

NPP-ITA: Normal-Prototype Prior for Unsupervised Video Anomaly Detection with Inference-Time Adaptation

Nayoung Kim*

Dept. of IT Engineering
Sookmyung Women's University
Seoul, Republic of Korea
yeooong@sookmyung.ac.kr

Minjeong Kim*

Dept. of IT Engineering
Sookmyung Women's University
Seoul, Republic of Korea
kmj2110546@sookmyung.ac.kr

Seonghoon Jeong

Division of Artificial Intelligence Engineering
Sookmyung Women's University
Seoul, Republic of Korea
seonghoon@sookmyung.ac.kr

Abstract—Video anomaly detection (VAD) aims to identify abnormal events such as violence, robbery, and explosions in surveillance videos. Existing approaches have often relied on weak supervision or one-class classification, requiring labeled data or normal-only training videos. In this paper, we propose NPP-ITA, an unsupervised video anomaly detection (UVAD) framework that leverages a normal-prototype prior for pseudo-label generation and inference-time adaptation (ITA). We assume that the initial segments of each video reflect the normal context of the corresponding camera environment. Based on this prior, NPP-ITA generates pseudo labels and adapts the model at inference time without ground-truth annotations. Experiments on UCF-Crime and ShanghaiTech show that NPP-ITA provides a practical UVAD framework and maintains competitive anomaly detection performance.

Index Terms—Video anomaly detection (VAD), unsupervised video anomaly detection (UVAD), normal-prototype prior, pseudo-labeling, inference-time adaptation (ITA)

I. INTRODUCTION

With the rapid expansion of smart surveillance cameras, video anomaly detection (VAD) has become essential for recognizing abnormal events in complex real-world scenes. As manual monitoring of large-scale surveillance footage is costly, VAD remains a practical research problem for reducing the burden of manual surveillance [1]–[5]. In particular, surveillance infrastructures in smart cities increasingly consist of large numbers of wirelessly connected cameras, making automated on-device video analytics an important capability for intelligent wireless networks.

Despite recent progress, many VAD methods still rely on video-level annotations or restrictive training assumptions [1], [6], [7]. Since abnormal events are rare and diverse, such annotations are expensive and difficult to obtain [2], [5], [8], [9]. Moreover, real-world surveillance videos vary in camera viewpoint, background, and illumination, which can cause unstable anomaly predictions under distribution shifts [2], [4].

To this end, we propose NPP-ITA, an unsupervised video anomaly detection (UVAD) framework that leverages a normal-prototype prior for both pseudo-label generation and

inference-time adaptation (ITA). The key idea is to exploit the normal-prototype prior: a short sequence of initial segments can provide a practical reference for the normal pattern of the corresponding camera environment. During training, this prior provides a criterion for pseudo-label generation. During inference, the same prior guides lightweight model adaptation. Since adaptation relies only on the few initial segments observed at each camera, NPP-ITA requires transmitting neither raw video nor annotations to a central server, which suits bandwidth-constrained wireless surveillance networks. The contributions of this paper can be summarized as follows:

- We formulate NPP-ITA, a UVAD framework that constructs training and adaptation signals using a normal-prototype prior without ground-truth annotations.
- We design a pseudo-label generator that computes prototype-distance scores from the normal prototype and refines them through a multi-stage refinement pipeline.
- We introduce a lightweight ITA strategy that exploits the same normal-prototype prior for model adaptation during inference without ground-truth annotations.

II. RELATED WORK

A. Video Anomaly Detection (VAD)

VAD aims to detect abnormal events and localize their temporal intervals in surveillance videos [9]. VAD methods can be categorized according to the level of supervision during training: supervised learning, weakly supervised learning, one-class classification, and fully unsupervised learning [2], [8].

Supervised methods use frame- or segment-level annotations, which are costly to construct for surveillance videos [2], [8]. Weakly supervised methods mitigate this by using only video-level annotations [1], [3], [7]. Although this approach reduces the need for precise frame-level annotations, it still requires annotators to watch entire videos to assign video-level labels, which remains costly. One-class classification learns normal patterns using only normal videos and regards inputs that deviate from the learned normality as anomalies [8]. However, this setting assumes access to normal-only training

*These authors contributed equally to this work.

data, which may not be available in real-world scenarios. In contrast, fully unsupervised methods use unlabeled training data that may contain both normal and abnormal videos [2], [8]–[10]. In this paper, we use UVAD to denote VAD in the fully unsupervised setting. Although this setting can reduce annotation costs, constructing stable training signals without annotations remains challenging [8], [10].

