments. In Figure 3(d), we observe that the performance consistently improves as no. of segments, n_s, increases from 16 to 64, but it begins to decline rapidly afterward. Therefore, we set n_s as 64.

Impact of the Size of External Data. To determine the optimal number of unlabeled external videos from the HACS dataset to integrate into the weakly-labeled training set of UCF-Crime, we conduct an ablation study, depicted in Figure 4. We observe that increasing the size of the external set increases the performance on XDV. However, this increase tends to plateau after the inclusion of 11,000 videos. Consequently, we do not include additional videos beyond the 11,000 threshold.

5. Conclusion

In this work, we demonstrated the effectiveness of integrating external, unlabeled data with weakly-labeled source data to enhance the cross-domain generalization of VAD models. To enable this integration, we proposed a weakly-supervised CDL (Cross-Domain Learning) framework that adaptively minimizes the prediction bias on external data by scaling it with the prediction variance, which serves as an uncertainty regularization score. The proposed method outperforms baseline models significantly in cross-domain and open-set settings while retaining competitive performance in in-domain settings.

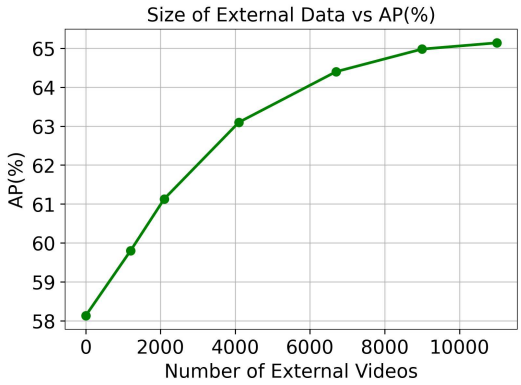

Figure 4. Ablation study on the impact of the size of external data.

Acknowledgement

This work was supported in part by U.S. NIH grants R01GM134020 and P41GM103712, NSF grants DBI-1949629, DBI-2238093, IIS-2007595, IIS-2211597, and MCB-2205148. This work was supported in part by Oracle Cloud credits and related resources provided by Oracle for Research, and the computational resources support from AMD HPC Fund. We thank Eshaan Mandal and Bhavay Malhotra for their assistance, which has been instrumental in completing this work.

References

[1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *CVPR*, 2022. 2, 7

[2] Abien Fred Agarap. Deep learning using rectified linear units (relu), 2019. 6

[3] Abhishek Aich, Kuan-Chuan Peng, and Amit K. Roy-Chowdhury. Cross-domain video anomaly detection without target domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2579–2591, 2023. 1, 2, 3, 6

[4] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *IJCNN*, 2020. 3, 4

[5] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *ICLR*, 2020. 3

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 3

[7] Antoni B. Chan and Nuno Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 909–926, 2008. 7

[8] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, 2023. 6

[9] MyeongAh Cho, Minjung Kim, Sangwon Hwang, Chaewon Park, Kyungjae Lee, and Sangyoun Lee. Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning. In *CVPR*, pages 12137–12146, 2023. 3

[10] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR*, pages 3449–3456, 2011. 1

[11] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. MIST: Multiple instance self-training framework for video anomaly detection. In *CVPR*, pages 14009–14018, 2021. 1, 2, 3, 6, 14

[12] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016. 3

[13] Mariana Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 4505–4523, 2022. 2

[14] Mahmudul Hasan, Jonghyun Choi, jan Neumann, Amit K Roy-Chowdhury, and Larry Davis. Learning temporal regularity in video sequences. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2016. 14

[15] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *CVPR*, 2016. 1, 2

[16] Kexin Huang, Vishnu Sresht, Brajesh Rai, and Mykola Bordyuh. Uncertainty-aware pseudo-labeling for quantum calculations. In *UAI*, pages 853–862, 2022. 3

[17] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *CVPR*, pages 7842–7851, 2019. 3

[18] Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *ICIP*, pages 3230–3234, 2023. 6

[19] Guoqiu Li, Guanxiong Cai, Xingyu Zeng, and Rui Zhao. Scale-aware spatio-temporal relation learning for video anomaly detection. In *ECCV*, pages 333–350, 2022. 6

