Unified Video Anomaly Detection Model for Detecting Different Anomaly Types

Kijung Lee, Youngwan Jo, Sunghyun Ahn, Sanghyun Park
Yonsei University, Seoul, Republic of Korea
{rlwjd4177, jyy1551, skd, sanghyun}@yonsei.ac.kr

Abstract

Video anomaly detection (VAD) is a crucial task for public safety and workforce reduction. Due to the rarity of abnormal events and the high cost of data collection, one-class classification (OCC) methods are extensively used. OCC methods are divided into object- and frame-centric approaches, each with its limitations. Object-centric methods fail to detect non-object anomalies because they focus solely on objects, whereas frame-centric methods struggle to identify abnormalities due to a higher background rate than the foreground rate in video frames. To this end, we define three types of abnormal events, namely, human, appearance, and non-object anomalies, and propose a unified VAD (UniVAD) model that effectively detects each defined anomaly type. UniVAD comprises three streams, namely, skeleton, local-visual, and global-visual, and each stream focuses on a specific type of anomaly. In addition, each stream uses an autoencoder; thus, we introduce the feature future past prediction task, which predicts past and future features based on present feature to suppress the strong generalization capacity of autoencoders. We validate the proposed model on three public benchmarks, ShanghaiTech, UBnormal, and NWPUCampus, and demonstrate that it achieves state-of-the-art performance by a significant margin.

1. Introduction

Video anomaly detection (VAD) aims to automatically identify events that deviate from normal patterns in videos and is essential for public safety and workforce reduction. Although a considerable amount of research has been conducted on VAD, it remains a challenging task. Due to the subjective definition of anomalies and the scarcity of anomalous samples, a fully supervised manner is not employed. Therefore, most VAD methods follow a one-class classification (OCC) approach, which utilizes only normal data during training. OCC methods model the distribution of normal data by predicting future frames or reconstructing present frames and consider samples that deviate from this distribution as anomalies.

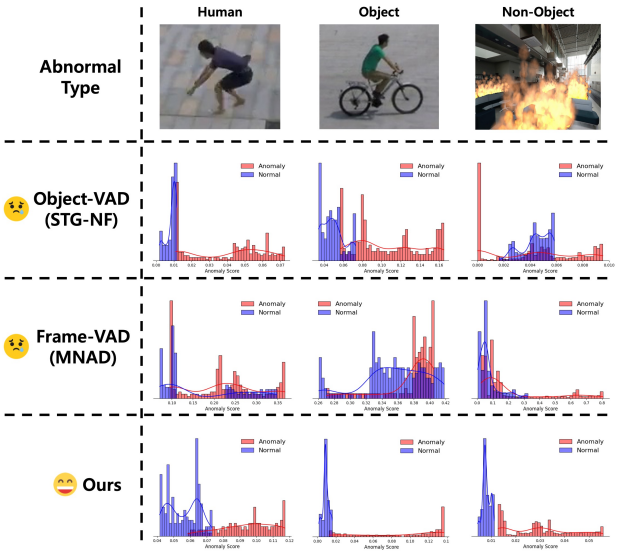

Figure 1. Examples of human, appearance, and non-object anomalies and comparisons of results from different methods. The results are presented as histograms of the anomaly scores for the normal and abnormal samples. From top to bottom: anomaly type, result of STG-NF [20] (object-centric method), result of MNAD [32] (frame-centric method), and result of ours.

OCC methods can be divided into object- and frame-centric approaches. Object-centric methods [3, 15, 20, 22, 23, 26, 27, 46] focus on objects, such as humans, cars, and bicycles, utilizing object detectors [47] or multi-object trackers [48]. Object-centric methods outperform frame-centric methods due to the higher occurrence rate of object anomalies. However, they may fail to detect anomalies if object detectors fail to detect objects or when non-object anomalies, such as fire and smoke, occur. As shown in the Object-VAD results presented in Fig. 1, Object-VAD effectively distinguishes between normal and abnormal samples when the abnormal type is human or appearance. However, it fails to make this distinction when the abnormal type is non-object. In contrast, frame-centric methods [5, 17, 25, 30, 32, 39, 41] consider the overall context

without relying on object detectors; they can detect non-object anomalies. However, because these methods utilize video frames in which the background rate is higher than the foreground rate, their performance is generally lower than that of object-centric methods. As shown in the Frame-VAD results presented in Fig. 1, Frame-VAD effectively distinguishes the distribution of normal and abnormal samples for non-object anomalies compared to Object-VAD. However, it fails to make a clear distinction between anomalies related to humans and appearance.

To address these issues, motivated by Sun et al. [38], we categorize all anomalies into the following types: human anomalies, such as violence or theft; appearance anomalies, such as vehicles or bicycles passing on sidewalks; non-object anomalies, such as fire or smoke. To effectively detect each defined anomaly type, we propose a unified VAD (UniVAD) model. The proposed UniVAD model comprises skeleton, local-visual, and global-visual streams, with each stream focusing on detecting a specific anomaly type. The skeleton stream uses skeleton data extracted from objects to detect human anomalies; the local-visual stream uses local-visual features extracted from objects to detect appearance anomalies; the global-visual stream uses global-visual features extracted from video frames to detect non-object anomalies. Following previous studies [19, 49], we utilize autoencoders (AEs) in all streams to model the distribution of normal data.

Because we use reconstruction-based AEs, it is essential to suppress their strong generalization capacity, which effectively reconstructs unseen anomalous samples during training. To this end, future prediction methods [25, 40, 44], which predict future frame based on past frames, and bidirectional prediction methods [7, 13, 24, 43], which predict present frame based on both past and future frames, have been developed. However, future prediction methods can accurately predict anomalous samples due to low variations between adjacent frames, and bidirectional prediction methods can interpolate present frame from both past and future frames, enabling effective reconstruction of anomalous samples. To effectively address this issue, we propose the feature future past prediction (FFPP) task. This task predicts both past and future features based on present feature, effectively suppressing the strong generalization capacity of AEs. In the FFPP task, variations between the input and output are significant, and present feature cannot be interpolated from past and future features.

All streams are designed to effectively detect anomalies by utilizing the FFPP task. The skeleton stream predicts both future and past skeletons based on present skeleton and introduces skeleton spatial loss (SSL) and skeleton temporal loss (STL) to learn the spatial and temporal relationships between predicted skeletons. The local- and global-visual streams predict past, present, and future visual features, which are extracted from past, present, and future frames or objects, based on the present visual feature.

UniVAD exhibits improvements of 2.4%, 10.5%, and 3.5% in the micro area under the receiver operating characteristic curve (AUC) compared to previous state-of-the-art (SOTA) methods on the ShanghaiTech, UBnormal, and NWPUCampus datasets, respectively, demonstrating that it is concise yet effective. Furthermore, extensive ablation studies and visual results demonstrate the well-founded design of UniVAD, highlighting that each stream effectively captures distinct types of anomalies. Our primary contributions can be summarized as follows:

- We propose UniVAD, which comprises three independent streams, and is designed to effectively detect all anomaly types.
- We propose the FFPP task, which predicts both past and future features based on present feature to effectively reduce the generalization capacity of AEs.
- We achieved new SOTA performance on three public benchmarks and provided extensive ablation experiments.

