VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection

Peng Wu[1], Xuerong Zhou[1], Guansong Pang[2*], Lingru Zhou[1], Qingsen Yan[1], Peng Wang[1*], Yanning Zhang[1]

[1]ASGO, School of Computer Science, Northwestern Polytechnical University, China
[2]School of Computing and Information Systems, Singapore Management University, Singapore
{xdwupeng, zxr2333}@gmail.com, gspang@smu.edu.sg, {lingruzhou, yqs}@mail.nwpu.edu.cn, {peng.wang, ynzhang}@nwpu.edu.cn

Abstract

The recent contrastive language-image pre-training (CLIP) model has shown great success in a wide range of image-level tasks, revealing remarkable ability for learning powerful visual representations with rich semantics. An open and worthwhile problem is efficiently adapting such a strong model to the video domain and designing a robust video anomaly detector. In this work, we propose VadCLIP, a new paradigm for weakly supervised video anomaly detection (WSVAD) by leveraging the frozen CLIP model directly without any pre-training and fine-tuning process. Unlike current works that directly feed extracted features into the weakly supervised classifier for frame-level binary classification, VadCLIP makes full use of fine-grained associations between vision and language on the strength of CLIP and involves dual branch. One branch simply utilizes visual features for coarse-grained binary classification, while the other fully leverages the fine-grained language-image alignment. With the benefit of dual branch, VadCLIP achieves both coarse-grained and fine-grained video anomaly detection by transferring pre-trained knowledge from CLIP to WSVAD task. We conduct extensive experiments on two commonly-used benchmarks, demonstrating that VadCLIP achieves the best performance on both coarse-grained and fine-grained WSVAD, surpassing the state-of-the-art methods by a large margin. Specifically, VadCLIP achieves 84.51% AP and 88.02% AUC on XD-Violence and UCF-Crime, respectively. Code and features are released at https://github.com/nwpu-zxr/VadCLIP.

Introduction

In recent years, weakly supervised video anomaly detection (WSVAD, VAD) has received growing concerns due to its broad application prospects. For instance, with the aid of WSVAD, it is convenient to develop more powerful intelligent video surveillance systems and video content review systems. In WSVAD, the anomaly detector is expected to generate frame-level anomaly confidences with only video-level annotations provided. The majority of current research in this field follows a systematic process, wherein the initial step is to extract frame-level features using pre-trained visual models, e.g., C3D (Tran et al. 2015; Sultani, Chen, and Shah 2018), I3D (Carreira and Zisserman 2017; Wu et al. 2020), and ViT (Dosovitskiy et al. 2020; Li, Liu, and Jiao 2022), followed by feeding these features into multiple instance learning (MIL) based binary classifiers for the purpose of model training, and the final step is to detect abnormal events based on predicted anomaly confidences. Despite their simple schemes and promising results, such a classification-based paradigm fails to take full advantage of cross-modal relationships, e.g, vision-language associations.

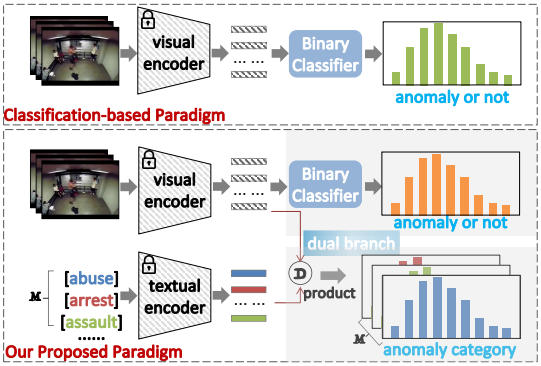

Figure 1: Comparisons of different paradigms for WSVAD.

During the past two years, we have witnessed great progress in the development of vision-language pre-training (VLP) models (Kim, Son, and Kim 2021; Jia et al. 2021; Wang et al. 2021; Chen et al. 2023a), e.g., CLIP (Radford et al. 2021), for learning more generalized visual representations with semantic concepts. The main idea of CLIP is to align images and texts by contrastive learning, that is, pull together images and matched textual descriptions while pushing away unmatched pairs in the joint embedding space. Thanks to hundreds of million noisy image-text pairs crawled from the web, such models pre-trained at a large scale really demonstrate their strong representation learning as well as associations between vision and language. In view of the breakthrough performance of CLIP, recently, building task-specific models on top of CLIP is becoming emerging research topics and applied to a broad range of vision tasks, and these models achieve unprecedented performance.

Although CLIP and its affiliated models demonstrate the great potential on various vision tasks, these methods mainly focus on the image domain. Therefore, how to efficiently adapt such a model learned from image-text pairs to more

*Corresponding Authors

complex video anomaly detection task under weak supervision deserves a thorough exploration. Recently, a few works (Joo et al. 2023; Lv et al. 2023) attempt to make use of the learned knowledge of CLIP, however, these methods limit their scope to directly using visual features extracted from the image encoder of CLIP, and neglect to exploit semantic relationships between vision and language.

In order to make effective use of generalized knowledge and enable CLIP to reach its full potential on WSVAD task, based on the characteristics of WSVAD, there are several critical challenges that need to be addressed. First, it is vital to explore ways to capture contextual dependencies across time. Second, it is essential to determine how to harness learned knowledge and the visual-language connections. Third, it is crucial to maintain optimal CLIP performance under weak supervision.

