Abnormal Ratios Guided Multi-Phase Self-Training for Weakly-Supervised Video Anomaly Detection

Haoyue Shi , *Student Member, IEEE*, Le Wang , *Senior Member, IEEE*, Sanping Zhou , *Member, IEEE*, Gang Hua , *Fellow, IEEE*, and Wei Tang , *Member, IEEE*

Abstract—**Weakly-supervised Video Anomaly Detection (W-VAD) aims to detect abnormal events in videos given only video-level labels for training. Recent methods relying on multiple instance learning (MIL) and self-training achieve good performance, but they tend to focus on learning easy abnormal patterns while ignoring hard ones, e.g., unusual driving trajectory or over-speeding driving. How to detect hard anomalies is a critical but largely ignored problem in W-VAD. To tackle this challenge, we propose a novel framework, termed Abnormal Ratios guided Multi-phase Self-training (ARMS), for W-VAD. It includes a new abnormal ratio-based MIL (AR-MIL) loss and a new multi-phase self-training paradigm. The AR-MIL loss guides the learning of hard anomalies by enforcing a minimum ratio of abnormal snippets in an abnormal video and no abnormal snippets in a normal video. Our multi-phase self-training paradigm sequentially performs bootstrapping, hard anomalies mining, and adaptive self-training so as to address pseudo labeling on easy anomalies, detect hard anomalies, and setting adaptive abnormal ratios for different videos in a unified framework. Experimental results on three benchmark datasets, i.e., ShanghaiTech, UCF-Crime, and XD-Violence, show that ARMS outperforms all previous state-of-the-art methods and has a great advantage in detecting hard anomalies.**

Index Terms—**Anomaly detection, weakly-supervised video anomaly detection, multiple instance learning.**

I. INTRODUCTION

VIDEO anomaly detection (VAD) aims to detect abnormal events in video sequences [1], [2], [3]. It has a wide range

Haoyue Shi is with the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China, and also with the Department of Computer Science, University of Illinois Chicago, ChicagoIL 60607 USA (e-mail: shyern@stu.xjtu.edu.cn).

Le Wang and Sanping Zhou are with the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: lewang@mail.xjtu.edu.cn; spzhou@mail.xjtu.edu.cn).

Gang Hua is with the Wormpex AI Research, Bellevue, WA 98004 USA (e-mail: ganghua@gmail.com).

Wei Tang is with the Department of Computer Science, University of Illinois Chicago, Chicago, IL 60607 USA (e-mail: tangw@uic.edu).

Digital Object Identifier 10.1109/TMM.2023.3336576

of real-world applications, e.g., intelligent surveillance [4], [5], [6], crime detecting [7], [8], [9], and road accident warning [10]. VAD falls into four major categories: full-supervised, unsupervised, one-class, and weakly-supervised video anomaly detection [10]. As obtaining frame-level annotations of abnormal events is very expensive, *weakly-supervised video anomaly detection* (W-VAD) has attracted increasing attention in recent years. It can achieve promising performance with only video-level annotations for training. However, W-VAD remains challenging as abnormal events are usually rare in videos and more complicated than normal events.

A common framework adopted by existing state-of-the-art methods [1], [11], [12], [13], [14] integrates multiple instance learning (MIL). The Maximum Score-based MIL (MS-MIL) learning objective utilizes a ranking loss to enlarge the margin between normal and abnormal video predictions. One effective MS-MIL loss [1] takes the snippet with the highest abnormal score as the video-level prediction. The top-k ranking loss-based MIL models [11], [14] take the mean score of the top-k predicted instances as the abnormal video prediction.

Despite their advancement, there is a critical limitation. The presence of hard anomalies (usually imperceptible and easily mispredicted as normal) may impact the overall model performance. Fig. 1 shows the example abnormal frames in real-world surveillance videos. Suspicious person moving frames from *Arson010* look similar to normal frames but should be abnormal frames in an arson event. Unusual driving tendency frames from *RoadAccidents133* are usually imperceptible but important to the road accident event warning. However, both top-1 and top-k ranking MIL losses tend to treat these hard abnormal frames as normal. This is because only several snippets with top-k or top-1 abnormal scores in an abnormal video are used to update the model. Then easy abnormal snippets that deviate significantly from normal events are selected at the beginning of training, and the subsequent optimization will keep increasing their abnormal scores while treating other snippets as normal. Thus, hard anomalies with lower scores are easily missed in MS-MIL based methods. In addition, it tends to be extremely sensitive to the selected k value [3].

Recent methods [3], [7], [15], [16], [17], [18], [19] have pushed forward the state-of-the-art anomaly detection performance. Self-training methods [7], [15] first initialize the snippet-level pseudo labels through the MS-MIL loss first, and then alternate between model re-training and pseudo labeling. Various feature integration methods [16], [17], [18], [19] design
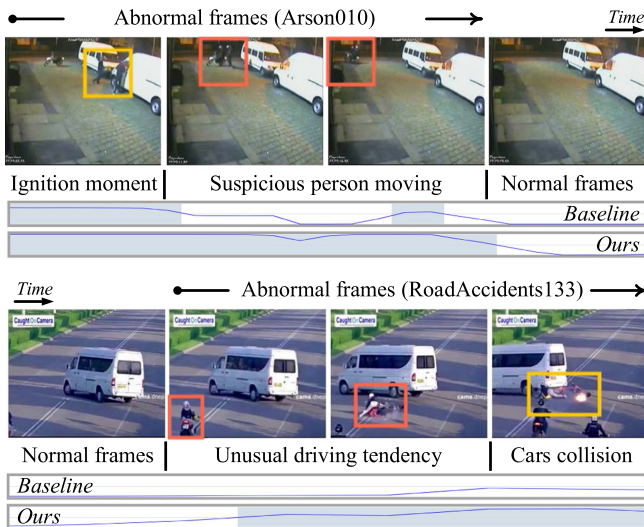
Fig. 1. Frames of the suspicious person moving (red rectangle) in the arson event (*Arson010*) and frames of unusual driving tendency (red rectangle) in the road accident event (*RoadAccidents133*) are hard anomalies and difficult to detect. They are misclassified in the traditional self-training method (*baseline*). By contrast, these abnormal frames are classified correctly by our method, which is designed to detect hard anomalies. The predictions and thresholded results of the baseline and our method are plotted.

different deep learning architectures and adopt the MS-MIL loss and other regularization terms. Sapkota et al. [3] aim to detect anomalies in multimodal scenarios for W-VAD. However, prior W-VAD methods neglect the detection of hard anomalies, which is one of the greatest challenges of W-VAD. Moreover, the lack of high-quality pseudo labels for anomalies prevents the previous self-training methods from learning hard anomalies and remedying this limitation.

To tackle this challenge, we propose a novel framework, termed Abnormal Ratios guided Multi-phase Self-training (ARMS), for W-VAD, as presented in Fig. 2. It includes a new Abnormal Ratio-based MIL (AR-MIL) loss and a new multi-phase self-training paradigm. The AR-MIL loss, as shown in Fig. 2, enforces that the ratio of abnormal snippets in an abnormal video should be larger than a margin. Compared with the MS-MIL loss, which selects only several snippets (with the highest or top-k abnormal scores) in an abnormal video for the model update, our AR-MIL loss takes all snippets in a video as normal or anomaly candidates, which are optimized jointly to meet the abnormal ratio. As a result, our AR-MIL loss will learn more comprehensive normal and abnormal patterns than other losses. This prevents the model from being trapped by the initial selection and helps in detecting hard anomalies.

Based on the AR-MIL loss, our ARMS framework includes three training phases: *bootstrapping*, *hard anomalies mining*, and *adaptive self-training*. Bootstrapping uses the AR-MIL loss with a relatively small abnormal ratio to train an initial model and obtain high-quality snippet-level pseudo labels. These pseudo labels mostly correspond to easy anomalies. Then, hard anomalies mining integrates the AR-MIL loss with a larger abnormal ratio and a classification loss to mine hard anomalies from the abnormal videos while maintaining good performance on easy

anomalies through the snippet-level pseudo labels. Till now, we have set the same abnormal ratio for all abnormal videos, but, in practice, some abnormal videos contain more abnormal snippets than others. To close this gap, the last training phase learns adaptive abnormal ratios for different abnormal videos and constantly updates them based on the estimates in the previous iteration. Different from the prior self-training methods and other feature integration methods that focus on easy normal and abnormal snippets, our method first integrates a new abnormal ratio based MIL loss to learn more comprehensive normal and abnormal patterns, and then propose a new multi-phase self-training paradigm that adopts three training phases and the AR-MIL loss with different abnormal ratios λ to help in detecting hard anomalies.

Experimental results on three benchmark datasets, i.e., ShanghaiTech [20], UCF-Crime [1], and XD-Violence [13], show that ARMS outperforms all previous state-of-the-art methods. In particular, ARMS has a great advantage of detecting hard anomalies, which is one of the greatest challenges of W-VAD. Fig. 1 illustrates the results obtained by our method and the baseline. The hard abnormal frames of suspicious person moving and unusual driving tendency are missed in the baseline but are detected by our framework.