2. Related work

Video Anomaly Detection. VAD has multiple training modes, including the OCC mode [17, 25, 27, 30, 32, 43, 45], which uses only normal data for training; the unsupervised mode [45], where both normal and abnormal data are included in the training set but without labels; the weakly supervised mode [9, 37, 45], where only video-level labels are available; the fully supervised mode [20], where frame-level labels are present. In this study, the OCC mode was adopted.

Existing VAD methods that utilize only normal training data are primarily grouped into reconstruction-based methods [6, 21, 28, 34] that learn to reconstruct present frames and prediction-based methods [25, 40, 44] that learn to predict future frame based on past frames, typically employing generative adversarial networks (GANs) [18] or AEs. Both methods classify samples with significant reconstruction or prediction errors as anomalies. However, they struggle with strong generalization capacity for anomalies (even when these samples are unseen during training), which results in poor anomaly detection performance. The use of a memory mechanism [17, 32], multimodal data (e.g., optical flow [25], skeleton [20] and text [9] data), diffusion mechanism [41] and the introduction of bidirectional prediction [23, 43] suppresses the generalization capacity to some extent; however, the improvements remain far from perfect. To address this, we propose the FFPP task, which predicts both past and future features based on present feature.

Object-centric VAD. Object-centric VAD uses objects extracted from frames via an object detector to perform object-level anomaly detection using self-supervised methods or

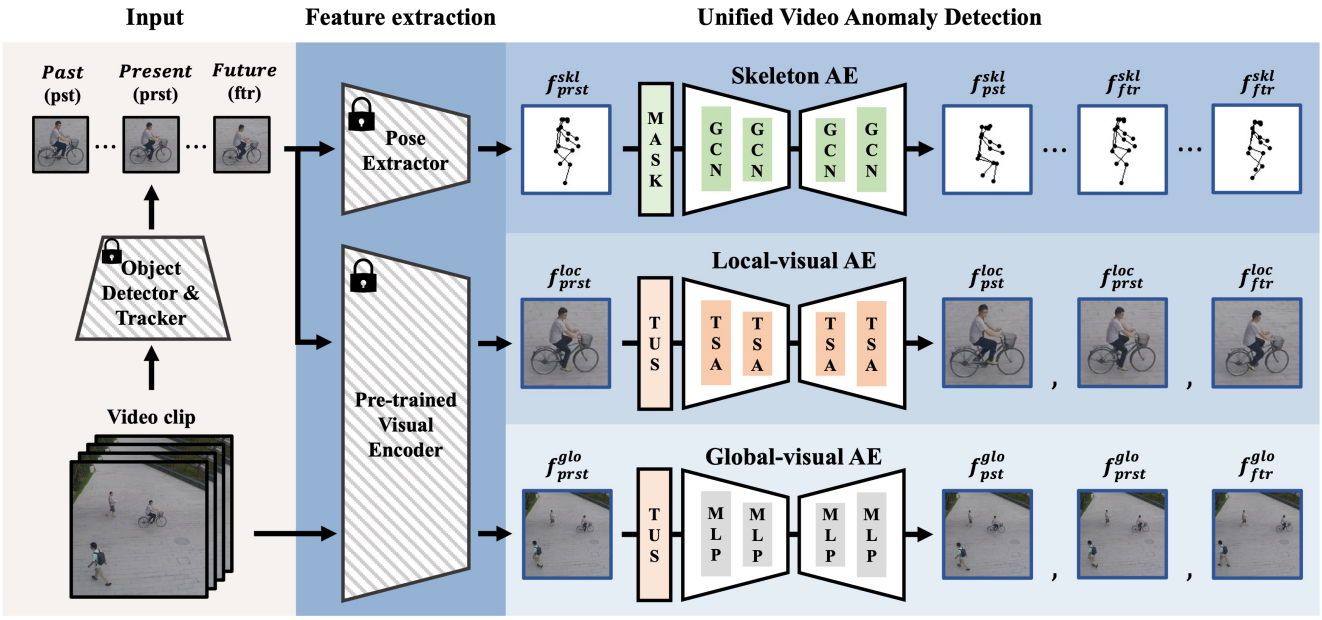

Figure 2. Overview of proposed method. The overall framework consists of preprocessing, feature extraction, and three streams. Each stream, using a preprocessing and feature extraction step, predicts all features from past to future based on present feature. The skeleton and local-visual streams use the object, whereas the global-visual stream uses the entire video clip.

various pretrained models. For example, Sun et al. [38] extracted appearance and motion information from objects using ViT [2] and PoseConv3D [11] and applied contrastive learning [8] to push different objects apart and bring similar objects closer together. Morais et al. [31] extracted skeleton data from objects using AlphaPose [12] to detect human anomalies and employed two GRU AE branches to account for the global and local decomposition of the skeleton. Wang et al. [39] divided objects into spatiotemporal cubes and reconstructed these mixed cubes as if solving a jigsaw puzzle, thereby enhancing the ability of the model to capture the spatial and temporal relationships of objects. However, these methods only detect object anomalies, such as humans, cars, and bicycles, and they struggle to detect non-object anomalies such as fire and smoke. In contrast, our approach can detect non-object anomalies using the global-visual stream.

Frame-centric VAD. Frame-centric VAD takes video frames as input and utilizes AE or GAN structures to learn prediction and reconstruction at the pixel level of images. These methods can detect all types of anomalies and have the advantage of high fps speeds, because they tend to rely less on pretrained models. For example, Liu et al. [25] used GAN structures to predict future frame based on short past frames and employed Flownet [10] to learn the optical flow between past and future frames, thereby enhancing the quality of the generated frame. Ristea et al. [34] proposed a self-supervised predictive convolutional attentive block to enhance VAD performance. However, these methods use image frames with a higher proportion of background than foreground, which increases the false positive rate for normal samples and reduces the detection performance for small objects. In contrast, our approach mitigates these issues using the skeleton and local-visual streams, which results in improved detection performance for small objects.

3. Method

Fig. 2 shows the overall architecture of the proposed method and the preprocessing step. In the preprocessing stage, a video is divided into multiple clips. The l-th video clip, represented as $v^l = \{v_t^l \mid v_t^l \in \mathbb{R}^{H\times W\times 3}\}_{t=1}^{T}$, consists of T frames and is provided as input. Within this clip, an object detector and tracker are used to identify and extract multiple continuous objects across the frames. For simplicity, we assume that only a single continuous object sequence $o^l = \{o_t^l \mid o_t^l \in \mathbb{R}^{H'\times W'\times 3}\}_{t=1}^{T}$ is extracted from v^l. In this context, we define the terms past (pst), present ($prst$), and future (ftr) correspond to specific time points $t = 1, t = T/2$,and $t = T$, respectively.