In this work, we propose a novel paradigm based on CLIP for WSVAD, which is dubbed as **VadCLIP**. VadCLIP consists of several components to overcome the above challenges. Specifically, **for the first challenge**, we present a local-global temporal adapter (LGT-Adapter), which is a lightweight module for video temporal relation modeling. LGT-Adapter involves two components, i.e., local temporal adapter and global temporal adapter, wherein the former mainly captures local temporal dependencies with high efficiency, since in most cases the current events are highly related to the adjacent events, and the latter smooths feature information in a more holistic view with less parameters. **For the second challenge**, unlike current methods (Joo et al. 2023; Lv et al. 2023) that solely use visual features, we encourage VadCLIP to also leverage textual features to preserve learned knowledge as much as possible. As shown in Figure 1, VadCLIP is devised as a dual-branch fashion, where one simply and directly utilizes visual features for binary classification (C-branch), while the other employs both visual and textual features for language-image alignment (A-branch). Moreover, such dual branch seamlessly achieves coarse-grained and fine-grained WSVAD (Wu, Liu, and Liu 2022). For A-branch, we build bridge between videos and video-level textual labels. Moreover, we propose two prompt mechanisms (Wu et al. 2023), i.e., learnable prompt and visual prompt, to specify that the succinct text is about the video. Learnable prompt does not require extensive expert knowledge compared to the handcrafted prompt, effectively transfers pre-trained knowledge into the downstream WSVAD task. Visual prompt is inspired by that visual contexts can make the text more accurate and discriminate. Imagine that if there is a car in the video, two types of abnormal events of "car accident" and "fighting" would be more easily distinguished. Hence, In the visual prompt, we focus on anomaly information in videos and integrate these anomaly-focus visual contents from C-branch with textual labels from A-branch for automatic prompt engineering. Such a practice seamlessly creates connections between dual branch. **For the third challenge**, multiple instance learning (MIL) (Sultani, Chen, and Shah 2018; Wu et al. 2020) is the most commonly used method. For the language-visual alignments in A-branch, we introduce a MIL-Align mechanism, the core idea is to select the most matched video frames for each label to represent the whole video.

Note that during training, the weights of CLIP image and text encoders are kept fixed, and the gradients are back-propagated to optimise these learnable parameters of the devised adapter and prompt modules.

Overall, the contributions of our work are threefold:
(1) We present a novel diagram, i.e., VadCLIP, which involves dual branch to detect video anomaly in visual classification and language-visual alignment manners, respectively. With the benefit of dual branch, VadCLIP achieves both coarse-grained and fine-grained WSVAD. To our knowledge, VadCLIP is the first work to efficiently transfer pre-trained language-visual knowledge to WSVAD.
(2) We propose three non-vital components to address new challenges led by the new diagram. LGT-Adapter is used to capture temporal dependencies from different perspectives; Two prompt mechanisms are devised to effectively adapt the frozen pre-trained model to WSVAD task; MIL-Align realizes the optimization of alignment paradigm under weak supervision, so as to preserve the pre-trained knowledge as much as possible.
(3) We show that strength and effectiveness of VadCLIP on two large-scale popular benchmarks, and VadCLIP achieves state-of-the-art performance, e.g., it obtains unprecedented results of 84.51% AP and 88.02% AUC on XD-Violence and UCF-Crime respectively, surpassing current classification based methods by a large margin.

Related Work

Weakly Supervised Video Anomaly Detection

Recently, some researchers (Zaheer et al. 2020; Feng, Hong, and Zheng 2021; Wu et al. 2021; Chen et al. 2023b) have proposed weakly supervised methods for VAD. Sultani et al. (Sultani, Chen, and Shah 2018) firstly proposed a deep multiple instance learning model, which considers a video as a bag and its multiple segments as instances. Then several follow-up works made effort to model temporal relations based on self-attention models and transformers. For example, Zhong et al. (Zhong et al. 2019) proposed a graph convolutional network (GCN) based method to model the feature similarity and temporal consistency between video segments. Tian et al. (Tian et al. 2021) used a self-attention network to capture the global temporal context relationship of videos. Li et al. (Li, Liu, and Jiao 2022) proposed a transformer based multi-sequence learning framework, and Huang et al. (Huang et al. 2022) proposed a transformer based temporal representation aggregation framework. Zhou et al. (Zhou, Yu, and Yang 2023) presented a global and local multi-head self attention module for the transformer layer to obtain more expressive embeddings for capturing temporal dependencies in videos. The above methods only detect whether video frames are anomalous, on the contrary, Wu et al. (Wu, Liu, and Liu 2022) proposed a fine-grained WSVAD method, which distinguishes between different types of anomalous frames. More recently, the CLIP model has also attracted great attentions in the VAD community. Based on visual features of CLIP, Lv et al. (Lv et al. 2023) proposed a new MIL framework called Unbiased MIL (UMIL)

to learn unbiased anomaly features that improve WSVAD performance. Joo et al. (Joo et al. 2023) proposed to employ visual features from CLIP to efficiently extract discriminative representations, and then model long- and short-range temporal dependencies and nominate the snippets of interest by leveraging temporal self-attention. All the above methods are based on the classification paradigm, which detect anomalous events by predicting the probability of anomalous frames. However, this classification paradigm does not fully utilize the semantic information of textual labels.