Challenge 1: Stable training signals without ground-truth annotations. These limitations motivate a UVAD framework that can construct stable training signals from unlabeled training videos. We address this challenge by generating pseudo labels without ground-truth annotations.

B. Unsupervised Video Anomaly Detection (UVAD)

Existing UVAD methods include reconstruction-based and pseudo-label-based methods.

a) *Reconstruction-based UVAD*: Reconstruction-based methods assume that normal data can be reconstructed well, whereas abnormal data cannot. In general, an encoder-decoder network reconstructs the input video, and the reconstruction error is used as an anomaly score [2]. Although this approach does not require ground-truth annotations, a powerful reconstruction model can reconstruct abnormal scenes well, reducing the discriminative power of the reconstruction error [6].

b) *Pseudo-label-based UVAD*: Pseudo-label-based methods use estimated normal and abnormal segments from unlabeled videos as training targets [8]–[10]. C2FPL [10] generates pseudo labels through hierarchical clustering and statistical testing. LANP [9] assumes that the start and end segments of a video are mostly normal and uses them as a normality prior for pseudo-label generation. Despite these advances, pseudo-label noise remains a major challenge because inaccurate labels can cause the model to confuse normal and abnormal segments [2].

Challenge 2: Reliable pseudo-label generation. Although prior work has shown that normality priors can help reduce pseudo-label noise [9], pseudo-label generation still requires both a reliable normal reference and refinement strategies to reduce noisy assignments. This motivates our normal-prototype prior, which uses the initial segments of each video as a normal prototype, together with refinement strategies.

C. Test-Time Adaptation (TTA)

Deep models often degrade when the test data differ from those in the training distribution [11]. Recent TTA methods address this problem by adapting a trained model using unlabeled test data, without requiring access to the original training data or training process [11]. This property is particularly relevant to VAD, since surveillance videos are captured by cameras with different viewpoints, backgrounds, and illumination conditions, while collecting annotations or retraining a detector for every unseen camera environment is impractical [3], [4].

Challenge 3: Annotation-free adaptation at inference time. As discussed above, LANP [9] mainly uses normality priors during training for pseudo-label generation. Meanwhile, cross-domain VAD methods such as CDL [4] address domain shifts during training with weak supervision and unlabeled external

data. In contrast, adapting to unseen camera environments at inference time without ground-truth annotations or additional training videos remains important for deployment-oriented UVAD. We therefore exploit the normal-prototype prior for both pseudo-label generation and inference-time adaptation.

III. PROPOSED METHOD

Fig. 1 illustrates the overall pipeline of the proposed NPP-ITA framework, which consists of three stages—a normal-prototype-based pseudo-label generator, anomaly detector, and inference-time adapter. The main idea of NPP-ITA is that the initial segments of a video can serve as a reliable anchor for the normality of the video in both training and inference phases. In the training phase, the initial segments of each video are used to construct a normal prototype, which is then used to generate pseudo labels to train the anomaly detector. In the inference phase, the same prior is used to adapt the input adapter using the initial segments of each test video, while keeping the trained anomaly detector frozen. The adapted model then predicts anomaly scores for the remaining segments.

A. Notation

Let $\mathcal{D} = \{V_i\}_{i=1}^N$ denote an unlabeled training set of N surveillance videos. Each video V_i is divided into a sequence of non-overlapping temporal segments, and the t -th segment of the i -th video is represented by a feature vector $\mathbf{x}_{i,t} \in \mathbb{R}^d$. We denote the feature sequence of video V_i as $\mathbf{X}_i = \{\mathbf{x}_{i,t}\}_{t=1}^{T_i}$, where T_i is the number of segments in V_i . For each video segment $\mathbf{x}_{i,t}$, the anomaly detector predicts a segment-level anomaly score $s_{i,t} \in [0, 1]$.

