Article

Not peer-reviewed version

AFR: Adaptive Feature Refinement for Fine-Grained Video Anomaly Detection

Dezhi An *, Wenqiang Liu , Jun Lu , Shengcai Zhang

Posted Date: 31 March 2025

doi: 10.20944/preprints202503.2351.v1

Keywords: video anomaly detection; small-object attention; contrastive language-image pre-training; small target anomalies; feature pyramid network

 doi:10.20944/preprints202503.2351.v1

Article

AFR: Adaptive Feature Refinement for Fine-Grained Video Anomaly Detection

Dezhi An *, Wenqiang Liu, Jun Lu and Shengcai Zhang

School of Cyber Security, Gansu University of Political Science and Law, Lanzhou 730070, Gansu, China
* Correspondence: andezhi@gsupl.edu.cn

Abstract: Video anomaly detection (VAD) remains a challenging task, especially when identifying occluded small scale and transient anomalies in complex contexts. Existing methods (such as optical flow analysis, trajectory modeling, and sparse coding) often ignore enhancement of fine-grained target features and lack adaptive precise attention mechanisms, thus limiting the robustness and generalization ability of anomaly detection. To this end, we propose an adaptive feature refinement (AFR), AFR method to improve the performance of small-scale anomaly detection. The AFR method integrates small-object attention module (SAM) into the feature pyramid network (FPN) of the clip-driven multi-scale instance learning architecture to adaptively enhance the feature representation of key areas. In addition, we combine the comparison language-image pre-training (CLIP) model to enrich semantic information and improve the generalization ability across scenes. Specifically, the SAM module guides the model to pay attention to the discriminant patterns of small-scale anomalies through channel recalibration and spatial attention mechanisms, while the semantic prior of the CLIP model further strengthens the expression ability of visual features. The AFR method combines optimization of SAM and CLIP to show superior generalization performance in cross-scale and cross-scene anomaly detection tasks. Extensive experiments on two common benchmark datasets UCF-Crime and XD-Violence show that the AFR method outperforms existing state-of-the-art methods in performance, verifying its effectiveness and migration in real-world video anomaly detection tasks.

Keywords: video anomaly detection; small-object attention; contrastive language-image pre-training; small target anomalies; feature pyramid network

1. Introduction

Video anomaly detection (VAD) [1–7] aims to detect anomalous events in untrimmed videos, such as accidents, explosions, and violence, which have a wide range of applications in intelligent surveillance [8] and violence alerting [9]. However, anomalous events are diverse and occur quite rarely, which makes it time- consuming and labor-intensive to annotate them in detail. To reduce annotation costs, weakly supervised VAD (WSVAD) has gained popularity, which relies solely on video-level labels (normal vs. abnormal) during the training phase.In this model, introducing a tandem joint attention mechanism between SCAM and SAM is the key to improving performance.

In VAD, videos consist of sequences of frames, resulting in complex, high-dimensional spatiotemporal data. Detecting anomalies within such complex data necessitates methods capable of effectively capturing spatial, temporal, spatiotemporal, and textual features. To address these challenges,numerous VAD methods and deep feature extractors have been introduced in academic research, which has played a significant role in advancing the current state-of-the-art.In the past, video anomaly detection methods are mainly divided into traditional feature modeling and deep learning.Methods based on traditional feature modeling rely on hand-designed features such as optical flow, locus, space-time interest points, etc., and detect anomalies through statistical models such as Gaussian mixture model, sparse coding, and topic model. Examples include the use of manual features, including spatiotemporal gradient [10], oriented gradient histogram (HOG) features [11,12] and optical flow

 doi:10.20944/preprints202503.2351.v1

histogram (HOF) [13] in the time rectangular body. These features were chosen for their effectiveness in capturing appearance and motion information in a spatiotemporal context.These tend to be global or capture only large-scale motion patterns, and are difficult to accurately characterize for small targets or transient local anomalies (such as a pedestrian suddenly falling, a small object being thrown). In addition, such methods usually use fixed Windows or areas for modeling, resulting in weak anomaly perception of local areas, easy to be affected by complex background and occlusion, and difficult to capture fine-grained abnormal behavior.

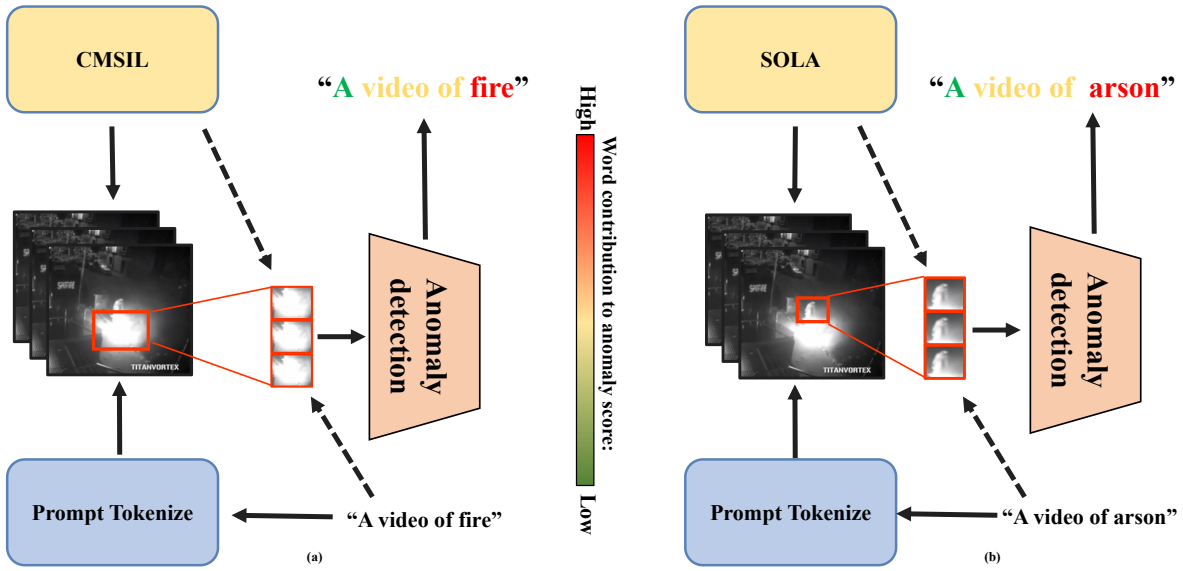

Figure 1. Illustrates the difference between the original CMSIL anomaly detection (a) and our proposed method (b). (a) Previously proposed CLIP-driven multi-scale instance learning tends to focus on close-range anomalous fragments, and fails to identify distant anomalous fragments. (b) Our SOLA anomaly recognizes small targets at a distance to make better identification.