Our main contributions are summarized as follows:

- Unlike existing anomaly detection methods, we contribute to addressing the important but largely ignored hard anomalies detection problem.
- We propose a new abnormal ratio-based multiple instance learning (AR-MIL) loss. Compared with the MS-MIL loss, it jointly optimizes all snippets in a video and better finds hard anomalies in abnormal videos.
- Based on the AR-MIL loss, we introduce a novel Abnormal Ratios guided Multi-phase Self-training (ARMS) framework for W-VAD. Its three training phases (bootstrapping, hard anomalies mining, and adaptive self-training) address pseudo labeling on easy anomalies, mining hard anomalies, and setting adaptive abnormal ratios for different videos in a unified framework.

The rest of the paper is organized as follows. Section II discusses related work. We present the technical details of the proposed method in Section III. Experimental results and discussions are presented in Section IV. Finally, we conclude the paper in Section V.

II. RELATED WORKS

In this section, we review previous works related to ours, which we categorize into two parts: 1) video anomaly detection and 2) weakly supervised video anomaly detection.

A. Video Anomaly Detection

Video Anomaly Detection (VAD) [1], [7], [10], [21], [22], [23], [24] is an important field in video and image understanding [25], [26], [27]. VAD aims to detect abnormal events in video sequences. It falls into four major categories: full-supervised, unsupervised, one-class, and weakly-supervised video anomaly detection [10]. It is labour-intensive and time-consuming for
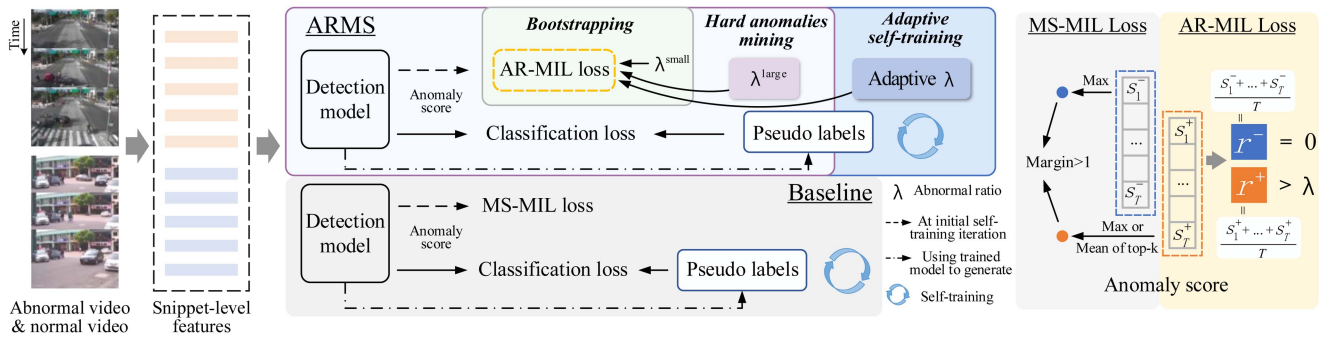
Fig. 2. Overview of our Abnormal Ratios guided Multi-phase Self-training (ARMS) framework for W-VAD. It first performs bootstrapping with a small abnormal ratio λ^{small}, then mines hard anomalies with a large abnormal ratio λ^{large} as well as snippet-level pseudo labels, and finally sets adaptive abnormal ratios for different videos for progressive improvement. The baseline method integrates the prior MS-MIL loss with the traditional self-training procedure. The prior MS-MIL loss enlarges the margin of normal and abnormal video predictions. The proposed AR-MIL loss uses the average of all snippets' abnormal scores in a video as the expectation of the ratios of abnormal snippets, then enforces a minimum ratio of abnormal snippets in an abnormal video and no abnormal snippets in a normal video.

full-supervised VAD methods [21], [28] to collect precise annotations. What's more, obtaining sufficient anomaly examples is quite cumbersome. Unsupervised VAD methods [10], [22], [29], [30] detect abnormal events without any labels. It only relies on some prior knowledge, such as anomalies are less frequent than the normal happenings. Unlike the unsupervised setting, the training set of one-class VAD contains only normal videos. Previous works [2], [8], [24], [31] usually build a reconstruction model to learn normal patterns. Then anomalies are viewed as the snippets that do not fit this normality model. However, both unsupervised and one-class VAD methods achieve inferior performance due to the lack of domain knowledge. In the current work, we explore weakly-supervised mode [1], [7] for video anomaly detection. Weakly supervised video anomaly detection (W-VAD) aims to detect abnormal frames with video-level annotations and differs from other unsupervised learning tasks without any annotations [32], [33]. Moreover, unlike other weakly-supervised or unsupervised learning task [34], [35], [36] that involves learning from unseen classes, all abnormal classes are included in the training set of W-VAD. W-VAD has attracted a lot of attention in recent years [1], [3], [7], [14] because it can achieve good performance with much lower labeling cost than fully supervised methods. Different from other video detection methods [25], [26] that use contrastive learning or temporal dynamics to learn discriminative embeddings, our method integrates a new multi-phase self-training paradigm to learn comprehensive normal and abnormal patterns.

B. Weakly-Supervised Video Anomaly Detection

Prior methods [3], [9], [11], [12], [13], [17] investigate different architectures of neural networks, objective functions, and training strategies to improve anomaly detection performance. As our proposed approach is related to multiple instance learning (MIL) and self-training, we review prior W-VAD methods from these two perspectives below.

Multiple instance learning (MIL) treats the entire video as a labeled bag containing multiple unlabeled instances (i.e., video snippets) [1], [12], [14]. Sultani et al. [1] introduce a maximum score-based multiple instance learning (MS-MIL) loss for this task. It first takes the snippet with the highest abnormal score as the video-level prediction, and then utilizes a ranking loss to enlarge the margin between normal and abnormal video predictions. Later, a temporal ranking loss [37] and a new complementary inner bag loss [11] are proposed to improve the top-1 ranking loss-based MIL model. Several methods [12], [13], [17] enhance the robustness of the above MS-MIL loss by aggregating the top-k abnormal scores as the abnormal video prediction, and utilizing a ranking loss or a cross-entropy loss as the objective function. In addition, various feature integration methods [16], [17], [18], [19] focus on designing more effective deep learning architectures while adopting the MS-MIL loss and other regularization terms. To use input features effectively, Tian et al. [14] propose temporal feature magnitude learning that calculates the MS-MIL loss over the feature magnitudes. Chang et al. [9] propose a novel model including a contrastive attention module and an attention consistency loss to boost detection performance further. Moreover, Sapkota et al. [3] conduct Bayesian nonparametric submodularity video partition for outlier and multimodal scenarios in W-VAD. However, all these previous MIL-based methods tend to ignore hard anomalies, e.g., unusual driving tendency or trajectory in *Road Accident*. This is because only several snippets with top-k abnormal scores in an abnormal video are used to update the detection model. Our abnormal ratio-based MIL loss will be able to find hard anomalies because it considers all snippets in the learning process. It keeps looking for anomalies as long as the ratio of detected abnormal snippets is not large enough.

Self-training has been extensively studied in semi-supervised learning [38], [39], [40] and recently extended to anomaly detection. It first trains a model only on labeled data to generate pseudo labels on unlabeled data, and then re-trains the model on both labeled and unlabeled data. Pang et al. [22] propose a self-trained deep ordinal regression network on the testing video directly for unsupervised VAD. Zhong et al. [41] propose an iterative training framework to optimize the action classifier and a novel noise cleaner, but the iterative optimization is inefficient. MIST [7] first obtains the frame-level pseudo labels based on the MIL

method and then fine-tunes a task-specific encoder by pseudo labels to learn discriminative representations. Li et al. [15] utilize a self-training method to refine the abnormal scores based on pseudo labels generated by their multi-sequence loss. Due to the lack of high-quality pseudo labels for hard anomalies, these previous self-training methods can only learn easy anomalies well. Different from all these methods, our ARMS framework includes three training phases, i.e., bootstrapping, hard anomalies mining, and adaptive self-training, which address pseudo labeling on easy anomalies, mining hard anomalies, and setting adaptive abnormal ratios for different videos in a unified framework.

III. METHOD

This section elaborates on the proposed framework. As shown in Fig. 2, given a pair of abnormal and normal videos, we extract snippet-level features through a visual encoder. Then we design a detection model to produce anomaly scores for each snippet in a video. Subsequently, our ARMS integrates the proposed AR-MIL loss with three specially designed training phases. The baseline method integrates the prior MS-MIL loss with the traditional self-training procedure. Finally, detection results are obtained by the trained anomaly detection model.

Section III-A first gives the definition of the weakly supervised video anomaly detection (WS-VAD) problem. Section III-B introduces the baseline method in detail. The following sections detail the proposed method, including the overview of our ARMS (Section III-C), the introduction of the abnormal ratio-based MIL loss (Section III-D), and multi-phase self-training (Section III-E).

A. Problem Statement

During training, we are given a set of training videos, each of which is annotated with a video-level label $y \in \{0, 1\}$. $y = 1$ means an abnormal video containing at least one abnormal event, and $y = 0$ means a normal video containing no abnormal event. After training, the model could detect abnormal frames in a new video.