Vision-Language Pre-training

Vision-language pre-training has achieved impressive progress over the past few years, which aims to learn the semantic correspondence between vision and language through pre-training on large-scale data. As one of the most representative works, CLIP has shown impressive performance on a range of vision-language downstream tasks, including image classification (Zhou et al. 2022a), image captioning (Mokady, Hertz, and Bermano 2021), object detection (Zhou et al. 2022b), scene text detection (Yu et al. 2023), dense prediction (Zhou et al. 2022c; Rao et al. 2022), and so on. Recently, some follow-up works attempted to leverage the pre-trained models for video domains. For example, CLIP4Clip (Luo et al. 2022) transferred the knowledge of CLIP model to the video-text retrieval, some works (Wang, Xing, and Liu 2021; Lin et al. 2022; Ni et al. 2022) attempted to take advantages of CLIP for video recognition, furthermore, CLIP is used to tackle the more complex video action localization task (Nag et al. 2022; Ju et al. 2022). More generally, Ju et al. (Ju et al. 2022) presented a simple yet strong baseline to efficiently adapt the pre-trained image-based visual-language model, and exploited its powerful ability for general video understanding. In this work, we deeply explore how to adapt pre-trained vision-language knowledge of CLIP from image-level into video-level downstream WSVAD efficiently.

Method

Problem Definition

The WSVAD task supposes that only video-level labels are available during the training stage. Given a video v, if all frames of this video do not contain abnormal events, this video is defined as normal with the label $y = 0$; Otherwise, if there is at least one frame contains abnormal events, this video is labeled as abnormal with the label $y = 1$. The goal of WSVAD task is to train a detection model that is able to predict frame-level anomaly confidences while only video-level annotations are provided.

Previous works generally make use of pre-trained 3D convolutional models, e.g., C3D (Tran et al. 2015) and I3D (Carreira and Zisserman 2017), to extract video features, and then feed these features into MIL-based binary classifiers, such paradigms are referred as the classification-based paradigm in this paper. Recently, CLIP, as a large-scale language-vision pre-trained model, has revolutionized many fields in computer vision, and has shown great generalization capabilities across a wide range of downstream tasks. Inspired by CLIP, our work not only uses the image encoder of CLIP as the backbone to extract video features, but also attempts to utilize the text encoder of CLIP to take full advantage of the powerful associations between visual contents and textual concepts. Our work is demonstrated in Figure 2.

Local and Global Temporal Adapter

As we know, CLIP is pre-trained on large-scale image-text pairs crawled from the web. In this section, we investigate how to model temporal dependencies and bridge the gap between the image domain and video domain for CLIP. Meanwhile, it is also significant to learn long-range and short-range temporal dependencies for WSVAD task (Zhou, Yu, and Yang 2023; Wu and Liu 2021). From the perspective of the efficiency and receptive field, we design a new temporal modeling method compatible with local and global receptive field.

Local Module. To capture local temporal dependencies, we introduce a transformer encoder layer on top of frame-level features $X_{clip} \in \mathbb{R}^{n \times d}$ from the frozen image encoder of CLIP, where n is the length of video, d is the dimension size, which is set as 512 in this work. Note that this layer differs from the ordinary transformer encoder layer since it limits self-attention computation to local windows (Liu et al. 2021) instead of the global scope. Specifically, we split frame-level features into equal-length and overlapping windows over temporal dimension, self-attention calculation is limited within each window, and no information exchange among windows. Such an operation possesses local receptive field like the convolution, and leads to the lower computation complexity.

Global Module. To further capture global temporal dependencies, we introduce a lightweight GCN module following local module, we adopt GCN to capture global temporal dependencies due to its widespread adoption and proven performance in VAD (Zhong et al. 2019; Wu et al. 2020; Wu and Liu 2021). Following the setup in (Zhong et al. 2019; Wu et al. 2020), we use GCN to model global temporal dependencies from the perspective of feature similarity and relative distance, it can be summarized as follows,

$$X_g = gelu\left(\left[Softmax\left(H_{sim}\right); Softmax\left(H_{dis}\right)\right] X_l W\right) \tag{1}$$

where H_{sim} and H_{dis} are the adjacency matrices, the Softmax normalization is used to ensure the sum of each row of H_{sim} and H_{dis} equals to one. X_l is the frame-level video feature obtained from local module, W is the only one learnable weight that is used to transform the feature space, this setup demonstrates the lightweight of global module.

Feature similarity branch is designed to generated a similarity relationship adjacency matrix for GCN. We use the frame-wise cosine similarity to calculate the adjacency matrix H_{sim}, which is presented as follows,

$$H_{sim} = \frac{X_l X_l^\top}{\|X_l\|_2 \cdot \|X_l\|_2} \tag{2}$$

we also use the thresholding operation to filter weak relations (Wu et al. 2020).