B. Normal-Prototype-Based Pseudo-Label Generator

Since ground-truth annotations are not available in the UVAD setting, we generate pseudo labels to train the segment-level anomaly detector. In the training phase, we define the normal prototype $\mathbf{p}_i$ as the average feature of the first k segments in each training video V_i :

$$\mathbf{p}_i = \frac{1}{k} \sum_{t=1}^k \mathbf{x}_{i,t}, \quad (1)$$

where k denotes the number of initial segments. The prototype distance of each remaining segment is then computed as

$$d_{i,t}^p = D(\mathbf{x}_{i,t}, \mathbf{p}_i), \quad t = k+1, \dots, T_i, \quad (2)$$

where $D(\cdot, \cdot)$ denotes the Euclidean distance. A larger prototype distance indicates that the segment deviates more from the normal pattern of the corresponding video and is therefore more likely to be abnormal.

Although prototype distance provides an initial criterion for separating normal and abnormal segments, directly thresholding the distance scores may produce noisy pseudo labels due to local score fluctuations. To mitigate the noise, we exploit the domain prior that abnormal events in videos tend to be temporally continuous. Based on this prior, we propose a multi-stage pseudo-label refinement pipeline comprising score-level temporal refinement, binary label-level post-processing, and confidence-based refinement.

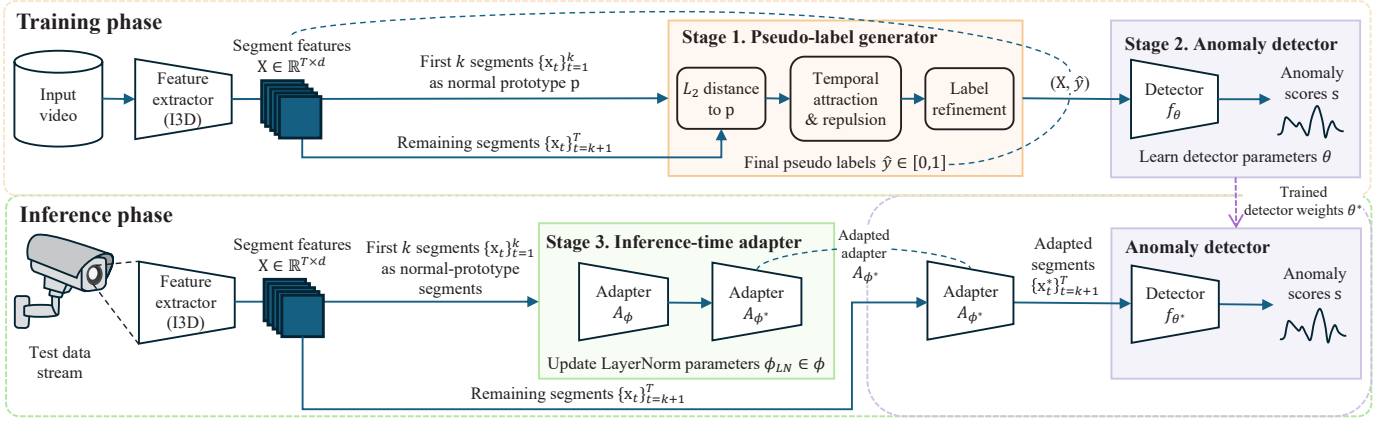


Fig. 1. Overall pipeline of the proposed NPP-ITA framework, comprising three stages: normal-prototype-based pseudo-label generator, anomaly detector, and inference-time adapter. For readability, the figure illustrates a per-video flow and omits the video index.

To reduce local score fluctuations before thresholding, we refine the prototype distance sequence using temporal attraction and repulsion. This process encourages temporally consistent abnormal responses and suppresses isolated noisy spikes, producing the refined prototype-based score $\tilde{d}_{i,t}$.

Based on the refined scores, we generate initial binary pseudo labels as follows:

$$\hat{y}_{i,t}^{(0)} = \begin{cases} 1, & \text{if } \tilde{d}_{i,t} \geq \tau, \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

where $\tau = \mu_{\tilde{d}} + \Phi^{-1}(1 - \alpha)\sigma_{\tilde{d}}$ is a global threshold computed from the refined scores $\tilde{d}_{i,t}$ over the training set. Here, $\mu_{\tilde{d}}$ and $\sigma_{\tilde{d}}$ denote the mean and standard deviation of these scores, respectively, and α controls the upper-tail probability. This reflects the assumption that anomalous segments occupy only a small portion of the training data.