Since feature extraction, anomaly detection, and training are commonly performed at the snippet level, we assume a video is divided into T non-overlapping snippets. Let $s_t^+ \in [0, 1]$ denote the predicted abnormal score of the t-th snippet in an abnormal video and let $s_t^- \in [0, 1]$ denote that in a normal video.

B. Baseline

The maximum score-based multiple instance learning (MS-MIL) loss [1] has been used for W-VAD. One effective MS-MIL loss takes the snippet with the highest abnormal score as the video-level prediction. It then enforces that the difference between the abnormal scores of an abnormal video and a normal video should be larger than a margin:

$$\mathcal{L}_{\text{MS-MIL}} = \max(0, 1 - \max_{1 \le t \le T} s_t^+ + \max_{1 \le t \le T} s_t^-), \quad (1)$$

The top-k ranking loss maximizing the margin between an average of k highest snippet predictions from an abnormal video
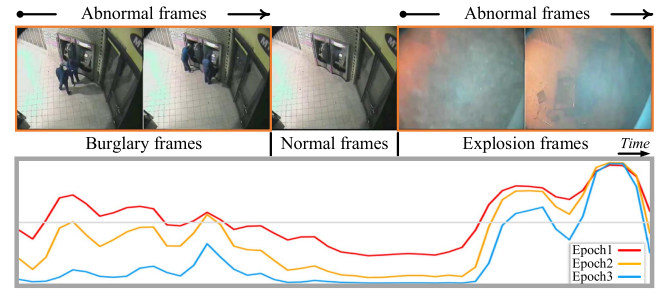

Fig. 3. Predictions of the MS-MIL loss (top-1) training in different training epochs on the abnormal training video (*Burglary010*). Because the hard anomalies learning has been neglected in the prior MS-MIL loss, the explosion frames are predicted with higher abnormal scores at epoch 1. Then the following training epochs keep increasing their scores, while predicting the burglary frames as normal frames with lower abnormal scores.

and the maximum snippet prediction from a normal video:

$$\mathcal{L}_{\text{MS-MIL}} = \max \left(0, 1 - \frac{1}{k} \sum_{i=1}^{k} s_i^+ + \max_{1 \le t \le T} s_t^- \right). \quad (2)$$

Our baseline integrates the prior MS-MIL loss with the traditional self-training procedure. Fig. 2 presents an overview of the baseline method. After training an initial model by using the MS-MIL loss, it first generates pseudo-labels according to the previous model predictions, and then trains a new model using the pseudo-labels. Finally, it alternates between pseudo labeling and model re-training. Here, once a (new) model is trained using the pseudo-labels (the MS-MIL loss), we finish a self-training iteration.

Though this baseline can achieve promising performance, it has a critical limitation. Both top-1 and top-k ranking MIL losses tend to mispredict the hard anomalies (similar to normal events and usually imperceptible) as normal. This is because only several snippets with top-k or top-1 abnormal scores are used to learn abnormal patterns of an abnormal video, the trained model can only capture a limited range of abnormal patterns in the training data. As a result, the easy abnormal snippets that deviate significantly from normal events are selected at the beginning of training. Moreover, the subsequent optimization will keep increasing their abnormal scores while treating other snippets as normal. Thus, hard anomalies with lower scores are easily missed in MS-MIL based methods. In addition, the lack of high-quality pseudo labels for hard anomalies further prevents the subsequent self-training iterations from learning hard anomalies and remedying this limitation.

To illustrate this phenomenon, Fig. 3 shows the predictions of the MS-MIL loss training in different training epochs on the abnormal training video (*Burglary010*). Because the hard anomalies learning has been neglected in the prior MS-MIL loss, the explosion frames are predicted as abnormal frames with higher scores at the beginning of training. The following training epochs keep increasing their abnormal scores, while predicting the burglary frames as normal frames with lower abnormal scores.

C. Overview of ARMS

We propose a novel framework, termed Abnormal Ratios guided Multi-phase Self-training (ARMS), for W-VAD. Fig. 2 presents an overview of ARMS. It includes a new abnormal ratio-based MIL (AR-MIL) loss and a new multi-phase self-training paradigm. The AR-MIL loss guides the learning of hard anomalies by enforcing a minimum ratio of abnormal snippets in an abnormal video and no abnormal snippets in a normal video. Based on the AR-MIL loss, ARMS first performs bootstrapping with a small abnormal ratio to detect easy anomalies and obtain high-quality snippet-level pseudo labels. It then mines hard anomalies via a large abnormal ratio and the pseudo labeled data. Finally, an adaptive self-training strategy sets adaptive abnormal ratios for different abnormal videos for more precise anomaly detection.

D. Abnormal Ratio-Based MIL Loss

As shown in Fig. 2, the abnormal ratio-based MIL (AR-MIL) loss enforces that the ratio of abnormal snippets in an abnormal video should be larger than a margin and no abnormal snippet exists in a normal video. As the abnormal score can be interpreted as the probability that a snippet is abnormal, the average of all snippets' abnormal scores in a video is the expectation of the ratio of abnormal snippets. Then the predicted anomaly ratio of the abnormal video r^+ and normal video r^- are formulated as:

$$r^+ = \frac{1}{T} \sum_{t=1}^{T} s_t^+, \quad r^- = \frac{1}{T} \sum_{t=1}^{T} s_t^-, \tag{3}$$

Then the AR-MIL loss is defined as:

$$\mathcal{L}_{\text{AR-MIL}} = r^- + \alpha \max(0, \lambda - r^+), \tag{4}$$

where λ sets the minimum ratio of abnormal snippets in an abnormal video, and α balances the impacts of the normal video and the abnormal video. Both λ and α are hyper-parameters. The first term in (4) enforces that the ratio of abnormal snippets detected in a normal video should be as small as possible, i.e., zero. The second term in (4) enforces that the ratio of abnormal snippets detected in an abnormal video should be no smaller than λ.

Our AR-MIL loss is different from the MS-MIL loss, which learns a limited range of normal and abnormal patterns from only a small number of snippets in a video. Our loss takes the average of all snippets' scores in a normal or abnormal video as the expectation of the ratio of abnormal snippets, then enforces it should be no smaller than the abnormal ratio λ in an abnormal video and be zero in a normal video. Specifically, all snippets in a normal video are used to update the model to learn comprehensive normal patterns. For the abnormal video, if the detected abnormal ratio in an abnormal video is smaller than λ, all snippets in this video will be used to update the model, and the gradient will be back-propagated to all snippets. In contrast, the existing top-k or top-1 ranking MIL loss will only back-propagate the gradient to a small number of snippets corresponding to the top-k or top-1 abnormal scores in a video. As a result, the self-training can be easily trapped by the bad initial

---

Algorithm 1: Multi-Phase Self-Training

Input: A set of training videos and their video-level labels, the number of self-training iterations K, two abnormal ratios λ^{small} and λ^{large}.

Output: An abnormal detection model.

1 **for** $k = 1$ *to* K **do**
2   Initialize the model parameters.
3   # Bootstrapping
4   **if** $k == 1$ **then**
5   Set $\lambda \leftarrow \lambda^{\text{small}}$.
6   Train the model with video-level labels by Eq. (4).
7   **end**
8   # Hard anomalies mining
9   **if** $k == 2$ **then**
10   Set $\lambda \leftarrow \lambda^{\text{large}}$.
11   Train the model with both video-level and snippet-level labels by Eq. (5).
12   **end**
13   # Adaptive self-training
14   **if** $k \geq 3$ **then**
15   Update λ for each abnormal video by Eq. (6).
16   Train the model with both video-level and snippet-level labels by Eq. (5).
17   **end**
18   Sample high-confidence snippet-level pseudo labels in abnormal videos.
19 **end**

---

selection of these top-k or top-1 abnormal snippets if they are mispredicted.

E. Multi-Phase Self-Training

Based on the AR-MIL loss, we introduce a novel multi-phase self-training paradigm for W-VAD. As illustrated in Fig. 2, it involves three training phases: bootstrapping, hard anomalies mining, and adaptive self-training, which address pseudo labeling on easy anomalies, mining hard anomalies, and setting adaptive abnormal ratios for different videos in a unified framework.

1) Bootstrapping: This phase uses the AR-MIL loss with a relatively small abnormal ratio λ^{small} to train an initial model and obtain high-quality pseudo labeled data. Because easy anomalies tend to have high abnormal scores in an abnormal video, which satisfies the AR-MIL loss, they will be detected confidently in this phase. After training, we select high-confidence predictions of snippets in abnormal videos as their snippet-level pseudo labels. As illustrated in Fig. 4(a), we take the snippets (in an abnormal video) with the highest quartile abnormal scores as abnormal snippets, while those with the lowest quartile scores as normal ones. Note that all snippets in a normal video are normal. Then, the pseudo labels obtained on abnormal videos and labels from normal videos are combined into a new snippet-level label set: $\{y_i\}_{i=1}^{N}$, where N is the total number of snippets associated with labels, and y_i is the label of a snippet.