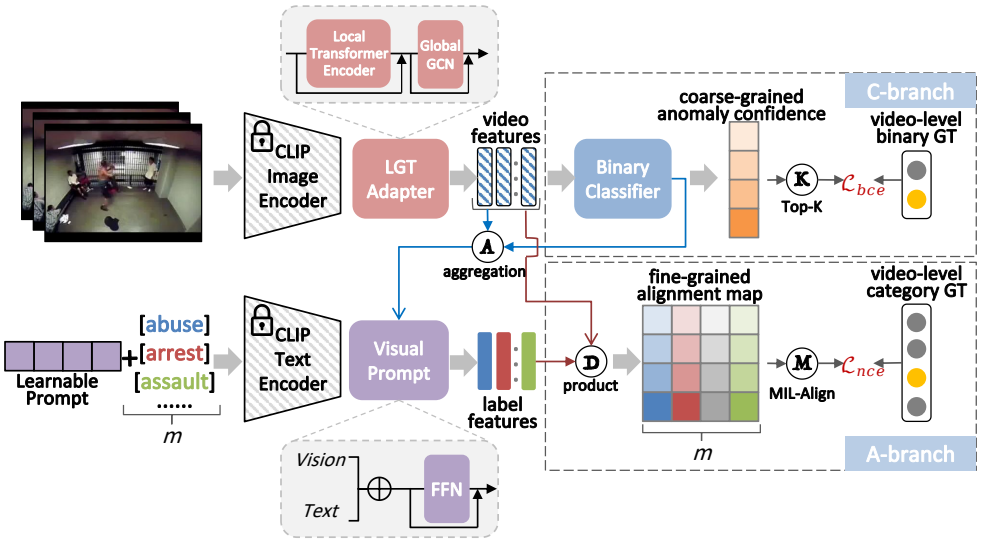

Figure 2: The framework of our proposed VadCLIP.

Position distance branch is used to capture long-range dependencies based on positional distance between each two frames. The proximity adjacency matrix is shown as follows:

$$H_{dis}(i,j) = \frac{-|i-j|}{\sigma} \tag{3}$$

the proximity relation between i^{th} and j^{th} frames only determined by their relative temporal position. σ is a hyper-parameter to control the range of influence of distance relation. Both local transformer and GCN layer employ residual connection to prevent feature over-smoothing.

Dual Branch and Prompt

Dual Branch. Unlike other previous WSVAD works, our VadCLIP contains dual branch, more precisely, in addition to the traditional binary classification branch (C-Branch), we also introduce a novel video-text alignment branch, dubbed as A-Branch. Specifically, after temporal modeling, the video feature X_g is fed into a fully connected (FC) layer to obtain the final video feature $X \in \mathbb{R}^{n\times d}$. In C-Branch, we feed X into a binary classifier that contains a feed-forward network (FFN) layer, an FC layer and a Sigmoid activation to obtain the anomaly confidence $A \in \mathbb{R}^{n\times 1}$.

$$A = Sigmoid\left(FC\left(FFN\left(X\right)+X\right)\right) \tag{4}$$

In A-Branch, textual labels, e.g., abuse, riot, fighting, etc, are no longer encoded as one-hot vectors, on the contrary, they are encoded into class embeddings using the text encoder of CLIP, we leverage the frozen pre-trained text encoder of CLIP throughout, as the text encoder can provide language knowledge prior for video anomaly detection. Then we calculate the match similarities between class embeddings and frame-level visual features to obtain the alignment map $M \in \mathbb{R}^{n\times m}$, where m is the number of text labels, such a setup is similar to that of CLIP. In A-Branch, each input text label represents a class of abnormal events, thus naturally achieving fine-grained WSVAD.

Learnable Prompt. In WSVAD, text labels are words or phrases, which are too succinct to summarize abnormal events very well. To learn robust transferability of text embedding, we take inspirations from CoOp (Zhou et al. 2022a), and add the learnable prompt to original class embeddings. Concretely, the original text labels are first transformed into class tokens through CLIP tokenizer, i.e., $t_{init} = Tokenizer(Label)$, where $Label$ is the discrete text label, e.g., fighting, shooting, road accident, etc. Then we concatenate t_{init} with the learnable prompt $\{c_1, ..., c_l\}$ that contains l context tokens to form a complete sentence token, thus the input of text encoder is presented as follows:

$$t_p = \{c_1, ..., t_{init}, ..., c_l\} \tag{5}$$

here we place the class token at the middle of a sequence. Then this sequence token is added to the positional embedding to obtain positional information, and finally, the text encoder of CLIP takes as input t_p and generates class embedding $t_{out} \in \mathbb{R}^d$.

Anomaly-Focus Visual Prompt. In order to further improve the representation ability of text labels for abnormal events, we investigate how to use visual contexts to refine the class embedding, since visual contexts can make the succinct text labels more accurate. To this end, we propose an anomaly-focus visual prompt, which focuses on the visual embeddings in abnormal segments, and aggregate these embeddings as the video-level prompt for class embeddings. We first use the anomaly confidence A obtained from C-Branch as the anomaly attention, then compute the video-level prompt by the dot product of anomaly attention and video feature X, which is presented as follows,

$$V = Norm\left(A^{\top}X\right) \tag{6}$$

where $Norm$ is the normalization, and $V \in \mathbb{R}^d$ is the anomaly-focus visual prompt. We then add V to the class

embedding t_{out} and obtain the final instance-specific class embedding T by a simple FFN layer and a skip connection.

$$T = FFN\left(ADD\left(V, t_{out}\right)\right) + t_{out} \tag{7}$$

where ADD is the element-wise addition. Such a implementation allows class embeddings to extract the related visual context from videos.