With the rapid development of deep learning, methods based on deep learning have gradually become the mainstream in the field of VAD video anomaly detection. Compared to traditional methods, these deep learning-based techniques show a strong advantage in the richness of feature representation, the ability to learn end-to-end, and the modeling of complex data patterns. They can automatically learn multi-level feature representations of data, from low-level pixel information to high-level semantic information, which greatly improves the adaptability and generalization ability of tasks. Over the past decade, deep learning methods have gained widespread adoption and popularity with the continuous optimization of deep neural network architectures, advances in computing hardware such as Gpus and Tpus, and the availability of large-scale data sets. Specifically, these methods utilize powerful feature extractors to capture meaningful spatio-temporal features. Examples include convolutional neural networks [14], autoencoders [15], gan [16], vision converters [17,18], and visual language models [19,20]. While deep learning methods have improved detection performance, they still present some challenges. For example, existing methods rely heavily on large-scale training data and lack generalization, especially when dealing with no exception classes or changes in data distribution. In addition, abnormal behaviors in video are often represented by small-range changes in local areas or short temporal events. These anomalies are usually small in scale and occupy a limited proportion in the whole video frame, which makes it difficult for traditional feature fusion methods to effectively capture these small targets, especially when the background is more complex or the abnormal targets are similar in appearance to normal behaviors. Models may have difficulty picking up on subtle abnormal changes.

In order to overcome the shortcomings of traditional methods in anomaly detection of small targets, this study introduces small-object attention module (SAM) into the feature pyramid network (FPN) structure of clip-driven multi-scale instance learning (CMSIL) [21] model. It is designed to

 doi:10.20944/preprints202503.2351.v1

improve the ability of the model to perceive fine-grained anomalies. Traditional video anomaly detection methods usually rely on global features or fixed area modeling, resulting in poor performance in the detection of small targets or transient anomalies, especially in complex background and occlusion environment, the model is often difficult to capture subtle abnormal behavior. SAM adaptively adjusts the importance of each channel and spatial position in the input feature map, so that the model can dynamically focus on key information according to the current task, and improve the response ability to local anomalies. Specifically, SAM includes channel attention module and spatial attention module. The former enhances the sensitivity of the model to key features by weighting the features of different channels, while the latter weights the spatial dimension to highlight abnormal areas and help the model identify small targets and local anomalies. Through this mechanism, SAM can improve the detection ability of small targets in the process of multi-scale feature fusion, while reducing the interference of complex background or occlusion on model performance. Especially in the face of fine-grained anomalies, SAM can dynamically adjust the attention according to the specific requirements of the task and enhance the robustness of the model. By introducing SAM, the anomaly detection capability of the model in complex scenes has been significantly improved, which not only effectively handles small target anomalies, but also enhances the sensitivity to local details, providing a more effective and adaptive solution for video anomaly detection.In a nutshell, our main contributions are as follows:

- We integrate SAM into the FPN structure of the CMSIL model, which effectively enhances the detection ability of small targets and local anomalies, overcoming the limitations of traditional methods.
- We enhance the model's robustness in complex backgrounds by integrating SAM into multi-scale feature fusion, effectively mitigating the impact of background noise on anomaly detection accuracy.
- We conduct extensive experiments on two challenging datasets. The results demonstrate that the proposed method outperforms several state-of-the-art approaches.

2. Related Work

2.1. Video Anomaly Detection

Video anomaly detection (VAD) has evolved from conventional handcrafted techniques to sophisticated deep learning models. Early works predominantly relied on statistical modeling and handcrafted features, such as optical flow, trajectory analysis, and histogram-based descriptors (e.g., HOG), often integrated with Gaussian Mixture Models (GMM) [22]. While effective in constrained settings, these approaches struggle to generalize to complex and dynamic environments due to limited capacity in capturing high-level semantics and long-range dependencies. The emergence of deep learning has brought significant advancements in VAD. Methods based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have demonstrated strong potential. For instance, C3D [23] utilizes 3D convolutions to model spatiotemporal patterns, and ConvLSTM [24] integrates temporal modeling via LSTM gates. However, CNNs are inherently limited to local receptive fields, and RNNs often suffer from vanishing gradients, hindering their ability to capture long-term dependencies and global scene context.

To overcome these limitations, Transformer-based models have been introduced in the VAD field. Vision Transformer (ViT)-style architectures, such as VAD-Transformer [25] and TADA [26], employ self-attention mechanisms to capture global dependencies and improve temporal reasoning. Nonetheless, these models are typically optimized for large-scale anomalies and lack tailored strategies for detecting fine-grained or small-target anomalies. This leads to performance degradation in scenes with subtle or transient abnormal events. Recent research has explored adaptive attention strategies to better capture such subtle cues. The integration of the Small-object Attention Module (SAM) into the Feature Pyramid Network (FPN) structure of the Clip-driven Multi-scale Instance Learning (CMSIL) framework aims to address this issue. Inspired by lightweight attention mechanisms [27], SAM adaptively recalibrates channel and spatial features, enhancing the model's sensitivity to localized

 doi:10.20944/preprints202503.2351.v1

anomalies. This aligns with prior work on multi-modal and fine-grained feature conditioning [28,29], as well as improvements in visual-semantic fusion for generation and retrieval tasks [30,31].

2.2. Small-Object Anomaly Detection

Detecting small-target anomalies in video surveillance is particularly challenging due to weak visual saliency, cluttered backgrounds, scale variation, and fragmented temporal cues. Traditional approaches based on optical flow [32] and trajectory analysis often fail under these conditions, particularly when objects are low-resolution or obscured. Deep learning-based alternatives [33] offer improved generalization but tend to prioritize high-level semantics over fine-grained details, leading to degraded performance for small-object anomalies. To bridge this gap, multi-scale architectures such as FPN [34] have been widely adopted for their ability to combine features across scales. However, standard FPNs use a static feature aggregation strategy that lacks adaptive focus on small anomalies. Recent advances have introduced spatial-channel attention mechanisms that enhance feature discrimination at multiple resolutions [35,36]. Nevertheless, these techniques are typically designed for general object detection and fall short in addressing anomaly-specific challenges such as transient motion patterns and semantic ambiguity.