UniVAD comprises the skeleton, local-visual, and global-visual streams. The skeleton and local-visual streams are designed to use the extracted object sequence o^l to detect object-related (human, appearance) anomalies, whereas the global-visual stream leverages the video frame sequence v_l to detect non-object anomalies. All streams apply the FFPP task, which models past and future fea-

tures based on present feature to suppress the strong generalization capacity of AEs. However, each stream exhibits slight differences in terms of its learning approach. The skeleton stream predicts all skeletons from past to future based on present skeletons, whereas the local- and global-visual streams predict past, present, and future visual features based on present visual feature. Since frames are affected by nuisance parameters such as background clutter, illumination changes, and other environmental factors [20, 29, 31, 36], visual features tend to exhibit lower temporal consistency compared to skeleton features. As a result, the visual streams are designed to predict features across all three time frames (past, present, and future). In the inference step, the final anomaly score is obtained by integrating the anomaly scores of each stream. The details of UniVAD are explained in the following section.

3.1. Skeleton stream

The skeleton stream leverages skeleton data to detect human-related anomalies and captures movement patterns and poses to identify unusual behavior. The continuous skeleton sequences $f^{skl} \in \mathbb{R}^{T\times J\times C}$ are extracted from o^l using a pose extractor. Here, T, J, and C represent the length of sequence, the number of joints, and the coordinates of the joints, respectively. To ensure the matching of the input and output shapes within this stream, the present skeleton $f^{skl}_{mask} \in \mathbb{R}^{T\times J\times C}$ is generated by masking the past and future skeletons, focusing only on the present state. This can be represented by the following equation:

$$f^{skl}_{mask} = \begin{cases} f^{skl}_t & \text{if } t = T/2 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Following previous studies [14, 20, 23], we use a graph-based AE Φ_{skl}, which utilizes spatial temporal graph convolution networks (GCN), to model the spatial and temporal relationships of the skeletons. Φ_{skl} predict all skeletons from past to future based on f^{skl}_{prst}. This follows the equation below.

$$\hat{f}^{skl} = \Phi_{skl}(f^{skl}_{mask}) \tag{2}$$

We introduce the SSL and STL to learn the spatial and temporal relationships of the predicted skeletons. SSL encourages the predicted joint coordinates to be similar to their target coordinates by penalizing coordinate differences. Specifically, we minimize the mean squared error (MSE) between the predicted skeletons $\hat{f}^{skl}$ and the target skeletons f^{skl}, calculated as follows:

$$L_{ssl} = \sum_{t}^{T}\sum_{j}^{J} \left\| f^{skl}_{t,j} - \hat{f}^{skl}_{t,j} \right\|_2^2 \tag{3}$$

STL encourages the displacement of the predicted joint coordinates to more closely match the displacement of the target joint coordinates by penalizing coordinate differences. To achieve this, we minimized the MSE between the differences in the predicted and target skeletons, as calculated below.

$$L_{stl} = \sum_{j}^{J} \left\| (f^{skl}_{T,j} - f^{skl}_{1,j}) - (\hat{f}^{skl}_{T,j} - \hat{f}^{skl}_{1,j}) \right\|_2^2 \tag{4}$$

Finally, the overall loss function is expressed as follows:

$$L_{skl} = \lambda * L_{ssl} + (1-\lambda) * L_{stl} \tag{5}$$

Here, λ denotes a hyperparameter that controls the contribution of each loss.

3.2. Local-visual stream

The local-visual stream leverages local-visual features to effectively detect appearance anomalies. The past, present, and future local-visual features is extracted from the past, present, and future object through the encoder Φ_{enc}. To model long-term temporal dependencies, these features are stacked in consecutive clips L as follows:

$$f^{loc}_{l,*} = \Phi_{enc}(o^l_*), \tag{6}$$

$$f^{loc}_{*} = \{f^{loc}_{l,*} | f^{loc}_{l,*} \in \mathbb{R}^D\}^{L}_{l=1}, \tag{7}$$

where '*' denotes either pst, $prst$, or ftr, D is dimension of the feature. We employ temporal upsampling (TUS) to generate e^{loc}_{pst}, e^{loc}_{prst} and $e^{loc}_{ftr} \in \mathbb{R}^{L\times D}$, which are then concatenated into a single feature $e^{loc} \in \mathbb{R}^{3L\times D}$. This concatenated feature is processed by a transformer-based AE Φ_{loc}, which utilizes temporal self-attention layers (TSA) to model the correlations between consecutive objects clips as follows:

$$e^{loc}_{pst}, e^{loc}_{prst}, e^{loc}_{ftr} = TUS(f^{loc}_{prst}) \tag{8}$$

$$\hat{f}^{loc}_{pst}, \hat{f}^{loc}_{prst}, \hat{f}^{loc}_{ftr} = \Phi_{loc}(e^{loc}) \tag{9}$$

The network is trained by minimizing the MSE between the predicted and target local visual features and ensures that the model accurately captures the desired feature correlations as follows:

$$L_{loc} = \left\| f^{loc}_{pst} - \hat{f}^{loc}_{pst} \right\|_2^2 + \left\| f^{loc}_{prst} - \hat{f}^{loc}_{prst} \right\|_2^2 + \left\| f^{loc}_{ftr} - \hat{f}^{loc}_{ftr} \right\|_2^2 \tag{10}$$

3.3. Global-visual stream

The global-visual stream uses global-visual features to effectively detect non-object anomalies within the video frames. In a similar manner to the local-visual stream, The past, present, and future local-visual features is extracted

from the past, present, and future frame through the encoder Φ_{enc} as follows:

$$f_*^{glo} = \Phi_{enc}(v_*^l), \tag{11}$$

We use temporal upsampling (TUS) and the multilayer-based AE Φ_{glo} to capture global features, and Φ_{glo} predicts the three global-visual feature based on f_{prst}^{glo} as follows:

$$e_{pst}^{glo}, e_{prst}^{glo}, e_{ftr}^{glo} = TUS(f_{prst}^{glo}) \tag{12}$$

$$\hat{f}_{pst}^{glo}, \hat{f}_{prst}^{glo}, \hat{f}_{ftr}^{glo} = \Phi_{glo}(e^{glo}) \tag{13}$$

The network is trained by minimizing the MSE between the predicted and target global features as follows:

$$L_{glo} = \left\| f_{pst}^{glo} - \hat{f}_{pst}^{glo} \right\|_2^2 + \left\| f_{prst}^{glo} - \hat{f}_{prst}^{glo} \right\|_2^2 + \left\| f_{ftr}^{glo} - \hat{f}_{ftr}^{glo} \right\|_2^2 \tag{14}$$

3.4. Inference

The anomaly score for each stream is calculated based on the loss of that stream. Because the skeleton and local-visual streams use objects, they yield an object-level anomaly score, whereas the global-visual stream, which uses video frames, yields a frame-level anomaly score. To obtain the final anomaly score, the highest object-level anomaly score among multiple objects is assigned to the corresponding frame. The total anomaly score of each frame is then calculated as follows:

$$Score_{tot} = \alpha * Score_{skl} + \beta * Score_{loc} + \gamma * Score_{glo} \tag{15}$$

Here, α, β, and $\gamma \in [0, 1]$ denote hyperparameters that set the importance of each stream, and they may vary slightly depending on the proportion of anomaly types in the dataset. Following [39], we apply a temporal one-dimensional Gaussian filter to smooth the values.