With X and T in hands, we calculate the match similarities between all class embeddings and frame-level visual features to obtain the alignment map M.

Objective Function

For C-Branch, we follow previous works (Wu et al. 2020) and use Top-K mechanism to select K high anomaly confidences in both abnormal and normal videos as the video-level predictions. Then we use the binary cross entropy between video-level predictions and ground-truth to compute classification loss $\mathcal{L}_{bce}$.

For A-Branch, we are confronted with new challenges: 1) there is no anomaly confidence; 2) facing multi-classes instead of binary classes. To address this dilemma, we propose the MIL-Align mechanism which is similar to vanilla MIL. Specifically, we consider the align map M since it expresses the similarity between frame-level video features and all class embeddings. For each row, we select top K similarities and compute the average to measure the alignment degree between this video and the current class. Then we can obtain a vector $S = \{s_1, ..., s_m\}$ that represents the similarity between this video and all classes. We hope the video and its paired textual label emit the highest similarity score among others. To achieve this, the multi-class prediction is firstly computed as follows,

$$p_i = \frac{exp\left(s_i/\tau\right)}{\sum_j exp\left(s_j/\tau\right)} \tag{8}$$

where p_i is the prediction with respect to the i^{th} class, and τ refers to the temperature hyper-parameter for scaling. Finally, the alignment loss $\mathcal{L}_{nce}$ can be computed by the cross entropy.

In addition to classification loss $\mathcal{L}_{bce}$ and alignment loss $\mathcal{L}_{nce}$, we also introduce a contrastive loss to slightly push the normal class embedding and other abnormal class embeddings away, here we first calculate cosine similarity between normal class embedding and other abnormal class embeddings, and then compute the contrastive loss $\mathcal{L}_{cts}$ as follows,

$$\mathcal{L}_{cts} = \sum_j max\left(0, \frac{t_n^\top t_{aj}}{\|t_n\|_2 \cdot \|t_{aj}\|_2}\right) \tag{9}$$

where t_n is the normal class embedding, and t_a is abnormal class embeddings.

Overall, the final total objective of VadCLIP is given by:

$$\mathcal{L} = \mathcal{L}_{bce} + \mathcal{L}_{nce} + \lambda \mathcal{L}_{cts} \tag{10}$$

Inference

VadCLIP contains dual branch that enables itself to address both fine-grained and coarse-grained WSVAD tasks. In regard to fine-grained WSVAD, we follow previous works (Wu, Liu, and Liu 2022) and utilize a thresholding strategy on alignment map M to predict anomalous events. In regard to coarse-grained WSVAD, there are two ways to compute the frame-level anomaly degree. The first one is to directly use the anomaly confidences in C-Branch, the second one is to use the alignment map in A-Branch, specifically, subtracting the similarities between videos and the normal class by one is the anomalous degree. Finally, we select the best of these two ways for computing the frame-level anomaly degree.

Experiments

Datasets and Evaluation Metrics

Datasets. We conduct experiments on two popular WSVAD datasets, i.e., UCF-Crime and XD-Violence. Notably, training videos only have video-level labels on both datasets.

Evaluation Metrics. For coarse-grained WSVAD, we follow previous works, and utilize the frame-level Average Precision (AP) for XD-Violence, and frame-level AUC and the AUC of anomaly videos (termed as AnoAUC) for UCF-Crime. For fine-grained WSVAD, we follow the standard evaluation protocol in video action detection and use the mean Average Precision (mAP) values under different intersection over union (IoU) thresholds. In this work, we use IoU thresholds ranging from 0.1 to 0.5 with a stride of 0.1 to compute mAP values. Meanwhile, we also report an average of mAP (AVG). Note that we only compute mAP on the abnormal videos in the test set.

Implementation Details

For network structure, frozen image and text encoders are adopted from pre-trained CLIP (ViT-B/16). FFN is a standard layer from Transformer, and ReLU is replaced with GELU. For hyper-parameters, we set σ in Eq.3 as 1, τ in Eq.8 as 0.07, and the context length l as 20. For window length in LGT-Adapter, we set it as 64 and 8 on XD-Violence and UCF-Crime, respectively. For λ in Eq.10, we set it as 1×10^{-4} and 1×10^{-1} on XD-Violence and UCF-Crime, respectively. For model training, VadCLIP is trained on a single NVIDIA RTX 3090 GPU using PyTorch. We use AdamW as the optimizer with batch size of 64. On XD-Violence, the learning rate and total epoch are set as 2×10^{-5} and 20, respectively, and on UCF-Crime, the learning rate and total epoch are set as 1×10^{-5} and 10, respectively.