To this end, we introduce a task-aware SAM into the CMSIL framework. SAM builds upon joint spatial-channel attention concepts [37], while being optimized for small-object anomaly detection through two core designs: (1) A localized spatial saliency mechanism that adaptively boosts feature responses in fine-grained regions without amplifying irrelevant noise; and (2) a channel-wise attention recalibration strategy that counteracts the suppression of low-level anomaly cues by dominant high-level features. These mechanisms draw on principles of targeted modulation observed in person generation [38] and remote sensing, where small-object discriminability is critical. Our proposed enhancements significantly improve anomaly recognition in benchmark VAD datasets, such as UCF-Crime and XD-Violence, particularly in challenging scenarios involving small targets. This validates the necessity of domain-specific attention strategies for fine-grained anomaly detection and underscores the general utility of SAM across vision tasks.

3. Proposed Method

3.1. Overview

Figure 2 shows the overall framework of the proposed approach. Given an input video clip (that is, a set of continuous frames), firstly, the ConvTransformer backbone network is used for feature extraction to model local spatial information and global temporal dependence at the same time, so as to obtain the advanced spatio-temporal features in the video clip, including object appearance, motion pattern and background information. Then, multi-scale feature fusion is carried out based on feature pyramid network (FPN) to enhance the ability of feature expression at different scales. In order to further improve the ability to perceive abnormal areas, the network introduces small-object channel attention module (SCAM) and small-object attention module (SAM) to adaptively optimize feature representation. Among them, SCAM emphasizes the key features related to abnormal patterns and suppresses redundant information by calculating the importance weight of channels; SAM highlights abnormal regions and improves the adaptability of the model to scene complexity by aggregating global and local spatial features. The attention-enhanced features are then input into the classification header (Cls Head) for anomaly detection and generate anomaly region prediction. In the reasoning phase, the model reconstructs the features of the test video clips by matching the stored normal pattern features, and identifies the abnormal regions based on the reconstruction error. Through the introduction of hierarchical attention mechanism, this method can effectively improve the compactness of normal patterns and the separability of abnormal patterns, thus enhancing the robustness and generalization ability of the model in complex scenes.

 doi:10.20944/preprints202503.2351.v1

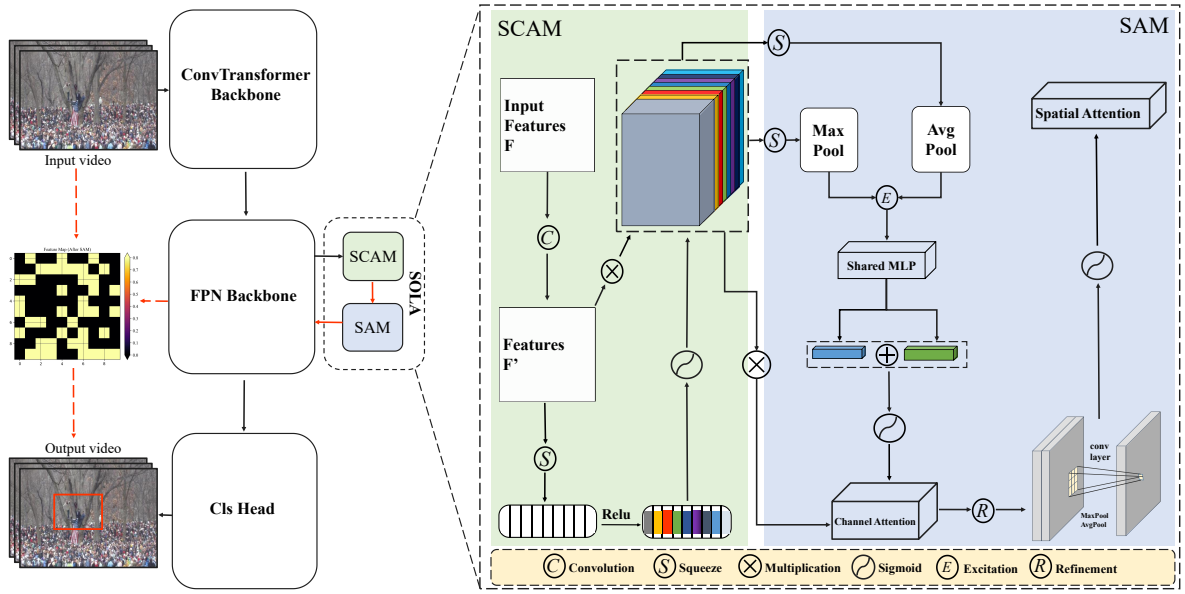

Figure 2. Overall framework.The SCAM and SAM modules work together to improve the performance of small target anomaly detection. SCAM represents the small target channel attention module, which is responsible for enhancing the channel characteristics of the small target. SAM represents the small-objective spatial attention module, captures the spatial characteristics of the small-objective and predicts anomaly scores through feature fusion.

3.2. Small Object Context-Aware Attention Module

The small target context aware attention module (SCAM) aims to optimize the feature expression of small target areas,enhance the response of abnormal areas, and reduce interference from background noise. Traditional methods based on global feature modeling usually find it difficult to accurately capture the information of small target areas, because during the global feature extraction process, the model tends to focus on large target areas or features with higher significance,and ignores smaller anomalies in space. To solve this problem, SCAM adopts a strategy based on channel attention and local feature extraction, so that the features of small target areas can get higher attention during feature extraction and fusion.Specifically,we first perform local feature extraction on the input feature map to enhance the fine-grained feature expression of small targets;then,we use the channel attention mechanism to adaptively adjust the weights of different channels,making the response of small target areas more prominent, thereby improving detection accuracy.