[20] Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. *AAAI*, pages 1395–1403, 2022. 3, 14

[21] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *CVPR*, pages 6536–6545, 2018. 1, 2, 3, 7

[22] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *ICCV*, pages 2720–2727, 2013. 2, 3, 7, 14

[23] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. Few-shot scene-adaptive anomaly detection. In *ECCV*, pages 125–141, 2020. 2, 3, 6

[24] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *ICME*, pages 439–444, 2017. 3

[25] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *CVPR*, pages 15425–15434, 2021. 1, 2, 3, 6

[26] David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *NAACL*, pages 152–159, 2006. 3

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 3

[28] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *ICLR*, 2021. 3

[29] Hitesh Sapkota and Qi Yu. Bayesian nonparametric submodular video partition for robust anomaly detection. In *CVPR*, pages 3212–3221, 2022. 2

[30] 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. In *NeurIPS*, 2020. 3

[31] Fahad Sohrab, Jenni Raitoharju, Moncef Gabbouj, and Alexandros Iosifidis. Subspace support vector data description. In *ICPR*, pages 722–727, 2018. 14

[32] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, pages 6479–6488, 2018. 1, 2, 3, 6, 7, 12

[33] 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 *ICCV*, pages 4975–4986, 2021. 1, 2, 6, 7, 14

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, page 6000–6010, 2017. 6, 12

[35] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. *ICME*, pages 1–6, 2020. 3

[36] Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *ICCV*, pages 8200–8210, 2019. 14

[37] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representa-

tion for video anomaly detection. In *ECCV*, pages 729–745, 2022. 6

[38] Peng Wu, jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *ECCV*, 2020. 6, 7, 12, 14

[39] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. In *BMVC*, pages 8.1–8.12, 2015. 2

[40] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL*, page 189–196, 1995. 3

[41] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 583–591, 2020. 2

[42] Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *CVPR*, pages 16271–16280, 2023. 1, 3, 6, 14

[43] Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *ICIP*, pages 4030–4034, 2019. 2

[44] Bin Zhao, Fei-Fei Li, and Eric Xing. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR*, pages 3313–3320, 2011. 2, 3

[45] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset. In *ICCV*, pages 8668–8678, 2019. 6, 12

[46] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision (IJCV)*, 2021. 3, 4, 5

[47] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *CVPR*, pages 1237–1246, 2019. 1, 3

[48] Yi Zhu and Shawn D. Newsam. Motion-aware feature for improved video anomaly detection. In *BMVC*, page 270, 2019. 2

[49] Yuansheng Zhu, Wentao Bao, and Qi Yu. Towards open set video anomaly detection. In *ECCV*, pages 395–412, 2022. 7, 14

Cross-Domain Learning for Video Anomaly Detection with Limited Supervision

Supplementary Material

6. Revisiting Multiple Instance Learning

Since acquiring frame-level labels requires significant time and effort, following Sultani *et al.* [32], 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:

$$\max_{\substack{X^i \in \mathcal{D}_l^a \\ 1 \le j \le n_s}} F(X^{i,j}|\theta) > \max_{\substack{X^i \in \mathcal{D}_l^n \\ 1 \le j \le n_s}} F(X^{i,j}|\theta), \tag{12}$$

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:

$$\begin{aligned}\mathcal{L}_{\text{rank}} = \max(0, 1 - \max_{\substack{X^i \in \mathcal{D}_l^a \\ 1 \le j \le n_s}} F(X^{i,j}|\theta) + \max_{\substack{X^i \in \mathcal{D}_l^n \\ 1 \le j \le n_s}} F(X^{i,j}|\theta)) \\ + \lambda_1 \mathcal{L}_{\text{Ts}} + \lambda_2 \mathcal{L}_{\text{Sp}},\end{aligned} \tag{13}$$

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

7. Datasets

UCF-Crime [32]: 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 videos annotated at the frame level.