4. Experiments

4.1. Experimental setup

We evaluated the proposed model on three public benchmarks: ShanghaiTech (ShT) [25], UBnormal (UB) [1], and NWPUCampus (NWPU) [5]. Based on the OCC setting, only normal data were used for training, whereas both normal and abnormal data were used for testing.

ShT. This dataset consists of 330 training videos and 107 test videos, containing 13 different scenes. Anomalous events include appearance-related anomalies, such as bicycles and cars, and human-related anomalies, such as throwing bags and running.

UB. This dataset comprises 543 synthetic videos across 29 different scenes. It contains human-related anomalies, such as hitting, running, and jumping, and non-object anomalies, such as fire and smoke.

NWPU. This dataset is the largest and comprises 305 training videos and 242 test videos across 43 different scenes. In addition to human- and appearance-related anomalies, it includes scene-dependent anomalies, such as walking on the sidewalk (normal) and walking on the road (abnormal).

Metrics. We used the micro- and macro-averaged AUC scores as metrics, following previous studies [30, 35, 39, 46]. The micro- and macro-AUC metrics are among the most commonly used evaluation metrics in the VAD task, and higher values indicate better discrimination between normal and abnormal events. Micro-AUC calculates the AUC score across the entire video set, whereas macro-AUC calculates the AUC score for each video individually and averages these values across all videos.

4.2. Implementation details

We used Yolov5 [47], Bytetrack [48], and AlphaPose [12] as the object detector, multi-object tracker, and pose extractor, respectively. CLIP(ViT-B/32) [33] was employed as the visual encoder, and the output of the last layer of CLIP was used as the visual feature, where D_{vis} was 512. During the preprocessing process, the scale of all frames or objects was set to 224×224. For the local- and global-visual streams, T was set to 8, and for the skeleton stream, T was set to 24 for the ShT and NWPU, and 16 for the UB. In the loss of the skeleton stream, λ was set to 0.5. To train the model, we used AdamW with a learning rate of 3e-4 and a batch size of 16 . The number of epochs was set to 200 across all datasets. The α, β, and γ used for the total anomaly score were set differently for each dataset, because the types of anomalies included varied across datasets. For ShT, UB, and NWPU, we used (1.0, 0.1, 0.01), (1.0, 0.01, 0.01), and (1.0, 1.0, 1.0), respectively.

4.3. Comparison with SOTA methods

We present the performance of UniVAD on three public benchmarks and compare it with previous SOTA methods in terms of micro- and macro-AUC in Tab. 1. Object-centric methods tend to outperform frame-centric methods. Among various object-centric approaches, some [1, 5, 15, 16, 27, 34, 35, 46] enhance performance by using a multitask model (masked with $\triangledown$ in Tab. 1) or scene-conditioned model(masked with $\triangle$ in Tab. 1) as the backbone or by training with a virtual dataset (masked with $*$ in Tab. 1). Since the ratio of anomaly types varies slightly across datasets, we implemented two versions of UniVAD: one with tuned values of α, β, and γ for each dataset, and the other using fixed values of 1.0, 0.1, and 0.1, respectively. We achieved higher performance than both frame-centric and object-centric methods without using these techniques.

Results on ShT. The proposed method achieved micro-

type	Method	Venue	ShT		UB		NWPU	
			Micro	Macro	Micro	Macro	Micro	Macro
frame-centric	Liu et al[25]	CVPR18	72.8	80.6	-	-	57.9	60.2
	Sultani et al[37]	CVPR18	-	76.5	50.3	76.8	-	-
	Gong et al[17]	CVPR19	71.2	-	-	-	61.9	62.5
	Park et al[32]	CVPR20	68.3	79.7	56.6	-	62.5	63.6
	Bertasius et al[4]	ICML21	-	-	68.5	80.3	-	-
	Lv et al[28]	CVPR21	73.8	-	-	-	64.4	-
	Wang et al[40]	TNNLS22	76.6	-	-	-	61.9	64.2
	Yan et al[41]	ICCV23	78.6	-	62.7	-	-	-
	Ristea et al[35]*	CVPR24	79.1	84.7	58.5	81.4	-	-
	Micorek et al[30]	CVPR24	81.3	85.9	72.8	85.5	-	-
	Yang et al[42]	ECCV24	85.2	-	71.9	-	-	-
object-centric	Ionescu et al[22]	CVPR19	78.7	84.9	-	-	59.3	63.4
	Liu et al[26]	ICCV21	74.2	83.2	-	-	63.7	-
	Georgescu et al[15]*	CVPR21	82.4	89.3	55.4	84.5	-	-
	Georgescu et al[16]*	TPAMI21	82.7	89.3	61.3	85.6	-	-
	Georgescu et al[16]°*	CVPR22	83.6	89.5	-	-	-	-
	Hirschorn et al[20]	ICCV23	85.9	-	71.8	-	-	-
	Liu et al[27]▽*	CVPR23	85.0	91.4	-	-	-	-
	Cao et al[5]△	CVPR23	79.2	-	-	-	68.2	-
	Barbalau et al[3]	CVIU23	83.8	90.5	62.1	86.5	-	-
	Micorek et al[30]	CVPR24	86.7	91.5	-	-	-	-
	Zhang et al[46]	CVPR24	85.1	89.8	-	-	67.3	70.9
	Zhang et al[46]▽*	CVPR24	87.5	**93.0**	-	-	-	-
	Zhang et al[46]△	CVPR24	-	-	-	-	70.1	72.2
	Ours	-	<u>89.3</u>	91.5	<u>79.3</u>	<u>90.1</u>	<u>72.2</u>	**87.6**
	Ours†	-	**89.5**	<u>91.6</u>	**82.7**	**91.4**	**73.4**	<u>87.5</u>

Table 1. Comparison wtih SOTA methods of the micro and macro AUC(%) on ShanghaiTech, UBnoraml and NWPUCampus datasets. The best-performing results are highlighted in bold, and the second-best results are underlined. ▽: Methods apply virtual dataset for training. ∗: Methods utilize multi-task model ([15] or [16]) as backbone. △: Methods utilize scene-conditioned model [5] as backbone. ○: Methods utilize SSPCAB block [34]. †: tunning α, β, and γ for each dataset.

and macro-AUC scores of 89.5% and 91.6% on the ShanghaiTech dataset, respectively, showing a 2.0% improvement in micro AUC compared to the results of Zhang et al. [46], even without using a virtual dataset or a multi-task model.