Suppose the input feature is $F \in \mathbb{R}^{C\times H\times W}$, C represents the number of channels, and H and W represent the height and width of the feature map respectively.In order to better enhance the local features of small goals, we first use local perceptual convolution operation $\varphi(\cdot)$ to transform the input features to obtain the enhanced features

$$F \in \mathbb{R}^{C\times H\times W}. \tag{1}$$

On this basis,we introduce a channel attention mechanism to extract global channel features through maximum pooling and average pooling operations to form a channel description vector b:

$$b = \mathrm{MLP}\Big(\mathrm{MaxPool}(F') + \mathrm{AvgPool}(F')\Big). \tag{2}$$

Then,the channel attention weight W_c is calculated using the sigmoid activation function and weight F':

$$W_c = \sigma(b), \tag{3}$$

$$F'' = F' \times W_c. \tag{4}$$

 doi:10.20944/preprints202503.2351.v1

This channel attention mechanism can adaptively adjust the information weights of different channels, making the response of small target areas more prominent in the feature map.

3.3. Small-Object Attention Module

Although SCAM improves the feature expression ability of small target areas through channel attention mechanism and context modeling, its information enhancement in the spatial dimension is still limited. Since small target anomalies often exhibit highly localized properties, spatial attention mechanisms are crucial for small target anomaly detection. Therefore, we further propose the small-objective attention module, whose core idea is to further optimize the characteristic response of the small-objective area after SCAM.

Suppose the output feature diagram of SCAM is $X \in \mathbb{R}^{C\times H\times W}$, SAM performs two key steps: channel attention enhancement and spatial attention enhancement in sequence.First, the channel attention enhancement part is similar to SCAM, and we extract the channel description vector through maximum pooling and average pooling operations a:

$$a = \mathrm{MLP}\Big(\mathrm{MaxPool}(X) + \mathrm{AvgPool}(X)\Big). \tag{5}$$

Then, the normalized weight is calculated by sigmoid and weighted X to improve the channel response of the small target area:

$$X' = X \times \sigma(a). \tag{6}$$

In this way, we can further optimize the feature expression of small target anomaly regions in the channel dimension. However, relying solely on channel attention may not be enough to comprehensively improve the abnormal detection capabilities of small targets.Since the spatial distribution of small target anomalies is usually sparse, we further introduce a spatial attention mechanism to highlight the spatial distribution of anomalies.We first compute the spatial attention map S of the feature map X':

$$S = \sigma\big(f(X')\big). \tag{7}$$

$f(\cdot)$ represents a lightweight spatial transformation function, such as a convolution operation or a local normalization operation. We then weighted the feature maps using S to improve spatial significance of the anomaly region:

$$X'' = X' \times S. \tag{8}$$

This spatial attention mechanism ensures that the weight of small target areas in the feature map is effectively enhanced while reducing the impact of background noise. Finally, SAM combines channel attention and spatial attention, greatly improves the model's perception of small-target anomaly areas, thereby improving the performance of anomaly detection.

3.4. Loss Function

The temporal feature pyramid Y consists of multiple levels, covering semantic information from low to high, allowing us to learn the features of short-term and long-term fragments at the same time and identify exception instances at different time scales. In our method, MIL is used to predict the most anomalies in each pyramid level separately. When $I = 1$, MSIL degenerates into the traditional MIL method.

When the time feature pyramid Y is constructed, we use the lightweight classification header f_{cls} to calculate the instance-level exception scores of each layer of the pyramid. Specifically, for the feature pyramid Y, the classification header f_{cls} evaluates each time step t of all I layers of the pyramid and predicts the probability of anomalies in that time step. Its expression is as follows:

$$P_i = \sigma\big(f_{\mathrm{cls}}(Z_i; \phi_{\mathrm{cls}})\big), \quad i = 1, \ldots, I, \tag{9}$$

 doi:10.20944/preprints202503.2351.v1

In our method, the classification header f_{cls} consists of three-layer one-dimensional convolution (kernel size is 3), layer normalization (applied to the first two layers) and ReLU activation function to ensure effective feature extraction and nonlinear transformation. Formally, f_{cls} is parameterized by the learnable parameter ϕ_{cls} and uses the sigmoid activation function σ to calculate the exception probability. The resulting $S_i \in \mathbb{R}^{T_i}$ represents the instance-level exception score independently predicted at each level of the pyramid. Where T_i represents the time step of the i layer. In order to train the model, we use the previously generated pseudo tag N as a direct supervision signal to constrain the instance-level prediction. It is worth noting that because different levels of the time pyramid may have different time resolutions, the pseudo tag N needs to be adjusted by the down sampling operator to align with the time matrix of each pyramid level to ensure consistent monitoring.

$$N_i = \begin{cases} N, & \text{when } i = 1, \\ \downarrow (N), & \text{when } i = 2, \ldots, I, \end{cases} \tag{10}$$

In our method, $N_i \in \mathbb{R}^{T_i}$ represents an instance-level pseudo-label, while $\downarrow(\cdot)$ represents a down sampling operator. Since different levels of the temporal feature pyramid may have different time scales, we need to down sample the pseudo-label N_i to ensure that it is consistent with the prediction results of the corresponding hierarchy in the time dimension, thereby achieving accurate supervised signal alignment. When all pyramid levels have instance-level pseudo-labels, we can formalize weakly supervised video anomaly detection (WSVAD) as a supervised learning problem.Specifically, we use adaptive focused loss (AF Loss) for training to effectively deal with the problem of uneven proportions of positive and negative samples.AF loss can adaptively adjust the weight of the difficult samples, making the model pay more attention to those anomalies with high uncertainty and great contributions, thereby improving the robustness of the detection.This optimization goal can be expressed as:

$$L_{\text{pse}} = \sum_{i=1}^{L} \text{Focal}(P_i, N_i). \tag{11}$$

Adaptive Focus Loss (AFL) is designed to mitigate the inherent imbalance in the data distribution, especially in anomaly detection tasks, where the number of anomaly samples (positive samples) is much smaller than that of normal samples (negative samples). This loss function adjusts the loss weight adaptively, emphasizes the learning of difficult-to-classify exception instances, while reducing the interference of easily classified samples to the optimization process. Such a design not only helps to improve the model's sensitivity to abnormal events, but also enhances its generalization ability in weakly supervised scenarios, thereby improving overall detection performance.