Comparison with State-of-the-Art Methods

VadCLIP can simultaneously realize coarse-grained and fine-grained WSVAD, therefore we present the performance of VadCLIP and compare it with several state-of-the-art methods on coarse-grained and fine-grained WSVAD tasks. For the sake of fairness, *all comparison methods use the same visual features extracted from CLIP as VadCLIP.*

Coarse-grained WSVAD Results. We show comparison results in Tables 1 and 2. Here Ju et al. (Ju et al. 2022) is a CLIP-based work for action recognition, which is significantly inferior to our method. Such results demonstrate

challenges on WSVAD task, and also show the strength of our method with respect to Ju et al. (Ju et al. 2022) for the specific WSVAD task. Besides, we found that VadCLIP significantly outperforms both semi-supervised methods and classification-based weakly supervised methods on two commonly-used benchmarks and across all evaluation metrics. More precisely, VadCLIP attains 84.51% AP and 82.08% AUC on XD-Violence and UCF-Crime, respectively, a new state-of-the-art on both datasets. By comparison, VadCLIP achieves an absolute gain of 2.3% and 2.1% in terms of AP over the best competitors CLIP-TSA (Joo et al. 2023) and DMU (Zhou, Yu, and Yang 2023) on XD-Violence, and on UCF-Crime, VadCLIP also outperforms them by 0.4% and 1.3% in terms of AUC. More importantly, among all comparison methods, AVVD (Wu, Liu, and Liu 2022) uses fine-grained class labels exactly, and it only achieves 78.10% AP and 82.45% AUC on XD-Violence and UCF-Crime, respectively, which lags behind VadCLIP by a large margin. Such a result shows simply using fine-grained labels cannot lead to performance gains, since excessive inputs of label increases the difficulty of binary classification. The performance advantage of VadCLIP is partially attributable to the vision-language associations, since all comparison baselines use the same visual features as VadCLIP.

Fine-grained WSVAD Results. For fine-grained WSVAD task, we compare VadCLIP with previous works AVVD and Sultani et al. (Sultani, Chen, and Shah 2018; Wu, Liu, and Liu 2022) in Tables 3 and 4. Here AVVD is the first work to propose the fine-grained WSVAD, and we reimplement it with visual features of CLIP, then we also finetune Sultani et al. based on the setup in AVVD for adapting fine-grained WSVAD. As we can see, the fine-grained WSVAD is a more challenging task with respect to coarse-fined WSVAD since the former needs to consider both multi-category classification accuracy and detection segment continuity. On this task, VadCLIP is also clearly superior to these excellent comparison methods on both XD-Violence and UCF-Crime datasets. For instance, On XD-Violence, VadCLIP achieves a performance improvement of 13.1% and 4.5% in terms of AVG compared to Sultani et al. and AVVD.

Ablation Studies

Extensive ablations are carried out on XD-Violence dataset. Here we choose the similarity map to compute the frame-level anomaly degree for coarse-grained WSVAD.

Effectiveness of LGT-Adapter. As shown in Table 5, firstly, without the assistance of LGT-Adapter for temporal modeling, the baseline model only achieves 72.22% AP and 15.64% AVG, this results in a considerably drop of 12.3% AP and 9.1% AVG. Secondly, only using global transformer encoder layer, local transformer encoder layer or GCN layer gets clear performance boosts, especially in terms of AP, which convincingly indicates transformer encoder and GCN both can efficiently capture temporal dependencies by means of the self-attention mechanism across video frames. Thirdly, the combination of global transformer encoder and GCN yields the slightly improved performance

Category	Method	AP(%)
Semi	SVM baseline	50.80
	OCSVM (1999)	28.63
	Hasan et al. (2016)	31.25
Weak	Ju et al. (2022)	76.57
	Sultani et al. (2018)	75.18
	Wu et al. (2020)	80.00
	RTFM (2021)	78.27
	AVVD (2022)	78.10
	DMU (2023)	82.41
	CLIP-TSA (2023)	82.17
	VadCLIP (Ours)	**84.51**

Table 1: Coarse-grained comparisons on XD-Violence.

Method	AUC(%)	Ano-AUC(%)
SVM baseline	50.10	50.00
OCSVM (1999)	63.20	51.06
Hasan et al. (2016)	51.20	39.43
Ju et al. (2022)	84.72	62.60
Sultani et al. (2018)	84.14	63.29
Wu et al. (2020)	84.57	62.21
AVVD (2022)	82.45	60.27
RTFM (2021)	85.66	63.86
DMU (2023)	86.75	68.62
UMIL (2023)	86.75	68.68
CLIP-TSA (2023)	87.58	N/A
VadCLIP (Ours)	**88.02**	**70.23**

Table 2: Coarse-grained comparisons on UCF-Crime.

Method	mAP@IOU(%)					
	0.1	0.2	0.3	0.4	0.5	AVG
Random	1.82	0.92	0.48	0.23	0.09	0.71
Sultani et al.	22.72	15.57	9.98	6.20	3.78	11.65
AVVD	30.51	25.75	20.18	14.83	9.79	20.21
VadCLIP	**37.03**	**30.84**	**23.38**	**17.90**	**14.31**	**24.70**

Table 3: Fine-grained comparisons on XD-Violence.