After obtaining the initial binary pseudo labels, we refine them separately for each video and omit the video index i hereafter. To enforce temporal continuity, we remove isolated abnormal spikes and fill short normal gaps. To handle abnormal initial segments that contaminate the normal prototype, we apply a label correction strategy by flipping the normal and abnormal labels when the proportion of abnormal pseudo labels is excessively high, thereby reducing pseudo-label noise.

We then compute a confidence score for each pseudo label based on the agreement between feature similarity and label consistency with neighboring segments. For each non-boundary segment t , we first select the closer temporal neighbor $n(t)$ between the previous and next segments in terms of feature distance. Then, we compute the feature similarity S_t , label consistency C_t , and confidence score W_t as follows:

$$n(t) = \arg \min_{r \in \{t-1, t+1\}} D(\mathbf{x}_t, \mathbf{x}_r), \quad (4)$$

$$S_t = \exp\left(-\frac{D(\mathbf{x}_t, \mathbf{x}_{n(t)})}{\beta}\right), \quad (5)$$

$$C_t = 1 - \left| \hat{y}_t^{(0)} - \hat{y}_{n(t)}^{(0)} \right|, \quad (6)$$

$$W_t = S_t C_t + (1 - S_t)(1 - C_t). \quad (7)$$

Here, β is a scale parameter and $\hat{y}_t^{(0)}$ denotes the initial binary pseudo label. For boundary segments, we compute

the confidence score using the only available neighbor. The confidence scores of normal-prototype segments are set to 1.

A high confidence score indicates agreement between feature similarity and pseudo-label consistency. Labels with $W_t < 0.5$ are flipped, while those with $0.5 \leq W_t < 0.7$ are softened using W_t . The final labels $\hat{y}_{i,t}$ are used for detector training.

C. Anomaly Detector

Using the pseudo labels $\hat{y}_{i,t}$, we train a temporal anomaly detector to predict segment-level anomaly scores. The detector is a temporal convolutional neural network (CNN) composed of gated residual blocks, where each block combines a convolution and a sigmoid gate to capture local temporal dependencies. The final feature is projected to a segment-level logit and converted to an anomaly probability by a sigmoid function.

For training, rather than uniformly sampling each video into a short fixed-length sequence [1], [7], we preserve temporal order using windows of length T_{win} . Videos longer than T_{win} are divided into multiple clips, whereas shorter videos are kept as single clips and padded within each mini-batch.

Let $\mathbf{X}_b$ denote the feature sequence of the b -th clip. The detector f_θ produces a sequence of segment-level anomaly scores

$$\mathbf{s}_b = f_\theta(\mathbf{X}_b), \quad (8)$$

where $s_{b,t}$ denotes the anomaly score of the t -th segment.

To handle noisy pseudo labels, we optimize the detector using a hybrid loss. Let $e_{b,t} = |s_{b,t} - \hat{y}_{b,t}|$ denote the prediction error. The loss uses a squared-error penalty for small errors and a binary cross-entropy (BCE) loss for large errors:

$$\mathcal{L}_{b,t} = \begin{cases} \frac{1}{2}(s_{b,t} - \hat{y}_{b,t})^2, & \text{if } e_{b,t} < \lambda, \\ \mathcal{L}_{\text{BCE}}(s_{b,t}, \hat{y}_{b,t}), & \text{otherwise,} \end{cases} \quad (9)$$

where $\lambda = 0.3$ is the error threshold. This design strengthens the supervision for large errors while limiting updates for small errors.

Let $m_{b,t} \in \{0, 1\}$ denote whether the t -th segment of the b -th training clip is valid. The final detector loss is computed only over valid segments:

$$\mathcal{L}_{\text{det}} = \frac{1}{\sum_{b,t} m_{b,t}} \sum_{b,t} m_{b,t} \mathcal{L}_{b,t}. \quad (10)$$

D. Inference-Time Adapter

To improve robustness in unseen camera environments, we apply ITA using the same normal-prototype prior as in training. ITA follows the standard TTA protocol, in which no ground-truth annotations are used during adaptation; we use the term ITA to emphasize its deployment-oriented use.