2) Hard Anomalies Mining: We integrate the AR-MIL loss with a larger abnormal ratio λ^{large} and a classification loss in this
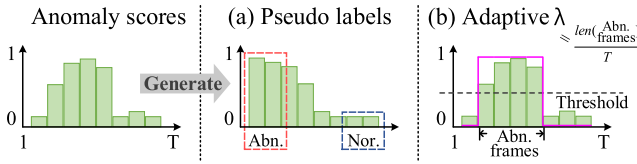
Fig. 4. Illustration of the snippet-level pseudo labeled data generation (a), and the adaptive abnormal ratio λ setting (b) for the abnormal video in the training set. The red and blue rectangles in (a) are the selected abnormal and normal snippets, respectively. The purple curve in (b) represents the threshold results.
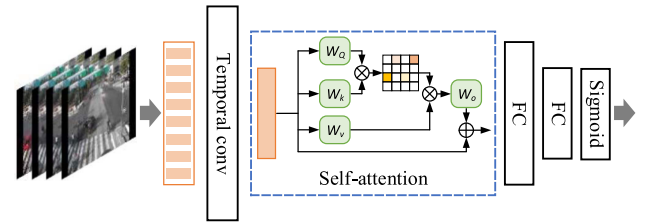

Fig. 5. Proposed network architecture. It includes a temporal convolution layer, a standard multi-head self-attention block, two linear layers, and a sigmoid function. The temporal convolution is applied to learn the local temporal dependencies between neighboring video snippets. The multi-head self-attention block is adopted to learn global temporal dependencies between non-local video snippets.

phase. On the one hand, the AR-MIL loss will keep looking for anomalies in an abnormal video as long as the ratio of detected abnormal snippets is smaller than λ^{large}. Thus, more hard anomalies will emerge in the training process. On the other hand, the classification loss on the snippet-level pseudo labels will help the model maintain good performance on easy anomalies. The objective function of this training phase is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{AR-MIL}} + \beta \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{BCE}}(s_i, y_i), \qquad (5)$$

where $\mathcal{L}_{\text{BCE}}$ is a binary cross-entropy loss, s_i is the predicted abnormal score of a snippet, y_i is from the first training phase, and β is a balancing hyper-parameter.

3) Adaptive Self-Training: We have set the same ratio of abnormal snippets for all abnormal videos in the first two phases. But some abnormal videos contain more abnormal snippets than others in practice. To close this gap, the last training phase learns adaptive abnormal ratios for different abnormal videos. Fig. 4(b) shows the adaptive abnormal ratios setting in detail. Given the predicted abnormal scores of each snippet in an abnormal video: $\{s_t^+\}_{t=1}^{T}$, we set the abnormal ratio λ for this video as the estimated ratio of abnormal snippets under a threshold μ:

$$\lambda = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}(s_t^+ \geq \mu), \qquad (6)$$

where $\mathbb{1}(s_t^+ \geq \mu)$ is an indicator function that outputs 1 if s_t^+ is larger than μ, otherwise 0. Considering that snippet-level predictions are generated by the previous model, we directly use the abnormal ratio used in the previous phase as the threshold μ. Then, based on pseudo labels generated in the previous training phase, we update the model in this phase using the loss function in (5).

In summary, we constantly update the pseudo labeled data and adaptive abnormal ratios based on the estimates in the previous self-training iteration to achieve better anomaly detection performance. We summarize our proposed multi-phase self-training in Algorithm 1.

The baseline method integrates the prior MS-MIL loss with the traditional self-training procedure. It includes several self-training iterations, and each iteration alternates between pseudo labeling and model re-training. However, the baseline method focuses on easy normal and abnormal snippets, making hard anomalies easily missed. We contribute a multi-phase self-training framework (ARMS) to address this important but

largely ignored problem. Our multi-phase self-training paradigm adopts three training phases and the AR-MIL loss with different abnormal ratios λ to help in detecting hard anomalies. The bootstrapping phase first uses a small abnormal ratio to detect easy anomalies and obtain a small number of high-confidence pseudo-labeled snippets. As the following hard anomalies mining phase uses a larger ratio, the AR-MIL loss becomes non-zero and keeps looking for hard anomalies to meet the abnormal ratio. Meanwhile, a small number of high-confidence pseudo-labeled snippets help the model maintain good predictions on easy normal and abnormal snippets. In the adaptive self-training phase, more suitable abnormal ratios for different abnormal videos are set adaptively to detect different numbers of abnormal snippets. With the proposed three phases, we can train a model to detect more hard anomalies while maintaining good performance on easy anomalies.

F. Network Architecture

We design a network architecture to capture both local and global temporal dependencies between video snippets. As shown in Fig. 5, we first extract RGB features for each snippet in a video via a pre-trained feature extraction network. Then, a temporal convolution is applied to learn the local temporal dependencies between neighboring video snippets. Afterward, a standard multi-head self-attention block [42] is adopted to learn global temporal dependencies between non-local video snippets. Finally, we apply two linear layers and a sigmoid function over the encoded features to predict the abnormal scores of each snippet. For better performance, we also apply a score correction module [15] in the inference stage to fine-tune the predicted snippet-level abnormal scores.

IV. EXPERIMENTS

We perform extensive experiments and compare experimental results with previous works in this section. Moreover, we conduct comprehensive ablation studies to verify our main contributions.

A. Datasets and Metrics

1) Evaluation Datasets: We evaluate our method on three popular video anomaly detection datasets: ShanghaiTech [43], UCF-Crime [1], and XD-Violence [13].

ShanghaiTech is a dataset of 13 scenes with complex lighting conditions and camera angles. It contains 437 campus surveillance videos (330 normal videos, 107 abnormal videos) and over 270,000 training frames. There are various types of anomalies, including cycling, car, fighting, running, etc. The original dataset [43] is proposed for one-class VAD that only uses normal videos for training. To adapt it to W-VAD, we follow the data organization in [41], i.e., 238 training videos and 199 testing videos. UCF-Crime is a large-scale video anomaly detection dataset [1] containing 128 hours of videos. It consists of 1900 long and untrimmed real-world surveillance videos (950 normal videos and 950 abnormal videos), with 13 realistic anomalies such as fighting, road accident, burglary, robbery, etc. The video length varies greatly, and each video may contain diverse backgrounds. By convention [1], [7], we use 1,610 videos in the training set for training and 290 videos in the testing set for evaluation.

XD-Violence is a large-scale and multi-scene dataset [13] with a total duration of 217 hours, containing 4,754 untrimmed videos (2,349 normal videos and 2,405 abnormal videos) with audio signals and weak labels. It contains six physically violent classes, namely, Abuse, Car Accident, Explosion, Fighting, Riot, and Shooting. The videos in this dataset are captured from multiple scenarios, e.g., real-life movies and surveillance cameras. Following the common practice [13], [14], we train our model on the training set with 3,954 videos and test it on the testing set with 800 videos.

2) Evaluation Metrics: Following previous works [1], [7], [13], we use the frame-level area under the ROC curve (AUC) as the primary metric for ShanghaiTech and UCF-Crime, and the frame-level area under the PR curve (AP) as the evaluation metric for XD-Violence. Note that larger AUC and AP values indicate better performance.

B. Implementation Details

We follow [14] to divide each video into 32 snippets and extract 2048-dimensional features for each snippet from the "*mix_5c*" of the pre-trained I3D [44] network. The kernel size of the temporal convolution is 3, and the number of self-attention heads is 8. Two fully connected layers have 32 and 1 nodes, respectively. We train the network with the Adam optimizer [45] using a weight decay of 0.0005. Each mini-batch consists of samples from 30 randomly selected normal and abnormal videos. Moreover, we train the new model in each self-training iteration for 300 epochs on ShanghaiTech with a learning rate of 0.001 and 50 epochs on UCF-Crime and XD-Violence with a learning rate of 0.0001. We empirically set λ^{small} as 0.1 for ShanghaiTech and XD-Violence, and 0.22 for UCF-Crime, λ^{large} as 0.4 for UCF-Crime and XD-Violence, and 0.42 for ShanghaiTech. To balance the loss magnitudes for normal and abnormal videos, we set α as the reciprocal of λ and set β as 0.5.