Method	mAP@IOU(%)					
	0.1	0.2	0.3	0.4	0.5	AVG
Random	0.21	0.14	0.04	0.02	0.01	0.08
Sultani et al.	5.73	4.41	2.69	1.93	1.44	3.24
AVVD	10.27	7.01	6.25	3.42	**3.29**	6.05
VadCLIP	**11.72**	**7.83**	**6.40**	**4.53**	2.93	**6.68**

Table 4: Fine-grained comparisons on UCF-Crime.

in terms of AP (+0.4%) over the combination of local transformer encoder and GCN, while the latter achieves significantly better performance in terms of AVG (+3.9%). We also attempt a combination of local Transformer encoder and global Transformer encoder, which results in significant performance degradation in terms of AP listed in the 5^{th} row. The possible reason is that, compared to Transformer, GCN can be regarded as a lightweight variant, and fewer parameters prevent learned knowledge of CLIP from being affected during the transfer process. Therefore, local transformer encoder and GCN are the optimum combination, which can

Method	AP(%)	AVG(%)
Baseline (w/o temporal modeling)	72.22	15.64
Global TF-Encoder	82.54	16.76
Local TF-Encoder	81.18	18.41
Only GCN	81.56	23.31
Local TF-Encoder+ Global TF-Encoder	79.91	19.78
Global TF-Encoder+GCN	**84.87**	20.84
LGT-Adapter	84.51	**24.70**

Table 5: Effectiveness of LGT-Adapter.

C-Branch	A-Branch	L-Prompt	V-Prompt	AP(%)
√				80.53
	√			68.15
√	√			75.03
√	√	√		78.27
√	√		√	82.35
√	√	√	√	**84.51**

Table 6: Effectiveness of dual branch.

	AP(%)	AVG(%)
Hand-crafted Prompt	81.06	22.46
Learnable-Prompt	**84.51**	**24.70**
Average-Frame Visual Prompt	81.34	21.57
Anomaly-Focus Visual Prompt	**84.51**	**24.70**

Table 7: Effectiveness of prompt.

capture different range temporal dependencies.

Effectiveness of Dual Branch. As shown in Table 6, our method with only C-Branch belongs to the classification-based paradigm, and can compete current state-of-the-art methods on XD-Violence. On the other hand, our method with only A-Branch achieves unsatisfactory performance in terms of AP since it is mainly focus on fine-grained WSVAD. With the assistance of coarse-grained classification on feature optimization in C-Branch, A-Branch obtains a leap of about 7% AP improvement. By further adding the learnable prompt and visual prompt that are ad-hoc designs in A-Branch, we notice that a consistent performance improvement can be achieved, leading to a new state-of-the-art. These results clearly show dual branch that contains coarse-grained classification paradigm and fine-grained alignment paradigm can boost the performance by leveraging the complementary of different granularity.

Effectiveness of Prompt. As shown in Table 7, using hand-crafted prompt results in a drop of 3.5% AP and 2.2% AVG, demonstrating that the learnable prompt has better potential for adapting pre-trained knowledge from the large language-vision model to WSVAD task. Furthermore, simply using the average of frame-level features in visual prompt (Ni et al. 2022) produces a drop of 3.2% AP and 3.1% AVG, such results show focusing on abnormal snippets in the video can support VadCLIP to obtain more accurate instance-specific text representations, which boosts the ability of video-language alignment that is useful for WSVAD task.

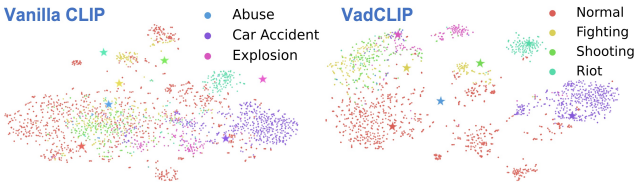

Figure 3: t-SNE visualizations for XD-Violence. Left: Raw CLIP features; Right: VadCLIP features.

Figure 4: Qualitative results of coarse-grained WSVAD.

Qualitative Analyses

Feature Discrimination Visualization. We visualize the feature distribution by using t-SNE for XD-Violence, and present results in Figure 3, where star icons denote textual label features. As we can see, although CLIP has learned generalized capacities based on image-text pairs, such capacities still cannot allow it to effectively distinguish different categories for WSVAD due to intrinsic problems on WSVAD task. After specialized optimization by VadCLIP, these visual features have more distinguishable boundaries and also surround the corresponding text class features.

Coarse-grained Qualitative Visualization. We illustrate the qualitative visualizations of coarse-grained WSVAD in Figure 4, where the blue curves represent the anomaly prediction, and the pink regions correspond to the ground-truth abnormal temporal location. As we can see, VadCLIP precisely detects abnormal region of different categories on two benchmarks, meanwhile, it also produces considerably low anomaly predictions on normal videos.

Conclusion

In this work, we propose a new paradigm named VadCLIP for weakly supervised video anomaly detection. To efficiently adapt the pre-trained knowledge and vision-language associations from frozen CLIP to WSVAD task, we first devise a LGT-Adapter to enhance the ability of temporal modeling, and then we design a series of prompt mechanisms to improve the adaptation of general knowledge to the specific task. Finally we introduce the MIL-Align operation for facilitating the optimization of vision-language alignment under weak supervision. We empirically verify the effectiveness of VadCLIP through state-of-the-art performance and sufficient ablations on two WSVAD benchmarks. In future, we will continue to explore vision-language pre-trained knowledge and further devote to open-set VAD task.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62306240, U23B2013, U19B2037, 62301432, 62101453), China Postdoctoral Science Foundation (No. 2023TQ0272), National Key R&D Program of China (No.2020AAA0106900), Shaanxi Provincial Key R&D Program (No.2021KWZ-03), Natural Science Basic Research Program of Shaanxi (No. 2021JCW-03, 2023-JC-QN-0685), and the Fundamental Research Funds for the Central Universities (No. D5000220431).