L_{pse} is not robust enough to effectively distinguish between normal and abnormal modes. During the training process, due to the possible volatility and inaccuracy of the generated pseudo-labels, the model may be misled by false high confidence pseudo-labels at subsequent stages, which will affect the learning effect. To alleviate this problem and improve the robustness of the model, we introduced video classification loss (L_{cls}) and ternary loss (L_{trip}) during the training process. Video classification loss aims to provide global supervision signals to ensure that the model can learn effective features that distinguish normal and abnormal samples; while ternary losses further enhance the model's ability to discriminate abnormal behavior by constraining the distance relationship between samples, thereby improving overall detection performance. This design effectively alleviates the impact of pseudo-label quality instability on model training, making the model more robust in video anomaly detection tasks for complex scenarios.

L_{cls} is a video-level classification loss, and its core goal is to aggregate instances of the same category, thereby improving the overall discrimination ability of the model to view video anomalies. To calculate the predicted score at the video level, we introduce a Top-k mean strategy. Specifically, let the video v_l consist of t_l fragments. We first sort the abnormal scores s_l of these fragments and select

 doi:10.20944/preprints202503.2351.v1

the highest $k = \left\lceil \frac{t_l}{\alpha} \right\rceil$ term for the mean calculation, so as to obtain the global abnormal score of the video. Among them, α is used as a hyperparameter to control the selected fragment ratio.

The core idea of this strategy is to adapt to the duration variations of different videos, rather than using fixed k values as in previous studies [39,40]. By dynamically adjusting the number of selected anomaly clips, we can more flexibly capture the performance of abnormal events in different videos, improving the model's adaptability to long and short videos. Finally, L_{cls} is calculated in the form of log loss , with the formula as follows:

$$L_{\text{cls}} = \log(s_l, y_l), \tag{12}$$

where $y_l \in \{0, 1\}$ is the video anomaly label, that is, whether the video contains an exception event. This design not only enhances the generalization ability of the model under different video lengths, but also avoids fixed information loss or misleading caused by k values, thereby improving the robustness and accuracy of abnormal detection.

The ternary loss (L_{trip}) is adopted in this study as defined in [41] and is further customized for the video anomaly detection task. This loss function is based on comparative representation learning, aiming to optimize the feature space so that the model can learn more discriminant feature distribution, thereby improving the ability to distinguish anomalies. In our task setting, we make specific adjustments to L_{trip} to effectively separate normal samples from exception samples in feature embedding space. Specifically, we constrain the embeddings learned by the model so that samples belonging to the same category (normal or abnormal) are gathered together, while samples of different categories are pushed away in the feature space, thereby improving the model's discriminatory ability. This design not only helps to reduce confusion among categories, but also enhances the model's generalization ability in complex scenarios, allowing it to detect abnormal behavior in videos more stably.

$$L_{\text{trip}} = \text{Triplet}(f_a, f_p, f_n), \tag{13}$$

where f_c (center feature) represents the average feature of the instance selected based on the model prediction score P_i. Similarly, f_r (correlation features) refers to the average feature of the instance selected based on the pseudo-label N_i , while f_d (distractor Features) represents the average feature of the instance in a normal video.Under this framework, the central feature f_c serves as a reference point, the relevant feature f_r should be as close to the central feature as possible, while the contrasting feature f_d should be away from the central feature, thereby forming a clearer category boundary in the feature space. This design can enhance the discriminant ability of the model, making it more robustly differentiate between normal and abnormal behavior in the abnormal detection task. In addition, by combining predicted scores and pseudo-label information, we further optimize feature representations, allowing the model to adaptively adjust the embedding space, thereby improving the generalization ability of complex video scenarios.

Given that the total loss function is essentially a weighted combination of multiple different subtasks, we draw on the method proposed in [42] to achieve an adaptive balance of various losses. By introducing a multi-task learning paradigm, we can dynamically adjust the weight of each loss item during the optimization process, so that the model can achieve collaborative learning between different tasks, thereby improving overall performance and generalization capabilities:

$$L_{\text{total}} = \frac{1}{\beta_1^2} L_{\text{cls}} + \frac{1}{\beta_2^2} L_{\text{pse}} + \frac{1}{\beta_3^2} L_{\text{trip}} + \sum_{i=1}^{3} \log(\beta_i + 1). \tag{14}$$

where β_i,$i \in \{1, 2, 3\}$, are learnable loss weights.

4. Experiment Results

To validate the proposed AFR method's superiority, it is compared with multiple state-of-the-art AFR approaches on two large-scale datasets, namely, XD-Violence and UCF-Crime.

 doi:10.20944/preprints202503.2351.v1

4.1. Datasets

XD-Violence XD-Violence is a large-scale, multimodal video anomaly detection dataset containing 4754 videos with a total duration of 217 hours, covering six categories of violence such as shootings, explosions, fights, demonstrations, burning, and car accidents, and provides frame level annotation. The data set combines both visual (RGB frame) and audio mode information, covering a variety of environments such as movies, network videos, surveillance videos, and sports events. The data set was divided into a training set (3954 videos) and a test set (800 videos), where the training set provided only video-level labels, while the test set provided frame-level labels, which was suitable for weakly supervised video anomaly detection tasks.

UCF-Crime UCF-Crime is a large data set for video anomaly detection, containing 1900 videos with a total duration of 128 hours, covering 13 crimes such as explosion, fighting, theft, arson, shooting, and road robbery, and providing frame level tags. The data is mainly from real surveillance video, the scene is close to practical application, and only contains visual (RGB) information. The data set is divided into a training set (950 videos) and a test set (950 videos). The training set has only video level labels, while the test set has frame level labels, which is suitable for weakly supervised video anomaly detection tasks.

4.2. Evaluation Metrics

For XD-Violence dataset, we mainly use frame-level average accuracy (AP) as a measure of coarse-grained detection performance, which can directly reflect the accuracy of the algorithm's recognition of violent behavior. For the UCFCrime dataset, we used the area under the frame-level receiver Operating Characteristic (ROC) curve (AUC) and the abnormal Video AUC (AnoAUC) to comprehensively evaluate the algorithm's ability to distinguish between normal and abnormal videos.A higher AP and AUC means the algorithm performs better and more robustly.

4.3. Implementation Details

The number of sampled snippets T for XD-Violence and UCF-Crime is set to 256 and 128.We used the AdamW optimizer with an initial learning rate of 5e-5 for XD-Violence and 1e-5 for UCF-Crime with a weight decay of 1e-2, respectively, and used TenCrop, a cropping method provided in Torchvision.Transforms to enhance data.We trained the XD-Violence dataset with a batch size of 30 and the UCF-Crime dataset with a batch size of 30 on the RTX 4090 GPU.