C. Comparison With State-of-the-Art Methods

In Table I, we compare our multi-phase self-training method with state-of-the-art weakly-supervised video anomaly detection methods [1], [3], [7], [9], [11], [12], [13], [14], [15], [37],

Sup.	Method	UCFC AUC (%)	STech AUC (%)
Un.	Zaheer et al. [10] (ResNet)	71.04	78.93
One-class	Lu et al. [46] (-)	65.51	68.00
	Sohrab et al. [47] (-)	58.50	-
	GODS [29] (I3D)	70.46	-
	AMMC-Net [2] (-)	-	73.70
Weak	Sultani et al. [1] (C3D)	75.41	-
	Zhang et al. [11] (C3D)	78.66	-
	Zhu et al. [37] (PWC)	79.00	-
	Zhong et al. [41] (C3D)	81.08	76.44
	Zhong et al. [41] (TSN^{Flow})	78.08	84.13
	Zhong et al. [41] (TSN)	82.12	84.44
	AR-Net [12] (I3D^{Flow})	-	82.32
	AR-Net [12] (I3D)	-	85.38
	AR-Net [12] (I3D$^{RGB, Flow}$)	-	91.24
	CLAWS [16] (C3D)	83.03	89.67
	Wu et al. [13] (I3D)	82.44	-
	Chang et al. [9] (I3D)	84.62	92.25
	MIST [7] (C3D)	81.40	93.13
	MIST [7] (I3D)	82.30	94.83
	RTFM [14] (C3D)	83.28	91.51
	RTFM [14] (I3D)	84.03	97.21
	Purwanto et al. [19] (Rel.)	85.00	96.85
	Li et al. [15] (C3D)	82.85	94.81
	Li et al. [15] (I3D)	85.30	96.08
	Li et al. [15] (VideoSwin)	85.62	97.32
	BN-SVP [3] (C3D)	-	96.00
	BN-SVP [3] (I3D)	83.39	-
	Baseline (I3D, top-1)	82.25	95.61
	Baseline (I3D, top-k)	83.40	95.63
	Ours (I3D)	**85.79**	**97.48**

We divide the methods into unsupervised (un.), one-class, and weakly-supervised (weak). Best results are in bold.

[41] on UCF-Crime and ShanghaiTech. Selected unsupervised and one-class methods [2], [10], [29], [46], [47] are presented for reference. We also report the results of our baseline that explores traditional self-training with the top-1 and top-k MS-MIL loss for W-VAD (we set k=3 following [14]). Note that we use the same network architecture for the baseline and the proposed method for a fair comparison. We observe that our method outperforms all the previous methods and establishes a new state of the art on UCF-Crime with 85.79% AUC and ShanghaiTech with 97.48% AUC. In particular, our method outperforms MIST [7] and Li et al. [15], which also utilize pseudo labels to guide the model self-training but neglect hard anomalies. Compared with the baseline, our method achieves a remarkable improvement of 3.5% on UCF-Crime and 1.9% on ShanghaiTech.

We also conduct experiments on XD-Violence, and the results are summarized in Table II. Again, our method obtains a new state-of-the-art performance of 83.11% AP, outperforming the latest works (e.g., RTFM [14] and Li et al. [15]). Note that our method even outperforms the multimodal method Wu et al. [13] that uses I3D RGB and audio features for training.

Our AR-MIL loss-based multi-phase self-training method performs better than top-1 and top-k ranking MIL loss-based methods on three datasets. This is because both the top-1 and top-k ranking MIL losses only take a small number of snippets corresponding to the top-k or top-1 abnormal scores to update the model. They can only learn a limited range of normal and

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS IN AP (%) ON
XD-VIOLENCE

Sup.	Method	AP (%)
Un.	SVM baseline (-)	50.78
	Hasan et al. [48] (-)	30.77
Weak	Wu et al. [13] (I3D)	75.41
	Wu et al. [13] (I3D$^{RGB, Audio}$)	78.64
	Chang et al. [9] (I3D)	76.90
	RTFM [14] (C3D)	75.89
	RTFM [14] (I3D)	77.81
	Li et al. [15] (C3D)	75.53
	Li et al. [15] (I3D)	78.28
	Li et al. [15] (VideoSwin)	78.59
	Baseline (I3D, top-1)	78.00
	Baseline (I3D, top-k)	79.04
	Ours (I3D)	**83.11**

Best results are in bold.

TABLE III
ABLATION STUDY ON DIFFERENT VALUES OF ABNORMAL RATIOS λ^{SMALL} AND
λ^{LARGE}

λ^{large} \ λ^{small}	0.1	0.22	0.3
0.3	85.61	85.75	-
0.4	85.72	85.79	85.19
0.5	85.20	85.59	85.42

TABLE IV
SCORE GAP (Δ_s) AND AUC (%) COMPARISON IN DIFFERENT SELF-TRAINING
ITERATIONS OF THE BASELINE AND OUR METHOD ON UCF-CRIME

	Iteration	1	2	3	4	5	6
Baseline	Δ_s	0.15	0.19	0.22	0.20	0.19	0.20
	AUC%	82.19	81.70	82.25	82.14	82.04	81.95
Ours	Δ_s	0.13	0.22	**0.25**	0.23	0.23	0.24
	AUC%	84.30	84.45	85.48	85.30	85.71	**85.79**

The baseline integrates the traditional self-training and the ms-mil loss. Best results are in
bold.

abnormal patterns. Moreover, the lack of high-quality pseudo labels for anomalies prevents these methods from learning hard anomalies and remedying this limitation. In contrast, our AR-MIL loss can learn comprehensive normal and abnormal patterns because all snippets in a video are used to update the model until the predicted abnormal ratio meets a specific ratio. Moreover, our multi-phase self-training paradigm adopts three training phases and the AR-MIL loss with different abnormal ratios λ to help in detecting more easy and hard anomalies, which leads to better performance than the other loss methods.

D. Ablation Studies

In this section, we conduct several ablation studies to verify our main contribution, including the value of λ^{small} and λ^{large}, performance on different iterations, ablation studies of different designs in ARMS, hard abnormal snippet mining, adaptive self-training analysis, model computational complexity, visualization results, and failure analysis. Note that the aforementioned network architecture is used in all experiments.

1) The Value of λ^{small} and λ^{large}: The abnormal ratio λ^{small} and λ^{large} are two important hyper-parameters in our method. We set a relatively small λ^{small} for the bootstrapping phase because the AR-MIL loss can be easily satisfied under this setting (easy anomalies tend to have high abnormal scores). A larger ratio λ^{large} in the hard anomalies mining phase enforces the AR-MIL loss is non-zero and keeps looking for hard anomalies. Here we test its sensitivity to these two hyperparameters on UCF-Crime. Considering that the abnormal events are rare, we set the range of λ^{large} as 0.3-0.5, and the range of λ^{small} as 0.1-0.3. The results are presented in Table III. Our method consistently achieves AUC higher than 85%. Thus, our method is insensitive to λ^{small} and λ^{large}.

2) Performance on Different Iterations: For our ARMS, there is one self-training iteration in the bootstrapping and hard anomalies mining phase, and more than one iteration in the adaptive self-training phase. To demonstrate the performance change of our method in different iterations, we report AUC and the Score Gap (Δ_s) on UCF-Crime in Table IV. Δ_s is calculated

by subtracting the average abnormal score of normal snippets from the average abnormal score of abnormal snippets. A larger score gap indicates that the model can better distinguish anomalies from normal events [7].

Results of the baseline are also reported for reference. We can see that our method achieves higher AUC and Δ_s performance than the baseline in each training iteration. This verifies the effectiveness of our AR-MIL loss and multi-phase self-training. Note that our method achieves a smaller Δ_s in iteration 1 (bootstrapping) compared with the baseline. The reason is that by training with a small abnormal ratio, only a small number of high-quality anomalies can be detected, leading to a small average score of abnormal snippets. Moreover, AUC and Δ_s of our method increase obviously from iteration 1 to 3 while having a small variation in self-training from iteration 3 to 6. This is because our method utilizes the AR-MIL loss and an increasing abnormal ratio in iterations 1–3 to guide the hard anomalies mining, which will discover more hard abnormal snippets. Iterations 3–6 aim to find the optimal abnormal ratios for different abnormal videos to achieve better performance. Here we take the iteration with the best testing performance for baseline to report, while the last iteration for our method because its performance increases with more iterations and finally converges.

3) Ablation Studies of Different Designs in ARMS: We conduct ablation studies of different designs in our method on UCF-Crime and ShanghaiTech, as shown in Table V. Our multi-phase self-training method involves three training phases: bootstrapping (BST), hard anomalies mining (HAM), and adaptive self-training (AS). The baseline that integrates the traditional self-training and the MS-MIL loss (top-1) achieves only 82.25% AUC on UCF-Crime and 95.61% AUC on ShanghaiTech, which are similar to the results obtained by another self-training method [7]. The performance of integrating traditional self-training and our AR-MIL loss is comparable to or even better than the baseline on UCF-Crime, which is a difficult dataset and contains complex abnormal events and diverse backgrounds. It validates the effectiveness of our AR-MIL
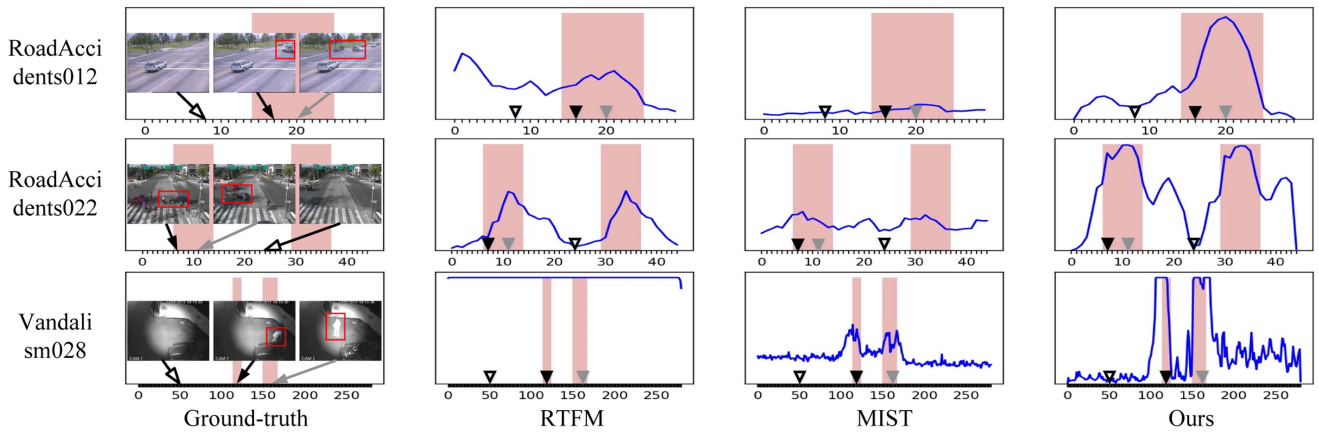
Fig. 6. Visualization cases of ground truth and anomaly score curves of different methods on UCF-Crime. The white, gray, and black triangles denote the locations of the normal, easy abnormal, and hard abnormal frames displayed on the left, respectively. The blue curves represent the anomaly predictions of different methods. The pink background corresponds to the ground-truth abnormal regions.