Results on UB. The proposed method achieved micro- and macro-AUC scores 82.7% and 91.4% on the UBnormal dataset. These scores represent 9.9% and 4.9% improvements in the micro- and macro-AUC, respectively, compared to the other SOTA methods. This significant margin indicates that the proposed method is particularly effective on the UBnormal dataset.

Results on NWPU. The proposed method achieved micro- and marco-AUC scores of 73.4% and 87.5% on the NWPUCampus dataset, respectively. Compared to the results of Zhang et al. [46], the proposed method surpasses their micro-AUC by more than 3.3% and their macro-AUC by 15.3%, even without relying on a scene-conditioned model.

4.4. Visual results

Fig. 3 illustrates examples where UniVAD detects all types of anomalies (human, appearance, and non-object) more effectively than frame-centric MNAD [32] and the object-centric STG-NF [20]. In Fig. 3 (a), the scene includes a non-object anomaly (fire) and a human anomaly (a person collapsing). MNAD detected both human and non-object anomalies but assigned high scores to normal events as well, resulting in a micro-AUC score of 25.5%. STG-NF effectively distinguished normal and abnormal human anomalies but failed to detect non-object anomalies, achieving only a micro-AUC score of 73.0%. In contrast, UniVAD separated normal and abnormal instances for both human and non-object anomalies, achieving the highest micro-AUC score of 100.0%. In Fig. 3 (b), the scene includes a human anomaly involving group bullying and an appearance anomaly of a bicycle passing on a pedestrian path. MNAD successfully detected the human and appearance anomalies

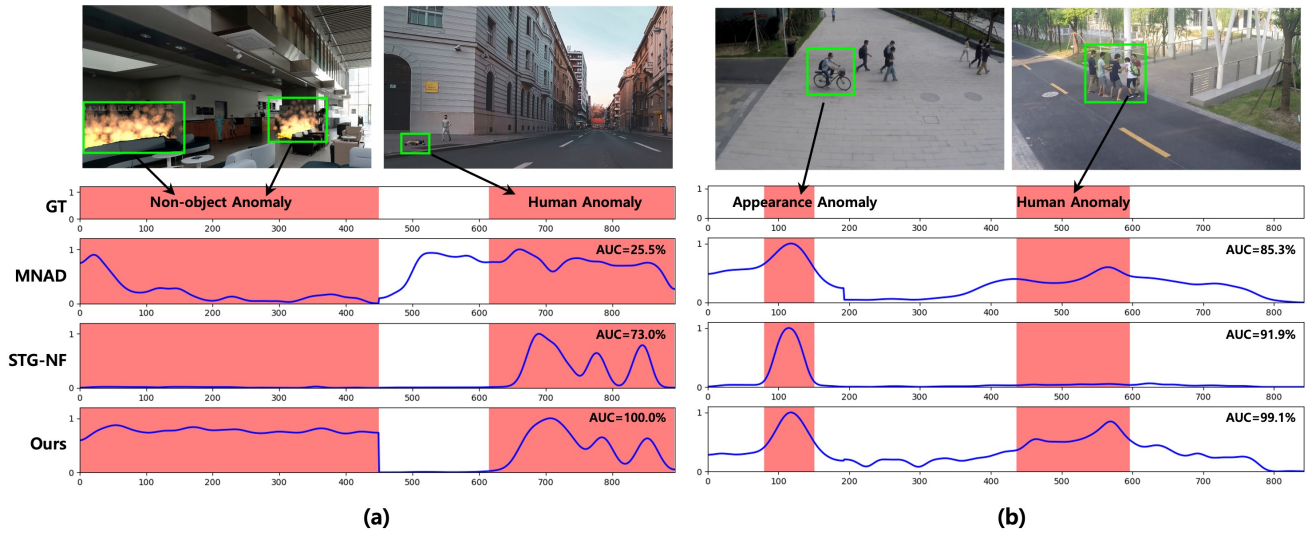

Figure 3. Comparisons of anomaly detection results for various types of anomalies across different methods. (a) Concatenate the video containing non-object anomalies with the video containing human anomalies in the UB dataset. (b) Concatenate the video containing appearance anomalies with the video containing human anomalies in the ShT dataset. Metric is micro AUC. From top to bottom: anomaly type, result of MNAD [32] (frame-centric method), result of STG-NF [20] (object-centric method) and result of ours.

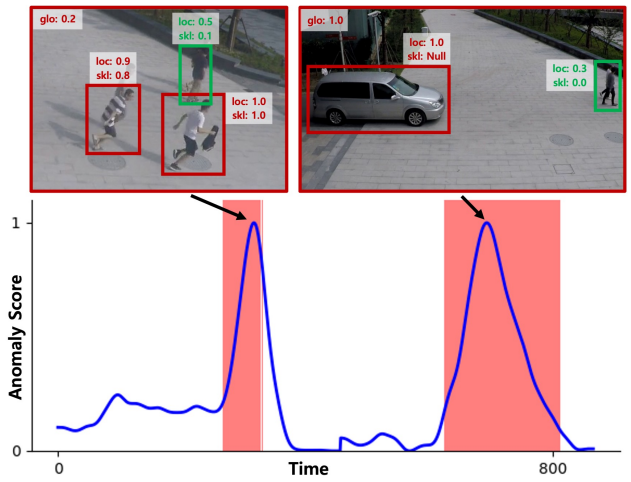

Figure 4. Visualization of anomaly scores from three streams for various types of anomalies. The example shows a concatenation of a video containing a theft with that of a car driving in the ShT.

but also assigned high scores to normal events, resulting in a micro-AUC score of 85.3%. STG-NF effectively distinguished normal and abnormal appearance anomalies but failed to separate normal and abnormal human anomalies, which require global information, resulting in a micro-AUC score of 91.9%. In contrast, UniVAD accurately distinguished normal and abnormal instances for both human and appearance anomalies, achieving the highest micro-AUC score of 99.1%.

Fig. 4 visualizes anomaly scores for normal and abnormal samples across each stream. This video frame includes a human anomaly of someone stealing a bag and an appearance anomaly of a car passing over a pedestrian path. In the global-visual stream, a low anomaly score of 0.2 was assigned to the human anomaly frame, where the anomaly area was small, whereas a high anomaly score of 1.0 was assigned to the appearance anomaly frame with a larger anomaly area. The local-visual stream generally assigned high scores to anomalous samples but failed to completely suppress the anomaly scores of the normal samples. The skeleton stream successfully distinguished running anomalous samples from walking normal samples but failed to detect the car, which is an appearance anomaly. Therefore, utilizing all three streams is effective for detecting various types of anomalies.