4.4. Comparison with State-of-the-art Methods

We compare with the state-of-the-art weakly supervised methods and report experimental results in Table 1.

Table 1. RESULTS ON THE XD-VIOLENCE AND UCF-CRIME DATASETS.

Method	Feature	AP(%) XD	AUC(%) UCF
P. Wu et al. (ECCV'20) [2]		-	82.44
RTFM (ICCV'21) [4]		77.81	84.03
CRFD (TIP'21) [20]	I3D-RGB	75.90	84.89
MSL (AAAI'22) [9]		78.28	-
MGFN (AAAI'23) [22]		79.19	86.98
AFR (Ours)	I3D-RGB	**81.74**	**92.73**

4.4.1. Comparisons on XD-Violence

We utilized the 5-crop I3D RGB features for our experiments and incorporated the Small-object Attention Module (SAM) to enhance fine-grained feature representation. By introducing SAM, our method achieved an average precision (AP) of **81.74%**, surpassing previous methods such as AnomalyCLIP and MGFN by a notable **2.55%** margin, thereby setting a new state-of-the-art performance

 doi:10.20944/preprints202503.2351.v1

benchmark on the XD-Violence dataset. The superior performance can be attributed to the synergistic effect of SAM and the spatio-temporal features extracted by the I3D backbone. Specifically, SAM enables the model to adaptively emphasize localized and small-scale anomalous patterns, which are often overlooked by conventional attention mechanisms. Moreover, SAM suppresses redundant background noise through channel-wise recalibration and spatial saliency refinement, allowing the model to capture more discriminative and task-relevant signals. Additionally, the effectiveness of SAM is particularly evident in scenarios involving high scene complexity, cluttered backgrounds, and occlusion, where traditional methods tend to misclassify subtle anomalies. Qualitative results show that our model consistently identifies small-scale anomalous behaviors, such as sudden object movements or transient violent actions, that previous approaches fail to capture. These improvements demonstrate the robustness and generalizability of our method for real-world anomaly detection tasks.

4.4.2. Comparisons on UCF-Crime

For evaluation on the UCF-Crime dataset, we adopted the 10-crop augmentation strategy on I3D RGB features, producing a 1024-dimensional representation for each input clip. The Small-object Attention Module (SAM) was then integrated into the feature extraction pipeline to enhance both spatial and channel-wise discriminability. With this enhancement, our model achieved a **remarkable AUC of 92.73%**, outperforming all previous state-of-the-art methods on this benchmark. SAM plays a critical role in refining feature responses that are indicative of abnormal behavior, especially in cases where the anomalies are subtle, small in size, or transient in nature. By emphasizing local saliency and adaptively calibrating the feature importance across spatial and channel dimensions, SAM enables the model to detect nuanced anomalies—such as short bursts of suspicious motion or minor deviant activities—that typically escape detection by global feature-based models.Furthermore, the UCF-Crime dataset contains a wide variety of real surveillance videos with high intra-class variance and low inter-class discriminability. In such challenging conditions, our SAM-enhanced model demonstrates strong robustness and generalization ability by effectively separating abnormal instances from normal background dynamics. This demonstrates that our method is not only effective in synthetic or curated datasets but also deployable in real-world surveillance scenarios where anomaly cues are often ambiguous and fleeting.

4.5. Ablation Studies and Analysis

Effectiveness of proposed modules. The core module composition of our method is Small-Object Attention Module (SAM). In this section, we deeply analyze the effect of this module and its internal structure on the improvement of model performance and verify its effectiveness through ablation experiments. Specifically, we compare the full model with a baseline model (see Table 2) that contains only the Small-Object Channel Attention Module (SCAM). Experimental results show that after the introduction of SAM, the AP of the XD-Violence dataset increased by 0.83%, while the AUC of the UCF-Crime dataset increased by 5.16%, significantly verifying the effectiveness of this module in the small-objective anomaly detection task. It is worth noting that when SAM is integrated into the overall architecture as an attention mechanism, the performance of both data sets reaches a new level. This result further proves the complementarity between SAM and SCAM, that is, the synergy between the two can capture small-objective anomaly characteristics more accurately, thereby improving the overall anomaly detection capability of the model.

Table 2. Effectiveness of proposed modules.sam:small-object attention module; ✓Signifies "Included".

Baseline	SAM	AP(%) - XD	AUC(%) - UCF
✓		80.91	87.57
✓	✓	**81.74**	**92.73**

 doi:10.20944/preprints202503.2351.v1

Complexity of the proposed module. To further explore the impact of SAM attention mechanisms on the complete model, we report the performance of its components in detail in Table 3. The baseline model relies solely on SCAM attention for feature supervision, while we compared the contributions of different attention mechanisms through ablation experiments. Experimental results show that both spatial attention (SAM-CA) and channel attention (SAM-SA) in SAM attention have positive effects on model performance, further verifying its effectiveness in capturing small-target anomaly behavior. Specifically, spatial attention enhances the model's perception of local anomalies, while channel attention helps the model focus on important dimensions of anomalies. Overall, these two refined attention mechanisms together constitute SAM and contribute to optimal performance in the experiment. This result shows that spatial attention and channel attention are complementary and indispensable in the small-objection anomaly detection task, providing strong support for improving the overall detection capability.

Table 3. Complexity of the proposed module.sam:small-object attention module;sam-ca:small-object channel attention module;sam-sa:small-object spatial attention module; ✓signifies "included".

Baseline	SAM-CA	SAM-SA	AP(%) - XD	AUC(%) - UCF
✓			80.91	87.57
✓	✓		79.02	84.11
✓		✓	80.88	85.64
✓	✓	✓	**81.74**	**92.73**

Performance Analysis of the Proposed Module. To verify the effectiveness of SAM attention mechanism in the video anomaly detection task, we conducted an in-depth comparison and analysis of NonLocal attention and summarized the experimental results in Table 4. Judging from the experimental results, although NonLocal attention can capture global features, its ability to detect small target anomaly is still limited. SAM attention further strengthens the model's focus ability on small target areas by fusing spatial attention and channel attention, thereby significantly improving the abnormal detection performance. Specifically, SAM's metrics on both datasets go beyond the NonLocal mechanism, fully verifying its effectiveness and applicability in small-objective anomaly detection tasks. In addition, this experiment further demonstrates that the SAM mechanism can more accurately model small-target anomaly behavior, thus providing a more targeted attention modeling strategy for small-target video anomaly detection.