To enable lightweight adaptation, we insert an input adapter A_ϕ before the trained anomaly detector. The adapter transforms the input feature sequence before passing it to the detector, while the detector itself remains frozen during adaptation.

Let f_{θ^*} denote the trained anomaly detector. During ITA, θ^* is fixed, and only the adapter’s LayerNorm affine parameters $\phi_{\text{LN}} \subset \phi$ are updated. For each test video $\mathbf{X}_i$, we use the first k segments as normal-prototype segments. Let $z_{i,t}$ denote the detector logit before the sigmoid output layer for segment t :

$$z_{i,t} = f_{\theta^*}^{\text{logit}}(A_\phi(\mathbf{X}_i))_t. \quad (11)$$

The ITA loss is defined over the first k segments as

$$\mathcal{L}_{\text{ITA}}^{(i)} = \sum_{t=1}^k \text{softplus}(z_{i,t}). \quad (12)$$

Minimizing this loss encourages the adapter to adjust the input feature distribution so that the frozen detector assigns lower anomaly logits to the normal-prototype segments. This constrained update reduces the risk of overfitting to a single test video. After adaptation, the remaining segments are transformed by the adapted adapter A_{ϕ^*} to obtain $\{\mathbf{x}_{i,t}^*\}_{t=k+1}^{T_i}$, and anomaly scores are computed by passing these adapted segments to the frozen detector f_{θ^*} . The normal-prototype segments used for adaptation are excluded from scoring and evaluation to avoid evaluation leakage.

IV. EXPERIMENTS

We evaluate the proposed method on two widely used VAD datasets: UCF-Crime [1] and ShanghaiTech (SHT) [6]. Following common practice, we use frame-level area under the curve (AUC) as the main evaluation metric. We compare our method with state-of-the-art UVAD methods and further analyze its effectiveness through ablation studies, sensitivity analysis, and qualitative results.

A. Datasets

UCF-Crime [1] is a large-scale dataset containing 1,900 untrimmed surveillance videos with 13 anomaly categories. Although video-level annotations are available, we assume a UVAD setting and do not use any annotations during training. Frame-level annotations are used only for evaluation. Since CDL [4] reports noise in the original UCF-Crime test annotations and provides corrected annotations, we report results under both the original annotations (UCF Orig.) and the CDL-corrected annotations (UCF CDL).

ShanghaiTech [6] contains 437 campus surveillance videos from 13 scenes with different camera viewpoints. The original training split contains only normal videos, which corresponds to a one-class setting. For the UVAD setting, we follow the data organization used in [12], where the training set includes both normal and abnormal videos.

TABLE I

COMPARISON WITH STATE-OF-THE-ART METHODS IN FRAME-LEVEL AUC (%) ON UCF-CRIME AND SHANGHAITECH UNDER THE UVAD SETTING. UCF ORIG. AND UCF CDL DENOTE EVALUATION WITH THE ORIGINAL FRAME-LEVEL ANNOTATIONS [1] AND THE CDL-CORRECTED ANNOTATIONS [4]. PARENTHESES IN FRD-UVAD INDICATE THE AUXILIARY SCORER USED IN THE ORIGINAL PAPER. OUR METHOD EXCLUDES THE FIRST k NORMAL-PROTOTYPE SEGMENTS USED FOR ITA FROM EVALUATION.

Method	UCF Orig.	UCF CDL	SHT
DyAnNet [14]	79.76	—	—
C2FPL [10]	80.65	—	—
LANP [9]	80.02	—	88.32
FRD-UVAD (None) [2]	77.56	74.47 [†]	—
FRD-UVAD (C2FPL) [2]	82.12	82.49 [†]	80.14
NPP-ITA (Ours)	81.62	83.07	78.06

[†] denotes our re-evaluation using the released checkpoint and the same saved frame-level prediction scores, replacing only the UCF Orig. with UCF CDL. **Bold** values indicate the best result in each column.

B. Implementation Details

We use a pre-trained I3D feature extractor [13] to obtain RGB segment features. Each video is divided into non-overlapping 16-frame segments, and ten-crop features are extracted and averaged over crops to obtain a 2048-dimensional feature vector per segment. The detector predicts segment-level anomaly scores, which are repeated over the corresponding 16 frames for frame-level evaluation.