4.5. Ablation study

Effect of the three streams. The VAD datasets contain various types of anomalies with different proportions for each type, which leads to performance differences among the three streams. As shown in Tab. 2, the ShT dataset has a higher proportion of human and appearance anomalies, which results in the largest performance improvement of 4.5% when the local-visual stream is added to the skeleton stream. In contrast, on the UB dataset, which has a higher proportion of human and non-object anomalies, adding the global stream to the skeleton stream led to the largest performance increase of 1.5%. Finally, NWPU is a large dataset composed of scene-dependent anomalies, where the skeleton stream, which does not account for the scene context,

Skeleton stream	Local stream	Global stream	ShT	UB	NWPU
✓			85.2	81.2	65.1
	✓		85.9	63.7	68.8
		✓	67.3	62.0	69.2
✓	✓		89.4	81.8	69.7
✓		✓	86.2	**82.7**	72.6
	✓	✓	86.0	63.8	72.3
✓	✓	✓	**89.5**	**82.7**	**73.4**

Table 2. Ablation experiments of the three streams in terms of the micro-AUC (%) on the ShT, UB, and NWPU datasets.

exhibited the lowest performance, whereas the global-visual stream, which considers the overall scene, achieved the highest performance. Nevertheless, the combination of the three streams resulted in the largest improvement of 8.3%.

Effect of the FFPP task. Tab. 3 presents the results of the ablation study conducted on the ShT and UB datasets to demonstrate that the proposed FFPP task effectively suppresses the generalization capability of AEs compared to other tasks. Rec reconstructs all features from past to future using them as input, whereas FP predicts future features by taking all features from past to present as input. BiP predicts the entire features by taking both past and future features as input, whereas the proposed FFPP task predicts past, present and future features based on the present feature. The FFPP task achieved a performance improvement of 0.8%-9.4% across all streams compared to the other tasks. This effect was particularly pronounced in the skeleton and local-visual streams, as they are object-related tasks.

Effect of skeletal losses. In Tab. 4, we analyze the impact of the loss function on training the skeleton stream through an ablation study on the ShT, UB, and NWPU datasets. STL resulted in a 0.2%-1.1% higher micro-AUC performance than SSL, indicating that learning temporal relationships is more important than learning spatial relationships for anomaly detection. The highest performance was achieved when both losses were used together.

5. Discussion

5.1. Limitation

The proposed UniVAD model has two main limitations. First, similar to most prior VAD methods, our model relies on several pretrained models, such as the object detector, visual encoder, and pose extractor. This dependency means that performance may vary depending on the pretrained models, which could lead to potential failures in specific domains. However, we believe this issue can be further addressed, and our work paves the way for future advancements in VAD. Second, UniVAD comprises three streams requiring the adjustment of many hyperparameters during model optimization. Although the training process is cumbersome, it has a lower optimization difficulty compared to existing methods that train multiple models end-to-end, ensuring higher performance.

Task	Skeleton		Local		Global	
	ShT	UB	ShT	UB	ShT	UB
Rec	82.3	71.8	83.2	58.3	62.3	61.2
FP	77.3	75.5	83.5	58.6	65.7	61.5
BiP	77.1	75.5	84.1	59.4	65.8	61.7
FFPP	**85.2**	**81.2**	**85.9**	**63.7**	**67.3**	**62.0**

Table 3. Ablation experiments on the contributions of each proxy task in the three streams. We report the micro-AUC (%) for the ShT and UB datasets. "Rec", "FP", "BiP" and "FFPP" represent feature reconstruction, future prediction, bidirectional prediction, and future past prediction, respectively.

SSL	STL	ShT	UB	NWPU
✓		83.7	79.1	64.6
	✓	84.8	80.1	64.8
✓	✓	**85.2**	**81.2**	**65.1**

Table 4. Ablation experiments of SSL and STL in terms of the micro-AUC (%) on the ShT, UB, and NWPU datasets.

5.2. Conclusion

We have proposed the UniVAD model, which effectively detects all anomaly types. UniVAD comprised the skeleton stream for detecting human anomalies, the local-visual stream for detecting appearance anomalies, and the global-visual stream for detecting non-object anomalies. In addition, we introduced the FFPP task, which uses present feature to predict both past and future features to suppress the generalization ability of AEs in each stream. This task is applied to all streams, thereby preventing the reconstruction of anomalies and enabling accurate anomaly detection. Furthermore, in the skeleton stream, we proposed SSL and STL loss functions for the temporal and spatial modeling of skeletons, respectively. Extensive experiments on three public VAD datasets demonstrated that UniVAD outperforms existing methods, thereby proving its effectiveness in anomaly detection.

Acknowledgments. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT).(No. RS-2025-02312833)

References

[1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20143–20153, 2022. 5

[2] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020. 3

[3] Antonio Barbalau, Radu Tudor Ionescu, Mariana-Iuliana Georgescu, Jacob Dueholm, Bharathkumar Ramachandra, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Ssmtl++: Revisiting self-supervised multi-task learning for video anomaly detection. *Computer Vision and Image Understanding*, 229:103656, 2023. 1, 6

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 6

[5] Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20392–20401, 2023. 1, 5, 6

[6] Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan. Clustering driven deep autoencoder for video anomaly detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 329–345. Springer, 2020. 2

[7] Chengwei Chen, Yuan Xie, Shaohui Lin, Angela Yao, Guannan Jiang, Wei Zhang, Yanyun Qu, Ruizhi Qiao, Bo Ren, and Lizhuang Ma. Comprehensive regularization in a bi-directional predictive network for video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 230–238, 2022. 2

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[9] Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, and David Aik-Aun Khoo. Tevad: Improved video anomaly detection with captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5559, 2023. 2

[10] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 3

[11] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2969–2978, 2022. 3

[12] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7157–7173, 2022. 3, 5

[13] Zhiwen Fang, Jiafei Liang, Joey Tianyi Zhou, Yang Xiao, and Feng Yang. Anomaly detection with bidirectional consistency in videos. *IEEE transactions on neural networks and learning systems*, 33(3):1079–1092, 2020. 2

[14] Alessandro Flaborea, Luca Collorone, Guido Maria D'Amely Di Melendugno, Stefano D'Arrigo, Bardh Prenkaj, and Fabio Galasso. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10318–10329, 2023. 4

[15] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12742–12752, 2021. 1, 5, 6

[16] Mariana Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4505–4523, 2021. 5, 6

[17] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1705–1714, 2019. 1, 2, 6

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[19] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016. 2

[20] Or Hirschorn and Shai Avidan. Normalizing flows for human pose anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13545–13554, 2023. 1, 2, 4, 6, 7

[21] Jinlei Hou, Yingying Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, and Hong Zhou. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8791–8800, 2021. 2

[22] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7842–7851, 2019. 1, 6

[23] Asiegbu Miracle Kanu-Asiegbu, Ram Vasudevan, and Xiaoxiao Du. Bipoco: Bi-directional trajectory prediction with pose constraints for pedestrian anomaly detection. *arXiv preprint arXiv:2207.02281*, 2022. 1, 2, 4

[24] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Bman: Bidirectional multi-scale aggregation networks for abnormal event detection. *IEEE Transactions on Image Processing*, 29:2395–2408, 2019. 2

[25] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018. 1, 2, 3, 5, 6

[26] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13588–13597, 2021. 1, 6

[27] Zuhao Liu, Xiao-Ming Wu, Dian Zheng, Kun-Yu Lin, and Wei-Shi Zheng. Generating anomalies for video anomaly detection with prompt-based feature mapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24500–24510, 2023. 1, 2, 5, 6