Table 4. Performance analysis of the proposed module.sam:small-object attention module; nonlocal: non-local neural module ✓signifies "included".

Baseline	SAM	NonLocal	AP(%) - XD	AUC(%) - UCF
✓			80.91	87.57
✓		✓	78.19	84.47
✓	✓		**81.74**	**92.73**

4.6. Visualization

To further analyze the effect of SAM attention mechanism on feature representation, we visualized the feature mapping changes of the model before and after the introduction of SAM in Figure 3. From the results, it can be observed that when SAM is not applied, the model's feature response is relatively discrete and it is not effective to focus on the abnormal target area. After the introduction of SAM, the feature activation distribution is more concentrated, indicating that the model's perception of small target anomalies has been significantly enhanced. This phenomenon verifies that SAM adaptively adjusts attention distribution, making the characteristics of the abnormal area more prominent, thereby improving the robustness of detection. In addition, from the perspective of feature distribution, SAM makes the features of abnormal targets more separable, improves the category discrimination ability,

 doi:10.20944/preprints202503.2351.v1

and further proves its effectiveness in small-objective anomaly detection task. Overall, the SAM mechanism enhances the model's recognition accuracy of abnormal behavior by strengthening the characterization ability of key regions, providing an efficient and robust attention modeling solution for small-objective anomaly detection.

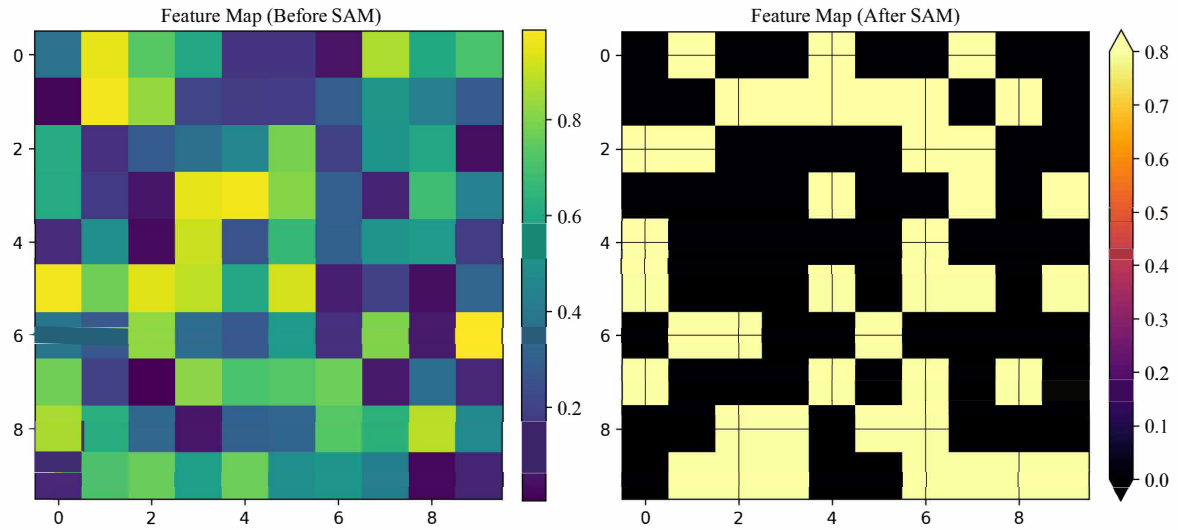

Figure 3. Feature mapping visualization before and after SAM attention mechanism

5. Conclusions

This study proposed a method that fused the Small-Object Attention Mechanism (SAM) and Contrastive Language-Image Pretraining (CLIP) to enhance small-object anomaly detection capabilities. By integrating SAM into the FPN structure of CMSIL, the model adaptively adjusted the importance of channel and spatial positions while leveraging CLIP for semantic understanding, enabling cross-scale and cross-scene anomaly detection. Experimental results demonstrated that the proposed method achieved state-of-the-art performance on the UCF-Crime and XD-Violence datasets. However, this approach had certain limitations, including high computational complexity and the need for further optimization in CLIP and visual feature fusion strategies. In future work, unsupervised learning, cross-modal anomaly detection, and more efficient feature fusion strategies will be explored to further improve the model's generalization ability and real-time performance. **Author Contributions:** Presenting the programme concept as well as the methodology of this paper and wrote the original manuscript, D.A.; Conducting a literature survey related to this paper as well as supervising the writing of the article, W.L.; Review of this manuscript, J.L.; Revision of this manuscript, S.Z. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The authors confirm that the data supporting the findings of this study are available within the article.

Conflicts of Interest: No potential conflict of interest was reported by the authors.

References

1. Ahn, S.; Jo, Y.; Lee, K.; Kwon, S.; Hong, I.; Park, S. AnyAnomaly: Zero-Shot Customizable Video Anomaly Detection with LVLM. *arXiv preprint arXiv:2503.04504* **2025**.
2. Ma, J.; Wang, J.; Luo, J.; Yu, P.; Zhou, G. Sherlock: Towards Multi-scene Video Abnormal Event Extraction and Localization via a Global-local Spatial-sensitive LLM. *arXiv preprint arXiv:2502.18863* **2025**.
3. Li, Z.; Zhao, M.; Yang, X.; Liu, Y.; Sheng, J.; Zeng, X.; Wang, T.; Wu, K.; Jiang, Y.G. STNMamba: Mamba-based Spatial-Temporal Normality Learning for Video Anomaly Detection. *arXiv preprint arXiv:2412.20084* **2024**.
4. Xu, A.; Wang, H.; Ding, P.; Gui, J. Dual Conditioned Motion Diffusion for Pose-Based Video Anomaly Detection. *arXiv preprint arXiv:2412.17210* **2024**.