TABLE V
ABLATION STUDIES IN AUC (%) ON UCF-CRIME (UCFC) AND
SHANGHAITECH (STECH)

| $\mathcal{L}_{MS-MIL}$ | $\mathcal{L}_{AR-MIL}$ | Traditional | ARMS | | | UCFC AUC (%) | STech AUC (%) |
			BST	HAM	AS		
✓		✓				82.25	95.61
	✓	✓				84.35	95.06
	✓		✓			84.30	91.68
	✓		✓	✓		84.45	94.29
	✓		✓		✓	84.49	96.09
	✓		✓	✓	✓	**85.79**	**97.48**

BST, HAM, and AS are short for bootstrapping, hard anomalies mining, and adaptive self-training of our multi-phase self-training, respectively. Best results are in bold.
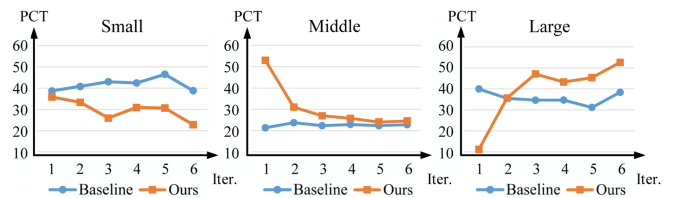

Fig. 7. Ablation studies of hard anomalies mining over different self-training iterations of the baseline and our method. The range of the abnormal score is divided into three groups: small S [0, 0.4], middle M (0.4, 0.8], and large L (0.8, 1]. PCT represents the percentage of true abnormal snippets' abnormal scores that fall into S interval, M interval, and L interval.

loss in terms of robustness to the hard anomalies in abnormal videos, while the MS-MIL loss fails to address such hard abnormal events. The bootstrapping phase achieves comparable or inferior performance on UCF-Crime and ShanghaiTech because this phase aims to initialize the training model with a small anomaly ratio. By mining hard anomalies with a larger abnormal ratio in the AR-MIL loss, the model can learn sufficient abnormal patterns to detect anomalies, especially the hard ones. The AUC is boosted to 84.45% and 94.29% on UCF-Crime and ShanghaiTech, respectively. When the adaptive self-training is directly applied after the bootstrapping, we obtain better performance than the fixed abnormal ratio in the hard abnormal mining phase on the two datasets. This is because different abnormal videos contain different numbers of abnormal snippets, and our adaptive self-training phase is able to learn suitable abnormal ratios for different abnormal videos to help in learning different numbers of abnormal snippets. When all three phases are included, the AUC increases to 95.79% and 97.48% on the two datasets, respectively.

4) Hard Abnormal Snippet Mining: To quantify the difficulty of abnormal snippets, we divide the estimated abnormal scores into three intervals as small S [0, 0.4], middle M (0.4, 0.8], and large L (0.8, 1]. We focus on the abnormal scores of true abnormal snippets and calculate the percentage of snippets that fall into these three groups. A large percentage of snippets falling
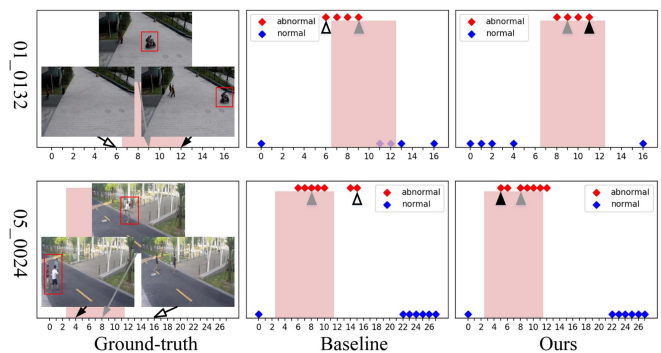

Fig. 8. Visualizations of ground truth and pseudo labels generated by different methods on ShanghaiTech. The red and blue diamonds denote snippets with abnormal and normal pseudo labels, respectively. The white, gray, and black triangles denote the locations of the normal, easy abnormal, and hard abnormal frames displayed on the left, respectively. The pink background corresponds to the ground-truth abnormal regions.

into S means that many hard abnormal snippets have been predicted incorrectly. We present the quantitative results of the baseline and our method in Fig. 7. We can see that, as the number of self-training iterations increases, the percentage in the L interval of our method has an obvious increase, while the percentages in the S and M intervals decrease greatly. In contrast, the
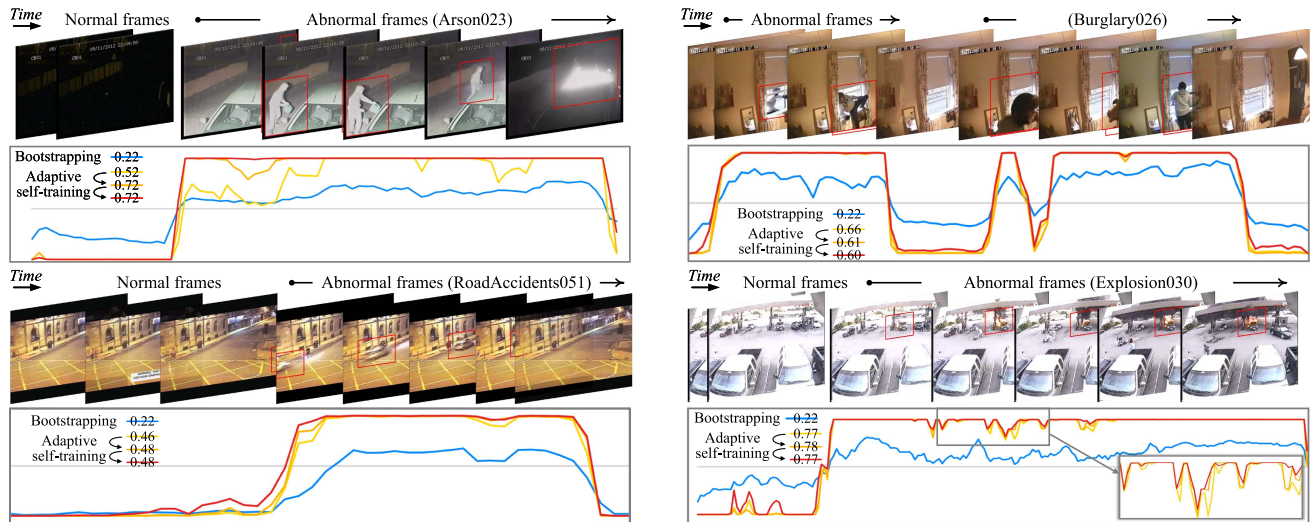
Fig. 9. Visualization of abnormal ratios (text on the legend) used in different self-training iterations of our method on training abnormal videos.
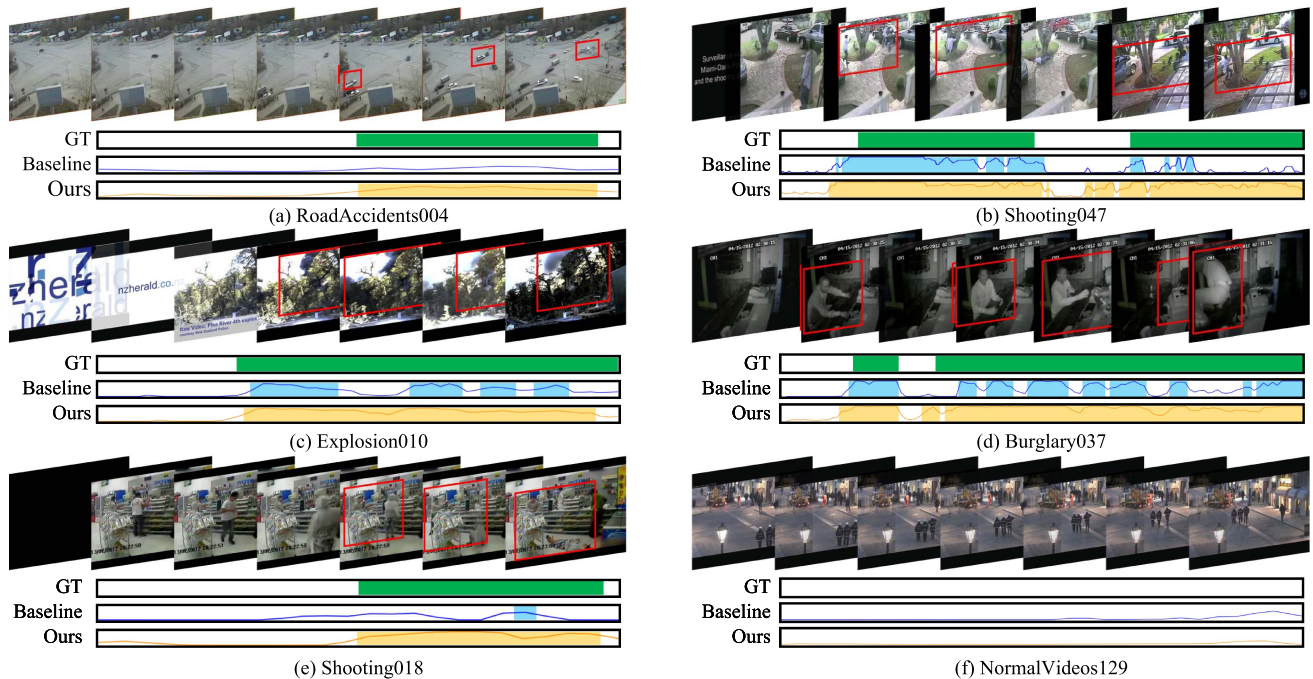