[28] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15425–15434, 2021. 2, 6

[29] Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. Graph embedded pose clustering for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10539–10547, 2020. 4

[30] Jakub Micorek, Horst Possegger, Dominik Narnhofer, Horst Bischof, and Mateusz Kozinski. Mulde: Multiscale log-density estimation via denoising score matching for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18868–18877, 2024. 1, 2, 5, 6

[31] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11996–12004, 2019. 3, 4

[32] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14372–14381, 2020. 1, 2, 6, 7

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[34] Nicolae-Cătălin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised predictive convolutional attentive block for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13576–13586, 2022. 2, 3, 5, 6

[35] Nicolae-C Ristea, Florinel-Alin Croitoru, Radu Tudor Ionescu, Marius Popescu, Fahad Shahbaz Khan, Mubarak Shah, et al. Self-distilled masked auto-encoders are efficient video anomaly detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15984–15995, 2024. 5, 6

[36] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2626–2634, 2020. 4

[37] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018. 2, 6

[38] Shengyang Sun and Xiaojin Gong. Hierarchical semantic contrast for scene-aware video anomaly detection. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 22846–22856, 2023. 2, 3

[39] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *European Conference on Computer Vision*, pages 494–511. Springer, 2022. 1, 3, 5

[40] Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu Wang, and Qi Qi. Robust unsupervised video anomaly detection by multipath frame prediction. *IEEE transactions on neural networks and learning systems*, 33(6):2301–2312, 2021. 2, 6

[41] Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffusion model for video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5527–5537, 2023. 1, 2, 6

[42] Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: reasoning for video anomaly detection with large language models. *arXiv preprint arXiv:2407.10299*, 2024. 6

[43] Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14592–14601, 2023. 2

[44] Jongmin Yu, Younkwan Lee, Kin Choong Yow, Moongu Jeon, and Witold Pedrycz. Abnormal event detection and localization via adversarial event prediction. *IEEE transactions on neural networks and learning systems*, 33(8):3572–3586, 2021. 2

[45] M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative

learning for unsupervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14744–14754, 2022. 2

[46] Menghao Zhang, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, and Jianxin Liao. Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17385–17394, 2024. 1, 5, 6

[47] Yu Zhang, Zhongyin Guo, Jianqing Wu, Yuan Tian, Haotian Tang, and Xinming Guo. Real-time vehicle detection based on improved yolo v5. *Sustainability*, 14(19):12274, 2022. 1, 5

[48] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer, 2022. 1, 5

[49] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1933–1941, 2017. 2

Unified Video Anomaly Detection Model for Detecting Different Anomaly Types

Supplementary Material

In this supplement, we provide the followings:

- Experiment 6 for effect of the prediction in visual streams.
- Experiment 7 for effect of the time T of objects or frames.
- Experiment 8 for effect of α, β, and γ in inference.
- More visual results 9 for skeleton, local-visual, and global-visual streams.
- Experiment 10 for visual feature extractors.
- Running time of UniVAD 11.

6. Different combinations of hyperparmeters

Since the types of anomalies vary across datasets, we experiment with different combinations of α, β, and γ to find the optimal hyperparameters, as shown in Tab. Tab. 5. Overall, due to the high rate of human anomalies in all datasets, setting α to 1.00 (IDs 0-9) yields better performance across all datasets compared to other settings (IDs 10-21). In the NWPU dataset, where capturing scene-dependency is crucial, ID 0 demonstrates significant improvements over IDs 1-2. In the ShT dataset, which contains numerous human anomalies and a few appearance anomalies, ID 5 outperforms IDs 2 and 8 significantly. Similarly, in the UB dataset, which includes numerous human anomalies and few nonobject and appearance anomalies, ID 8 achieves significant gains compared to IDs 6-7.

ID	α	β	γ	ShT	UB	NWPU
0	**1.00**	**1.00**	**1.00**	86.3	70.1	**73.4**
1	1.00	1.00	0.10	88.5	69.3	71.3
2	1.00	1.00	0.01	88.4	68.7	70.2
3	1.00	0.10	1.00	78.8	71.3	72.2
4	1.00	0.10	0.10	89.3	79.3	72.2
5	**1.00**	**0.10**	**0.01**	**89.5**	78.6	69.5
6	1.00	0.01	1.00	75.9	71.1	71.9
7	1.00	0.01	0.10	86.6	81.6	71.6
8	**1.00**	**0.01**	**0.01**	86.7	**82.7**	69.0
9	1.00	0.00	0.00	85.2	81.2	65.1
10	0.10	1.00	1.00	84.4	66.7	72.5
11	0.10	1.00	0.10	86.5	65.2	70.7
12	0.10	1.00	0.01	86.3	64.4	69.5
13	0.10	0.10	1.00	73.3	64.6	70.4
14	0.10	0.01	1.00	69.1	63.7	69.8
15	0.01	1.00	1.00	84.1	66.3	72.3
16	0.01	1.00	0.10	86.1	64.6	70.5
17	0.01	1.00	0.01	85.9	63.9	69.3
18	0.01	0.10	1.00	72.5	63.4	70.1
19	0.01	0.01	1.00	68.1	62.4	69.3
20	0.00	1.00	0.00	85.9	63.7	68.8
21	0.00	0.00	1.00	67.3	62.0	69.2

Table 5. Ablation experiments of the hyperparameters α, β, and γ in terms of micro-AUC (%) on the ShT, UB, and NWPU datasets. α, β, and γ are chosen from the set [1.00, 0.10, 0.01, 0.00].

7. Effect of the prediction in visual streams

Tab. 6 shows the results of the ablation study conducted on the ShT and UB datasets to demonstrate that predicting past, present, and future features effectively improves performance for visual streams compared to predicting other features. Predicting key frame features (past, present, and future) efficiently trains the model better than predicting all features, resulting in improved detection performance.

8. Effect of the time of objects or frames

In Tab. 7, we conduct an ablation study on the ShT and UB datasets to analyze the impact of the time of objects or frames on detection performance. In general, as the time T increases, prediction becomes more challenging, leading to decreased performance. In the skeleton stream, the optimal T values are 24 and 16 for the ShT and UB datasets, respectively. In the local-visual and global-visual streams, the optimal T value is 8 for both the ShT and UB datasets.

9. Visual results for three streams

Fig. 5 presents the results of anomaly detection by the skeleton, local-visual, and global-visual streams for various types of anomalies. In Fig. 5 (a), since the anomaly is a human anomaly involving running in a place where running is not allowed, the skeleton stream performs better than local-visual, global-visual stream. In Fig. 5 (b), since the anomaly is a combination of a human anomaly (a person riding a bike) and an appearance anomaly (the bike), the skeleton and local-visual stream perform good performance. In Fig. 5 (c), the anomaly is a nonobject anomaly involving smoke caused by a car accident. The skeleton and local-visual streams fail to detect the anomaly, while the global-visual stream successfully detects it.