XD-Violence (XDV) [38]: 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 [45]: 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 [34]. 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

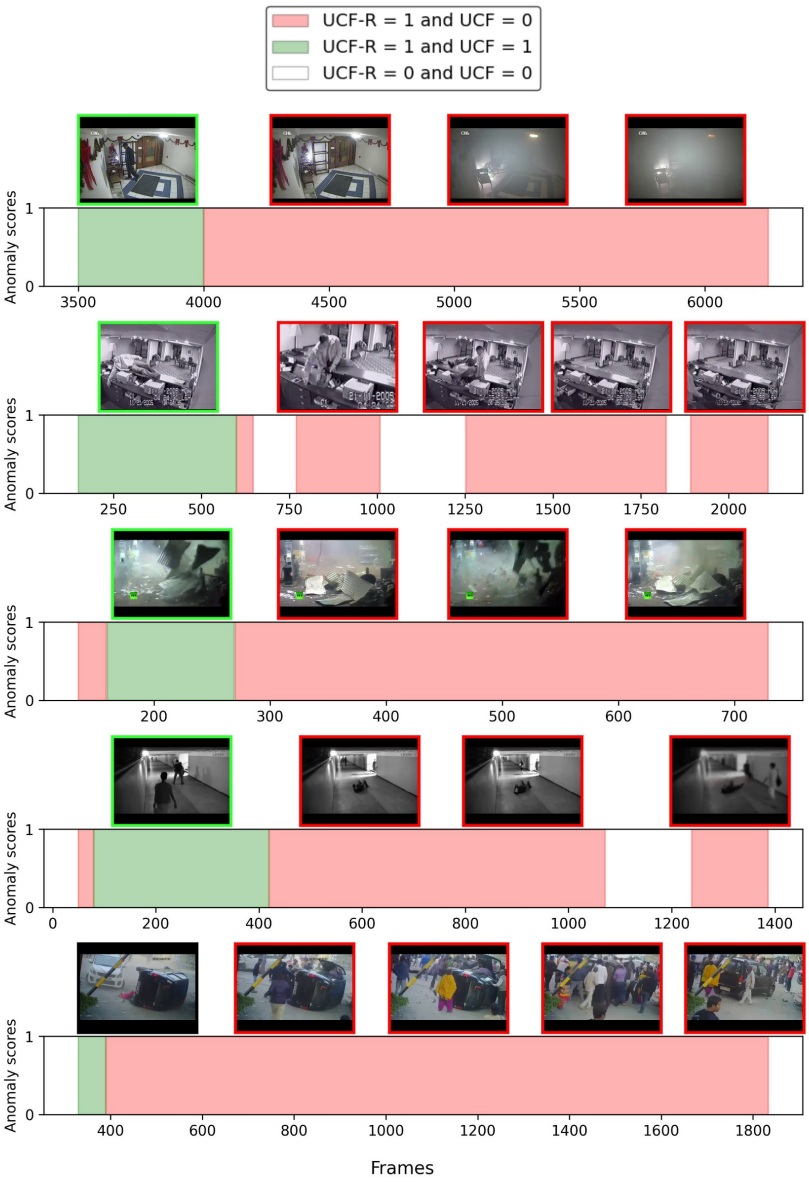

Figure 5. A comparison between the original annotations (UCF) and the proposed annotations (UCF-R). The green region represents frames labeled as anomalous by both the original and proposed annotations. The red region indicates frames labeled as anomalous by the proposed annotations but not by the original 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.

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.

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

Figure 5 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:

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 [14]	50.60	-	-	-	-
	Sohrab *et al*. [31]	58.50	-	-	-	-
	Lu *et al*. [22]	65.51	-	-	-	-
	BODS [36]	68.26	-	-	-	-
	GODS [36]	70.46	-	-	-	-
Weakly-Sup.	Wu *et al*. [38] (offline)	-	73.22	75.15	78.46	79.96
	Wu *et al*. [38] (online)	-	73.78	74.64	77.84	79.11
	RTFM [33]	-	75.91	76.98	77.68	79.55
	Zhu *et al*. [49]	-	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

https://rb.gy/4vkr1r.

11. Limitations

Similar to some recent weakly-supervised VAD works [11, 20, 42], 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.