Fig. 10. Visualization of the testing results of the baseline and ARMS on UCF-Crime. Predictions higher than 0.5 are selected as abnormal frames. Transparent frames represent normal frames.

percentages in all three intervals of the baseline have small variations in different self-training iterations. This verifies that our method has the ability to discover more hard abnormal snippets in each new training iteration, thanks to our AR-MIL loss and multi-phase self-training. Meanwhile, the better AUC and Score Gap (Δ_s) performance in Table IV indicate that the normal snippets can also be detected well.

In order to provide a more comprehensive comparison for detecting hard anomalies, Fig. 6 visualizes the detection results of the proposed ARMS and other state-of-the-art methods [7], [14]. As we can see, our method detects more easy and hard abnormal frames than other two methods, especially in *RoadAccidents022* where our method predicts higher scores for the hard anomalies of the car accident than other two methods. Moreover, the visualized detection results in Figs. 1 and 10 also verify that our method predicts more hard abnormal frames than baseline, e.g., the shooting frames in *Shooting047*, the imperceptible explosion frames in *Explosion010*.

5) Pseudo Label Analysis: In order to analyze the generated pseudo labels, we present the precision of pseudo labels (Prec.) and testing AUC in three self-training iterations of our method and baseline in Table VI. Results are obtained on the training

TABLE VI
PRECISION OF PSEUDO LABELS (PREC.) AND TESTING AUC IN THREE
SELF-TRAINING ITERATIONS OF OUR METHOD (GREEN) AND BASELINE
(TRANSPARENT) ON SHANGHAITECH

Iter.	1	2	3	1	2	3
Prec. %	-	78.21	78.46	-	78.97	**80.76**
AUC %	95.61	94.97	93.45	91.68	94.29	**97.48**

TABLE VII
COMPARISON OF MODEL SIZE AND COMPUTATIONAL COMPLEXITY ON
SHANGHAITECH

Method	#Params (M)	FLOPs (G)
MIST [7]	31.00	45.68
RTFM [14]	24.72	0.79
ARMS (Ours)	10.53	0.34

set of ShanghaiTech, which provides frame-level annotations. We also provide some examples of pseudo labels generated by the baseline and our method in Fig. 8. We can see that higher-quality pseudo labels improve the performance, and our method performs better than baseline with higher-quality pseudo labels. Moreover, the visualized cases demonstrate that both baseline and our method can recognize easy anomalies, but our method generates pseudo labeled data with more true normal and hard abnormal frames than baseline. For example, in *01_0132*, our method generates a correct label for the hard abnormal frame with a running bike present at the panel boundaries.

6) Adaptive Self-Training Analysis: We present several examples in Fig. 9 to show the predicted results and the adaptive abnormal ratios (text on the legend) in different training phases of our method, where the visualized videos are abnormal videos in the UCF-Crime training set. As the adaptive self-training continues, we actually generate appropriate abnormal ratios λ for different abnormal videos in the AR-MIL loss. In the last iteration of the adaptive self-training, we also achieve the best anomaly detection performance.

7) Model Size, Speed, and Computational Complexity: We present the comparison of model size and computational complexity between our method and other methods [7], [14] on ShanghaiTech in Table VII. The results of all methods are obtained by running their official code on a single NVIDIA RTX 3090 GPU. We report the results of MIST [7] including only its second stage, where an I3D network is fine-tuned by pseudo labels. RTFM [14] and our method directly adopt I3D features as the model input. This explains why our method and RTFM have much lower FLOPs than MIST. Moreover, when the I3D extraction time is included in our method and RTFM, three methods have similar speeds as 21 snippets per second. The results in Table VII demonstrate that our model is light and efficient.

8) Visualization Results: We visualize the predictions of the baseline and our method in Fig. 10. Our method can detect the abnormal events exactly and predict abnormal scores of the normal frames very close to zero. In Fig. 10(a), the baseline method predicts small abnormal scores for the abnormal frames of *Road-Accidents004*, while our method detects all abnormal frames, including hard abnormal frames of over-speed car rushing and easy abnormal frames of two cars crashing. This is because the
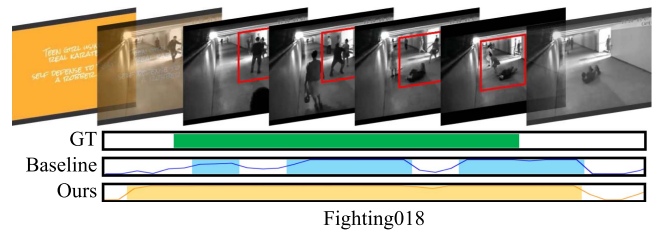

Fig. 11. Failure case of our method on UCF-Crime. Predictions higher than 0.5 are selected as abnormal frames. Transparent frames represent normal frames.
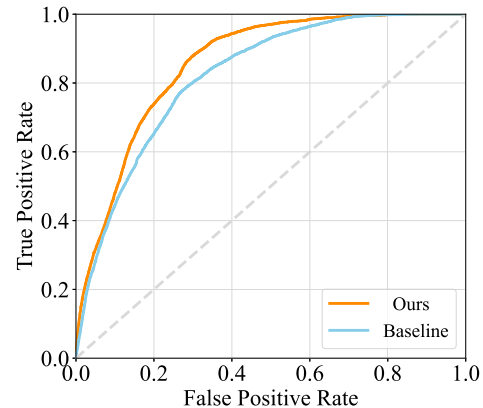

Fig. 12. Receiver Operating Characteristic (ROC) curves of the baseline and our method on the testing set of UCF-Crime.

abnormal ratio-based loss in our ARMS helps discover more hard anomalies. Furthermore, our method can detect almost all explosion frames in Fig. 10(c) *Explosion010*, while the baseline method has an incomplete prediction.

9) Failure Analysis: We present a failure case in Fig. 11. As we can see, there could be over-predictions at the abnormal/normal boundary w.r.t. annotation. This is because our method focuses on detecting more hard anomalies and the abnormal event boundary is prone to have hard anomalies. What's more, the boundary is ambiguous (even to humans), and precisely locating the boundary is still an unsolved problem. However, our method does not lead to increased false positives. We draw Receiver Operating Characteristic (ROC) curves of the baseline and our method in Fig. 12. It shows that our method decreases FPR at all TPRs. As our AR-MIL loss looks for both easy and hard anomalies in the whole video, more anomalies (true positives) and thus fewer normal snippets (false positives) will be retrieved at a given number of anomaly detections.

E. Limitation and Future Work

Although our method has the great advantage of detecting hard anomalies, there are still some limitations in our method. As the abnormal ratios in the first two training phases are set manually, we still need to predefine these two values, though our multi-phase self-training method is insensitive to λ in the first two phases (as shown in Table III). In addition, as our method focuses on detecting hard anomalies and the abnormal event boundary is prone to have hard anomalies, our method could

have over-predictions at the abnormal boundary w.r.t. annotations. In our further work, we will improve our study in two aspects. On the one hand, we will consider the intra-video and inter-video differences to explore the intrinsic normal and abnormal differences to overcome the over-prediction problem. On the other hand, we will take into account the self-paced learning paradigm to explore the knowledge of continuous epochs to learn the abnormal ratios and the abnormal scores at the same time.