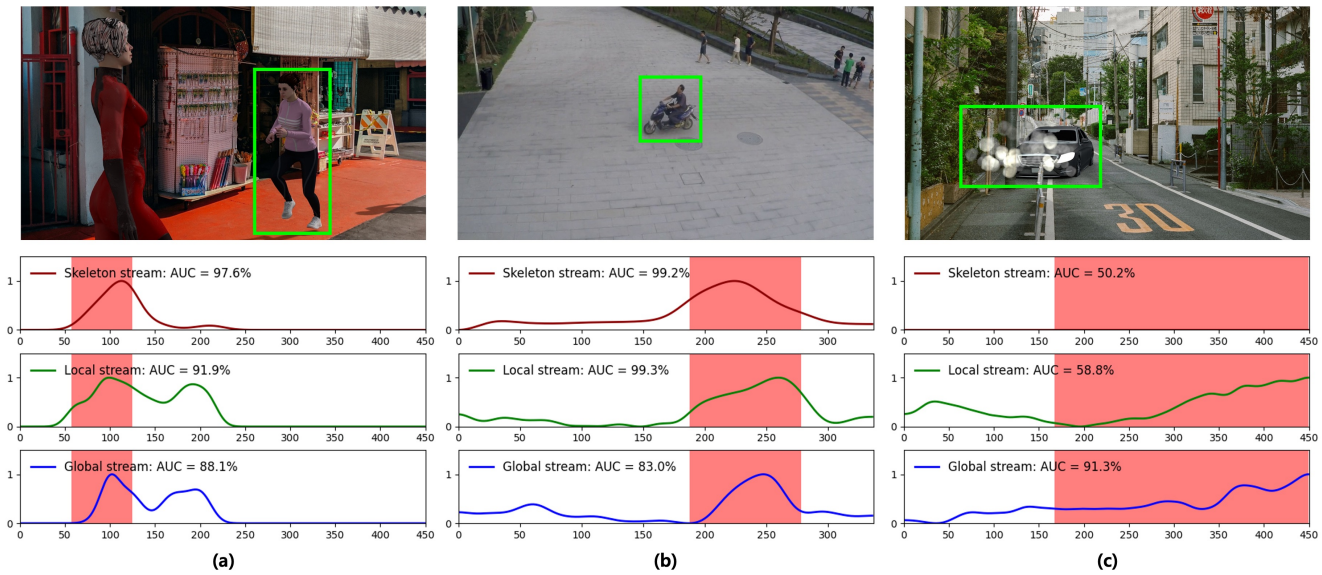

Figure 5. Comparisons of anomaly detection results for various types of anomalies across three streams. (a) The video containing human anomalies in the UB dataset. (b) The video containing appearance anomalies in the ShT dataset. (c) The video containing nonobject anomalies in the UB dataset. The metric used is micro-AUC. From top to bottom: results from the skeleton stream, results from the local-visual stream, and results from the global-visual stream.

Prediction	Local		Global	
	ShT	UB	ShT	UB
All features	82.8	62.4	52.6	58.5
Present features	83.2	58.3	62.3	61.2
Future, Past features	84.8	63.0	65.3	61.5
Future, Present, Past features	**85.9**	**63.7**	**67.3**	**62.0**

Table 6. Ablation experiments on the prediction of features in visual streams. We report the micro-AUC (%) for the ShT and UB datasets. Here, "All features" refers to predicting T features from past to future, "Present features" refers to predicting only the present features, "Future, Past features" refers to predicting both past and future features, and "Future, Present, Past features" refers to predicting past, present and future features.

T	Skeleton		Local		Global	
	ShT	UB	ShT	UB	ShT	UB
8	78.5	79.2	**85.9**	**63.7**	**67.3**	**62.0**
16	84.6	**81.2**	83.8	62.9	64.6	61.6
24	**85.2**	79.9	83.0	62.2	61.2	60.7
32	84.2	78.5	82.6	61.8	59.7	60.0

Table 7. Ablation experiments on the time T of the sequence of objects or frames in the three streams. We report the micro-AUC (%) for the ShT and UB datasets.

10. The choice of feature extractors

In Tab. 8 and Tab. 9, we compare the performance of local-, global- stream and UniVAD used with different image feature extractors. We observe, that CLIP(ViT-L-14) mostly outperforms CLIP(ViT-B-32) in experiments, but runs considerably slower. Among various CLIP versions, we selected CLIP (ViT-B/32) due to its favorable trade-off between computational efficiency and accuracy.

11. Running time

We conducted all our experiments on an NVIDIA RTX 3090 GPU. The object detection and tracker take approximately 39 milliseconds (ms) per frame. The pose extractor takes approximately 19 ms, and the visual encoder takes approximately 2.5 ms. Computing skeleton data, local-visual features, and global-visual features across all streams takes approximately 0.02 ms. UniVAD runs at 16.5 FPS with an average of 5 objects per frame.

Backbone	Local stream			Global stream			UniVAD		
	ShT	UB	NWPU	ShT	UB	NWPU	ShT	UB	NWPU
CLIP(Resnet-50)	76.8	62.7	67.9	67.8	60.9	68.0	85.7	78.3	70.6
CLIP(Resnet-101)	81.9	64.9	67.0	68.4	59.1	66.7	87.7	78.0	69.2
CLIP(ViT-B/32)	85.9	63.7	**68.8**	67.3	62.0	69.2	89.3	79.3	**72.2**
CLIP(ViT-B/16)	85.1	65.8	64.0	66.5	60.6	67.7	88.7	78.7	69.8
CLIP(ViT-L/14)	**87.4**	**67.3**	67.2	**70.6**	**63.7**	**69.5**	**89.4**	**80.8**	71.4

Table 8. Micro AUC-ROC (%) comparison. For each stream feature representation CLIP(ViT-B/32), CLIP(ViT-B/16), CLIP(ViT-L/14), CLIP(Resnet-50), and CLIP(Resnet-101), we mark the best scores bold.

Backbone	Local stream			Global stream			UniVAD		
	ShT	UB	NWPU	ShT	UB	NWPU	ShT	UB	NWPU
Clip(Resnet-50)	84.4	83.7	83.2	74.0	78.9	83.7	90.3	90.1	87.4
Clip(Resnet-101)	86.1	84.8	83.4	**75.9**	78.4	83.4	90.9	90.3	87.5
Clip(ViT-B/32)	86.8	84.8	83.8	75.1	80.0	83.6	91.5	90.1	87.6
Clip(ViT-B/16)	87.6	**85.6**	84.6	74.8	79.4	84.2	91.2	90.0	88.5
Clip(ViT-L/14)	**89.2**	85.2	**85.0**	75.1	**82.2**	**84.5**	**91.6**	**91.4**	**88.7**

Table 9. Macro AUC-ROC (%) comparison. For each stream feature representation CLIP(ViT-B/32), CLIP(ViT-B/16), CLIP(ViT-L/14), CLIP(Resnet-50), and CLIP(Resnet-101), we mark the best scores bold.