 doi:10.20944/preprints202503.2351.v1

5. Zhang, H.; Xu, X.; Wang, X.; Zuo, J.; Huang, X.; Gao, C.; Zhang, S.; Yu, L.; Sang, N. Holmes-vau: Towards long-term video anomaly understanding at any granularity. *arXiv preprint arXiv:2412.06171* **2024**.
6. Tan, X.; Wang, H.; Geng, X. Frequency-Guided Diffusion Model with Perturbation Training for Skeleton-Based Video Anomaly Detection. *arXiv preprint arXiv:2412.03044* **2024**.
7. Ye, M.; Liu, W.; He, P. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. *arXiv preprint arXiv:2412.01095* **2024**.
8. Bao, Q.; Liu, F.; Liu, Y.; Jiao, L.; Liu, X.; Li, L. Hierarchical scene normality-binding modeling for anomaly detection in surveillance videos. In Proceedings of the Proceedings of the 30th ACM international conference on multimedia, 2022, pp. 6103–6112.
9. Kaneko, Y.; Miah, A.S.M.; Hassan, N.; Lee, H.S.; Jang, S.W.; Shin, J. Multimodal Attention-Enhanced Feature Fusion-based Weekly Supervised Anomaly Violence Detection. *arXiv preprint arXiv:2409.11223* **2024**.
10. Almahadin, G.; Subburaj, M.; Hiari, M.; Sathasivam Singaram, S.; Kolla, B.P.; Dadheech, P.; Vibhute, A.D.; Sengan, S. Enhancing video anomaly detection using spatio-temporal autoencoders and convolutional lstm networks. *SN Computer Science* **2024**, *5*, 190.
11. Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A.K.; Davis, L.S. Learning temporal regularity in video sequences. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 733–742.
12. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Ieee, 2005, Vol. 1, pp. 886–893.
13. Dalal, N.; Triggs, B.; Schmid, C. Human detection using oriented histograms of flow and appearance. In Proceedings of the Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part II 9. Springer, 2006, pp. 428–441.
14. Nejad, S.S. Weakly-Supervised Anomaly Detection in Surveillance Videos Based on Two-Stream I3D Convolution Network. Master's thesis, The University of Western Ontario (Canada), 2023.
15. Liu, Z.; Nie, Y.; Long, C.; Zhang, Q.; Li, G. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 13588–13597.
16. Li, D.; Nie, X.; Gong, R.; Lin, X.; Yu, H. Multi-branch GAN-based abnormal events detection via context learning in surveillance videos. *IEEE Transactions on Circuits and Systems for Video Technology* **2023**, *34*, 3439–3450.
17. Sun, S.; Gong, X. Hierarchical semantic contrast for scene-aware video anomaly detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 22846–22856.
18. Li, S.; Liu, F.; Jiao, L. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 1395–1403.
19. Ni, B.; Peng, H.; Chen, M.; Zhang, S.; Meng, G.; Fu, J.; Xiang, S.; Ling, H. Expanding language-image pretrained models for general video recognition. In Proceedings of the European conference on computer vision. Springer, 2022, pp. 1–18.
20. Ju, C.; Han, T.; Zheng, K.; Zhang, Y.; Xie, W. Prompting visual-language models for efficient video understanding. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 105–124.
21. Qian, Z.; Tan, J.; Ou, Z.; Wang, H. CLIP-Driven Multi-Scale Instance Learning for Weakly Supervised Video Anomaly Detection. In Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2024, pp. 1–6.
22. Goyal, K.; Singhai, J. Review of background subtraction methods using Gaussian mixture model for video surveillance systems. *Artificial Intelligence Review* **2018**, *50*, 241–259.
23. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.
24. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.c. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems* **2015**, *28*.

 doi:10.20944/preprints202503.2351.v1

25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
26. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. PmLR, 2021, pp. 8748–8763.
27. Shen, F.; Shu, X.; Du, X.; Tang, J. Pedestrian-specific Bipartite-aware Similarity Learning for Text-based Person Retrieval. In Proceedings of the Proceedings of the 31th ACM International Conference on Multimedia, 2023.
28. Shen, F.; Tang, J. IMAGPose: A Unified Conditional Framework for Pose-Guided Person Generation. In Proceedings of the The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.
29. Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; Tang, J. IMAGDressing-v1: Customizable Virtual Dressing. *arXiv preprint arXiv:2407.12705* **2024**.
30. Shen, F.; Ye, H.; Liu, S.; Zhang, J.; Wang, C.; Han, X.; Yang, W. Boosting Consistency in Story Visualization with Rich-Contextual Conditional Diffusion Models. *arXiv preprint arXiv:2407.02482* **2024**.
31. Shen, F.; Wang, C.; Gao, J.; Guo, Q.; Dang, J.; Tang, J.; Chua, T.S. Long-Term TalkingFace Generation via Motion-Prior Conditional Diffusion Model. *arXiv preprint arXiv:2502.09533* **2025**.
32. Brox, T.; Bruhn, A.; Papenberg, N.; Weickert, J. High accuracy optical flow estimation based on a theory for warping. In Proceedings of the Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part IV 8. Springer, 2004, pp. 25–36.
33. Liu, W.; Luo, W.; Lian, D.; Gao, S. Future frame prediction for anomaly detection–a new baseline. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6536–6545.
34. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.
35. Zhang, M.; Wang, J.; Qi, Q.; Sun, H.; Zhuang, Z.; Ren, P.; Ma, R.; Liao, J. Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 17385–17394.
36. Li, H.; Zhang, R.; Pan, Y.; Ren, J.; Shen, F. LR-FPN: Enhancing Remote Sensing Object Detection with Location Refined Feature Pyramid Network. *arXiv preprint arXiv:2404.01614* **2024**.
37. Weng, W.; Wei, M.; Ren, J.; Shen, F. Enhancing Aerial Object Detection with Selective Frequency Interaction Network. *IEEE Transactions on Artificial Intelligence* **2024**, *1*, 1–12.
38. Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; Yang, W. Advancing pose-guided image synthesis with progressive conditional diffusion models. *arXiv preprint arXiv:2310.06313* **2023**.
39. Sultani, W.; Chen, C.; Shah, M. Real-world anomaly detection in surveillance videos. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6479–6488.
40. Zhou, H.; Yu, J.; Yang, W. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2023, Vol. 37, pp. 3769–3777.
41. Balntas, V.; Riba, E.; Ponsa, D.; Mikolajczyk, K. Learning local feature descriptors with triplets and shallow convolutional neural networks. In Proceedings of the Bmvc, 2016, Vol. 1, p. 3.
42. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7482–7491.