The number of initial segments for the normal-prototype prior is fixed at $k = 5$ (80 frames, about 2.7 seconds at 30 fps) in all experiments. These segments are used both to construct the normal prototype during pseudo-label generation and to define the normal-prototype segment set for the ITA loss.

For pseudo-label generation and refinement, we set $\alpha = 0.05$ and $\beta = 10$. We set $T_{\text{win}} = 500$ for UCF-Crime and $T_{\text{win}} = 20$ for ShanghaiTech. The anomaly detector is implemented as a CNN with two gated residual convolutional blocks using temporal kernel sizes of 3 and 5. The detector is trained for 30 epochs with a batch size of 32 on UCF-Crime and for 60 epochs with a batch size of 8 on ShanghaiTech. We use SGD with a learning rate of 1×10^{-3} , momentum of 0.9, Nesterov acceleration, and weight decay of 1×10^{-4} .

For ITA, the detector is frozen and the adapter is reset for each test video. We update only the LayerNorm affine parameters using Adam with a learning rate of 1×10^{-2} for 30 adaptation steps per test video. The initial k segments used for adaptation are excluded from evaluation.

C. Results and Analysis

Quantitative results. Table I reports frame-level AUC results on UCF-Crime and ShanghaiTech under the UVAD setting. Our goal is not to claim detector-only superiority under the conventional full-video protocol, but to evaluate whether the normal-prototype prior provides a practical annotation-free deployment mechanism for surveillance videos. Since most previous methods report results under UCF Orig., we use the same protocol for comparison and additionally report results under UCF CDL. For UCF-Crime, we mainly compare with RGB I3D-based methods. On UCF-Crime, NPP-ITA achieves

TABLE II

ABLATION ANALYSES ON UCF-CRIME IN TERMS OF FRAME-LEVEL AUC (%). THE SUFFIX-ONLY PROTOCOL EVALUATES ONLY THE SEGMENTS AFTER THE FIRST $k = 5$ SEGMENTS. THE COMPONENT-WISE ABLATION USES THE CDL-CORRECTED ANNOTATIONS (UCF CDL).

(a) Evaluation protocol ablation			(b) Component-wise ablation			
Setting	UCF Orig.	UCF CDL	Temp.Att.	Con.Ref.	ITA	AUC
Full-video w/o ITA	80.64	81.99	×	×	×	60.43
Suffix-only w/o ITA	80.56	81.95	✓	×	×	80.19
Suffix-only w/ ITA	81.62	83.07	✓	✓	×	81.95
			✓	✓	✓	83.07

Temp.Att., Con.Ref., and ITA denote temporal attraction, confidence-based refinement, and inference-time adaptation, respectively.

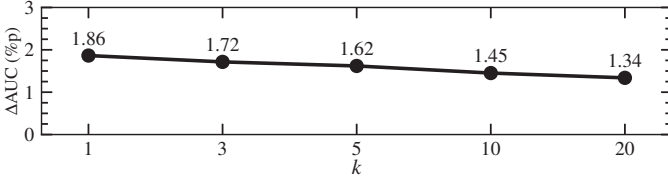


Fig. 2. Sensitivity of ITA to the number of normal-prototype segments k . Δ AUC denotes the gain over the fixed suffix-only baseline.

81.62% AUC under UCF Orig. and the best AUC of 83.07% under UCF CDL, showing that the proposed prior remains effective under the corrected-annotation protocol. On ShanghaiTech, NPP-ITA achieves 78.06% AUC, below the strongest methods on this dataset. This suggests that the proposed normal-prototype prior may be more beneficial for datasets with diverse camera environments and more pronounced train-test feature shifts, such as UCF-Crime.

Ablation analysis. Table II presents two ablation analyses on UCF-Crime. In Table II(a), the suffix-only baseline without ITA is almost identical to the full-video baseline, while ITA clearly improves AUC. This shows that the gain mainly comes from ITA, not simply from excluding the first k segments. Table II(b) further evaluates the effect of each proposed component. Temporal attraction improves the pseudo-label quality over the basic prototype-distance thresholding, and confidence-based refinement yields additional gains. The full model with ITA achieves the best performance, confirming the contribution of each component.