References

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chen, F.-L.; Zhang, D.-Z.; Han, M.-L.; Chen, X.-Y.; Shi, J.; Xu, S.; and Xu, B. 2023a. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1): 38–56.

Chen, Y.; Liu, Z.; Zhang, B.; Fok, W.; Qi, X.; and Wu, Y.-C. 2023b. MGFN: Magnitude-Contrastive Glance-and-Focus Network for Weakly-Supervised Video Anomaly Detection. volume 37, 387–395.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Feng, J.-C.; Hong, F.-T.; and Zheng, W.-S. 2021. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14009–14018.

Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A. K.; and Davis, L. S. 2016. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 733–742.

Huang, C.; Liu, C.; Wen, J.; Wu, L.; Xu, Y.; Jiang, Q.; and Wang, Y. 2022. Weakly Supervised Video Anomaly Detection via Self-Guided Temporal Discriminative Transformer. *IEEE Transactions on Cybernetics*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916. PMLR.

Joo, H. K.; Vo, K.; Yamazaki, K.; and Le, N. 2023. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, 3230–3234. IEEE.

Ju, C.; Han, T.; Zheng, K.; Zhang, Y.; and Xie, W. 2022. Prompting visual-language models for efficient video understanding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, 105–124. Springer.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 5583–5594. PMLR.

Li, S.; Liu, F.; and Jiao, L. 2022. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1395–1403.

Lin, Z.; Geng, S.; Zhang, R.; Gao, P.; de Melo, G.; Wang, X.; Dai, J.; Qiao, Y.; and Li, H. 2022. Frozen clip models are efficient video learners. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, 388–404. Springer.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2022. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304.

Lv, H.; Yue, Z.; Sun, Q.; Luo, B.; Cui, Z.; and Zhang, H. 2023. Unbiased Multiple Instance Learning for Weakly Supervised Video Anomaly Detection. *arXiv preprint arXiv:2303.12369*.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Nag, S.; Zhu, X.; Song, Y.-Z.; and Xiang, T. 2022. Zero-shot temporal action detection via vision-language prompting. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, 681–697. Springer.

Ni, B.; Peng, H.; Chen, M.; Zhang, S.; Meng, G.; Fu, J.; Xiang, S.; and Ling, H. 2022. Expanding language-image pretrained models for general video recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, 1–18. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; and Lu, J. 2022. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18082–18091.

Schölkopf, B.; Williamson, R. C.; Smola, A.; Shawe-Taylor, J.; and Platt, J. 1999. Support vector method for novelty detection. *Advances in neural information processing systems*, 12.

Sultani, W.; Chen, C.; and Shah, M. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of*

the IEEE conference on computer vision and pattern recognition, 6479–6488.

Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J. W.; and Carneiro, G. 2021. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4975–4986.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Wang, M.; Xing, J.; and Liu, Y. 2021. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Wu, J.; Zhang, W.; Li, G.; Wu, W.; Tan, X.; Li, Y.; Ding, E.; and Lin, L. 2021. Weakly-supervised spatio-temporal anomaly detection in surveillance video. *arXiv preprint arXiv:2108.03825*.

Wu, P.; and Liu, J. 2021. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, 30: 3513–3527.

Wu, P.; Liu, J.; He, X.; Peng, Y.; Wang, P.; and Zhang, Y. 2023. Towards Video Anomaly Retrieval from Video Anomaly Detection: New Benchmarks and Model. *arXiv preprint arXiv:2307.12545*.

Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; and Yang, Z. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, 322–339. Springer.

Wu, P.; Liu, X.; and Liu, J. 2022. Weakly supervised audio-visual violence detection. *IEEE Transactions on Multimedia*, 1674–1685.

Yu, W.; Liu, Y.; Hua, W.; Jiang, D.; Ren, B.; and Bai, X. 2023. Turning a CLIP Model into a Scene Text Detector. *arXiv preprint arXiv:2302.14338*.

Zaheer, M. Z.; Mahmood, A.; Astrid, M.; and Lee, S.-I. 2020. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, 358–376. Springer.

Zhong, J.-X.; Li, N.; Kong, W.; Liu, S.; Li, T. H.; and Li, G. 2019. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1237–1246.

Zhou, H.; Yu, J.; and Yang, W. 2023. Dual Memory Units with Uncertainty Regulation for Weakly Supervised Video Anomaly Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022b. Detecting twenty-thousand classes using image-level supervision. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, 350–368. Springer.

Zhou, Z.; Zhang, B.; Lei, Y.; Liu, L.; and Liu, Y. 2022c. ZegCLIP: Towards Adapting CLIP for Zero-shot Semantic Segmentation. *arXiv preprint arXiv:2212.03588*.