V. Conclusion

This paper introduces a novel Abnormal Ratios guided Multiphase Self-training (ARMS) for W-VAD. It includes a new abnormal ratio-based MIL (AR-MIL) loss and a new multi-phase self-training paradigm. The AR-MIL loss enforces a minimum ratio of abnormal snippets in an abnormal video and no abnormal snippets in a normal. It better finds hard anomalies in abnormal videos than the prior MS-MIL loss. Three training phases (bootstrapping, hard anomalies mining, and adaptive self-training) in our method address pseudo labeling on easy anomalies, mining hard anomalies and setting adaptive abnormal ratios for different videos in a unified framework. Extensive experiments indicate that ARMS outperforms prior state-of-the-art methods and has a great advantage of detecting hard anomalies, which is one of the greatest challenges in W-VAD.

References

[1] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6479–6488.

[2] R. Cai, H. Zhang, W. Liu, S. Gao, and Z. Hao, "Appearance-motion memory consistency network for video anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 938–946.

[3] H. Sapkota and Q. Yu, "Bayesian nonparametric submodular video partition for robust anomaly detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3202–3211.

[4] Y. Zhai et al., "Action coherence network for weakly-supervised temporal action localization," *IEEE Trans. Multimedia*, vol. 24, pp. 1857–1870, 2022.

[5] K. Xia et al., "Exploring action centers for temporal action localization," *IEEE Trans. Multimedia*, early access, Mar. 03, 2023, doi: 10.1109/TMM.2023.3252176.

[6] S. Zhou et al., "Large margin learning in set-to-set similarity comparison for person reidentification," *IEEE Trans. Multimedia*, vol. 20, pp. 593–604, 2018.

[7] J.-C. Feng, F.-T. Hong, and W.-S. Zheng, "MIST: Multiple instance self-training framework for video anomaly detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14004–14013.

[8] K. Xu, T. Sun, and X. Jiang, "Video anomaly detection and localization based on an adaptive intra-frame classification network," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 394–406, Feb. 2020.

[9] S. Chang, Y. Li, S. Shen, J. Feng, and Z. Zhou, "Contrastive attention for video anomaly detection," *IEEE Trans. Multimedia*, vol. 24, pp. 4067–4076, 2022.

[10] M. Z. Zaheer et al., "Generative cooperative learning for unsupervised video anomaly detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14724–14734.

[11] J. Zhang, L. Qing, and J. Miao, "Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection," in *Proc. Int. Conf. Image Process.*, 2019, pp. 4030–4034.

[12] B. Wan, Y. Fang, X. Xia, and J. Mei, "Weakly supervised video anomaly detection via center-guided discriminative learning," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2020, pp. 1–6.

[13] P. Wu et al., "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 322–339.

[14] Y. Tian et al., "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4955–4966.

[15] S. Li, F. Liu, and L. Jiao, "Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 1395–1403.

[16] M. Z. Zaheer, A. Mahmood, M. Astrid, and S.-I. Lee, "CLAWS: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 358–376.

[17] H. Lv et al., "Localizing anomalies from weakly-labeled videos," *IEEE Trans. Image Process.*, vol. 30, pp. 4505–4515, 2021.

[18] P. Wu and J. Liu, "Learning causal temporal relation and feature discrimination for anomaly detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3513–3527, 2021.

[19] D. Purwanto, Y.-T. Chen, and W.-H. Fang, "Dance with self-attention: A new look of conditional random fields on anomaly detection in videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 173–183.

[20] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 341–349.

[21] B. Wan, W. Jiang, Y. Fang, Z. Luo, and G. Ding, "Anomaly detection in video sequences: A benchmark and computational model," *IET Image Process.*, vol. 15, no. 14, pp. 3454–3465, 2021.

[22] G. Pang, C. Yan, C. Shen, A. v. d. Hengel, and X. Bai, "Self-trained deep ordinal regression for end-to-end video anomaly detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12170–12179.

[23] D. Gong et al., "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1705–1714.

[24] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14360–14369.

[25] K. Ying et al., "CTVIS: Consistent training for online video instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 899–908.

[26] L. Wu, Y. Wang, H. Yin, M. Wang, and L. Shao, "Few-shot deep adversarial learning for video-based person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 1233–1245, 2020.

[27] S. Yang, L. Wu, A. Wiliem, and B. C. Lovell, "Unsupervised domain adaptive object detection using forward-backward cyclic adaptation," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 124–142.

[28] K. Liu and H. Ma, "Exploring background-bias for anomaly detection in surveillance videos," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 1490–1499.

[29] J. Wang and A. Cherian, "GODS: Generalized one-class discriminative subspaces for anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8200–8210.

[30] G. Yu et al., "Deep anomaly discovery from unlabeled videos via normality advantage and self-paced refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13967–13978.

[31] M.-I. Georgescu et al., "Anomaly detection in video via self-supervised and multi-task learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12737–12747.

[32] L. Wu et al., "Pseudo-pair based self-similarity learning for unsupervised person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 4803–4816, 2022.

[33] M. Li et al., "Video pivoting unsupervised multi-modal machine translation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3918–3932, Mar. 2023.

[34] C. Yan et al., "ZeroNAS: Differentiable generative adversarial networks search for zero-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9733–9740, Dec. 2022.

[35] L. Zhang et al., "TN-ZSTAD: Transferable network for zero-shot temporal activity detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3848–3861, Mar. 2023.

[36] D. Liu et al., "Generative metric learning for adversarially robust open-world person re-identification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 1, pp. 1–19, 2023.

[37] Y. Zhu and S. Newsam, "Motion-aware feature for improved video anomaly detection," in *Proc. Brit. Mach. Vis. Conf.*, 2019.

[38] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Int. Conf. Mach. Learn. Workshop*, 2013, paper 896.

[39] J. Kim et al., "Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 14567–14579.

[40] S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang, "Semi-supervised temporal action detection with proposal-free masking," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 663–680.

[41] J.-X. Zhong et al., "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1237–1246.

[42] A. Vaswani et al., "Attention is all you need," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[43] 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.*, 2018, pp. 6536–6545.

[44] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4724–4733.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[46] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MAT-LAB," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2013, pp. 2720–2727.

[47] F. Sohrab, J. Raitoharju, M. Gabbouj, and A. Iosifidis, "Subspace support vector data description," in *Proc. Int. Conf. Pattern Recognit.*, 2018, pp. 722–727.

[48] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 733–742.

Haoyue Shi (Student Member, IEEE) received the B.S. degree in software engineering from Northwest A&F University, Yangling, China in 2019. She is currently working toward the Ph.D. degree with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China. Her research interests include computer vision, image/video processing, analysis and understanding.
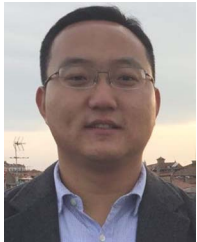
Le Wang (Senior Member, IEEE) received the B.S. and Ph.D. degrees in control science and engineering from Xi'an Jiaotong University, Xi'an, China, in 2008 and 2014, respectively. From 2013 to 2014, he was a Visiting Ph.D. Student with the Stevens Institute of Technology, Hoboken, NJ, USA. From 2016 to 2017, he was a Visiting Scholar with Northwestern University, Evanston, IL, USA. He is currently a Professor with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research interests include computer vision, pattern recognition, and machine learning.
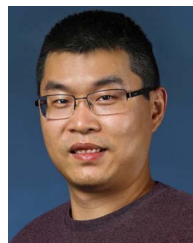
Sanping Zhou (Member, IEEE) received the Ph.D. degree in control science and engineering from Xi'an Jiaotong University, Xi'an, China, in 2020. From 2018 to 2019, he was a Visiting Ph.D. Student with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. He is currently an Assistant Professor with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research interests include machine learning, deep learning, and computer vision, with a focus on person re-identification, salient object detection, medical image segmentation, image classification, and visual tracking.

Gang Hua (Fellow, IEEE) received the B.S. and M.S. degrees in automatic control engineering from Xi'an Jiaotong University, Xi'an, China, in 1999 and 2002, respectively, and the Ph.D. degree in electrical engineering and computer science from Northwestern University, Evanston, IL, USA, in 2006. He is the author of more than 200 peer reviewed publications in prestigious international journals and conferences. He holds 19 U.S. patents and has 15 more U.S. patents pending. He is currently the Vice President and Chief Scientist of Wormpex AI Research. He is an Associate Editor for IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and MVA. He is a General Chair of ICCV'2025, the Program Chair of CVPR' between 2019 and 2022, and the Area Chair of CVPR' between 2015 and 2017 and ICCV' between 2011 and 2017. He was the recipient of the 2015 IAPR Young Biometrics Investigator Award. He is an IAPR Fellow, and an ACM Distinguished Scientist.

Wei Tang (Member, IEEE) received the B.E. and M.E. degrees from Beihang University, Beijing, China, in 2012 and 2015 respectively, and the Ph.D. degree in electrical engineering from Northwestern University, Evanston, IL, USA, in 2019. He is currently an Assistant Professor with the Department of Computer Science, University of Illinois Chicago, Chicago, IL. His research interests include computer vision, pattern recognition and machine learning.