Sensitivity to the number of normal-prototype segments k at inference time. Fig. 2 shows the sensitivity of ITA to k . To isolate the effect of k from the evaluation region, we fix the evaluation region by excluding the first 20 segments for all settings and vary only the number of segments used for ITA. ITA consistently improves AUC over the fixed suffix-only baseline for all tested values of k , indicating that the proposed adaptation is robust to k and is not tuned to a specific choice.

Qualitative analysis. Fig. 3 shows representative anomaly score curves after ITA. The predicted scores increase around the ground-truth abnormal intervals, while remaining relatively low in normal regions.

V. CONCLUSION

We proposed NPP-ITA, a UVAD framework that exploits a normal-prototype prior for both pseudo-label generation and inference-time adaptation. By using the initial segments of each video, our framework constructs pseudo labels and performs lightweight adaptation without ground-truth annotations.

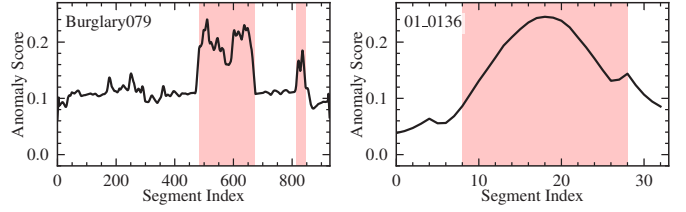


Fig. 3. Per-segment anomaly score curves on example UCF-Crime (left) and ShanghaiTech (right) test videos. Red shading marks the ground-truth abnormal intervals.

Experiments demonstrate clear performance gains on UCF-Crime, while the results on ShanghaiTech suggest that a more adaptive prototype selection strategy is needed for diverse cross-dataset environments. Building upon these insights, future work will focus on refining the prototype selection to enhance robustness across varying camera viewpoints and evaluating the framework on broader benchmarks to further validate its generalization capabilities.

REFERENCES

- [1] W. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2018, pp. 6479–6488.
- [2] C. Tao, C. Wang, S. Lin, S. Cai, D. Li, and J. Qian, “Feature reconstruction with disruption for unsupervised video anomaly detection,” *IEEE Trans. Multimedia*, vol. 26, pp. 10 160–10 173, 2024.
- [3] M. Cho, T. Kim, M. Shim, D. Wee, and S. Lee, “Towards multi-domain learning for generalizable video anomaly detection,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, 2024, pp. 50 256–50 284.
- [4] Y. Jain, A. Dabouei, and M. Xu, “Cross-domain learning for video anomaly detection with limited supervision,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2025, pp. 468–484.
- [5] T. Zhu, Q. Yu, X. Dong, S. Li, Y. Liu, J. Jiang, and L. Shu, “ProDisc-VAD: An efficient system for weakly-supervised anomaly detection in video surveillance applications,” in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, 2025, pp. 2532–2537.
- [6] W. Liu, W. Luo, D. Lian, and S. Gao, “Future frame prediction for anomaly detection – a new baseline,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2018, pp. 6536–6545.
- [7] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, “Weakly-supervised video anomaly detection with robust temporal feature magnitude learning,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, October 2021, pp. 4975–4986.
- [8] M. Z. Zaheer, A. Mahmood, M. H. Khan, M. Segu, F. Yu, and S.-I. Lee, “Generative cooperative learning for unsupervised video anomaly detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2022, pp. 14 744–14 754.
- [9] H. Shi, L. Wang, S. Zhou, G. Hua, and W. Tang, “Learning anomalies with normality prior for unsupervised video anomaly detection,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2025, pp. 163–180.
- [10] A. Al-lahham, N. Tastan, M. Z. Zaheer, and K. Nandakumar, “A coarse-to-fine pseudo-labeling (C2FPL) framework for unsupervised video anomaly detection,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, January 2024, pp. 6793–6802.
- [11] Y. Yuan, B. Xu, L. Hou, F. Sun, H. Shen, and X. Cheng, “TEA: Test-time energy adaptation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2024, pp. 23 901–23 911.
- [12] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, “Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2019, pp. 1237–1246.
- [13] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the Kinetics dataset,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, July 2017, pp. 4724–4733.
- [14] K. V. Thakare, Y. Raghuwanshi, D. P. Dogra, H. Choi, and I.-J. Kim, “DyAnNet: A scene dynamicity guided self-trained video anomaly detection network,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, January 2023, pp. 5541–5550.