CLIP-TSA: CLIP-Assisted Temporal Self-Attention for Weakly-Supervised Video Anomaly Detection

Hyekang Kevin Joo *hkjoo@cs.umd.edu*
Department of Computer Science
University of Maryland
College Park, MD

Khoa Vo *khoavoho@uark.edu*
Department of Computer Science & Computer Engineering
University of Arkansas
Fayetteville, AR

David W. Jacobs *dwj@cs.umd.edu*
Department of Computer Science
University of Maryland
College Park, MD

Ngan Le *thile@uark.edu*
Department of Computer Science & Computer Engineering
University of Arkansas
Fayetteville, AR

Abstract

Video anomaly detection (VAD) – commonly formulated as a multiple-instance learning problem in a weakly-supervised manner due to its labor-intensive nature – is a challenging problem in video surveillance where the frames of anomaly need to be localized in an untrimmed video. In this paper, we first propose to utilize the ViT-encoded visual features from CLIP, in contrast with the conventional C3D or I3D features in the domain, to efficiently extract discriminative representations in the novel technique. We then model long- and short-range temporal dependencies and nominate the snippets of interest by leveraging our proposed Temporal Self-Attention (TSA). The ablation study conducted on each component confirms its effectiveness in the problem, and the extensive experiments show that our proposed CLIP-TSA outperforms the existing state-of-the-art (SOTA) methods by a large margin on two commonly-used benchmark datasets in the VAD problem (UCF-Crime and ShanghaiTech Campus). The source code will be made publicly available upon acceptance.

1 Introduction

Video understanding is a growing field and a subject of intense research that requires analysis of both spatial and temporal information, *e.g.*, action recognition (Pareek & Thakkar, 2021; Vu et al., 2021a;b; Sun et al., 2022; Vu et al., 2022), action detection (Xu et al., 2020; Zeng et al., 2019; Vo et al., 2021a; Zhang et al., 2022), video captioning (Lei et al., 2020a; Dai et al., 2019; Yamazaki et al., 2022), video retrieval (Snoek et al., 2009; Gabeur et al., 2020; Wang et al., 2021; Wray et al., 2021). One of the challenging problems in video understanding is video anomaly detection (VAD), which is the task of localizing anomalous events in a given video. VAD is an area of research that has several years of history, and it has been gaining more attraction in recent years (Hasan et al., 2016; Sultani et al., 2018; Wu & Liu, 2021). Generally, there are three main paradigms in VAD, namely, fully-supervised (Liu & Ma, 2019), unsupervised (Gong et al., 2019; Zaheer et al., 2022), and weakly-supervised (Thakare et al., 2022; Sultani et al., 2018;

Table 1: Comparison among multiple VAD approaches.

Supervision	Normal	Abnormal	Annotation	Approaches
Fully-Supervised	✓	✓	Frame-Level	Liu & Ma (2019)
Weakly-Supervised	✓	✓	Video-Level	Sultani et al. (2018); Thakare et al. (2022); Purwanto et al. (2021); Tian et al. (2021); Zaheer et al. (2020); Sapkota & Yu (2022)
Unsupervised	✓	✗	✗	Hasan et al. (2016); Gao et al. (2021) Wang & Cherian (2019); Lu et al. (2013) Zaheer et al. (2022); Wu & Liu (2021)

Tian et al., 2021). While it generally yields high performance, the supervised VAD requires fine-grained anomaly labels (*i.e.,* frame-level normal/abnormal annotations in the training data). However, the problem has traditionally been difficult to solve in a fully supervised manner due to the labor-expensive nature of data collection. In general, anomaly detection annotation requires the annotator to localize and label anomalies in a video, or a large set of sequential frames. Unfortunately, this is a very strenuous labor for the annotator because, as anomalies can happen at any moment, almost all of the frames need to be observed carefully, leading to massive time consumption. Because of its time-consuming and labor-intensive nature, collecting a fully-annotated large-scale dataset is a difficult task for the supervised VAD. In unsupervised VAD learning, one-class classification (OCC) problem (Zaheer et al., 2020) is a common approach, in which the model is trained on only normal class samples with the assumption that unseen abnormal videos have high reconstruction errors. However, the performance of unsupervised VAD is usually poor because of its lack of prior knowledge of abnormality as well as its inability to capture all normality variations (Chandola et al., 2009). Compared to both unsupervised and supervised VAD, the weakly-supervised VAD is considered the most practical approach by many for VAD because of its competitive performance and annotation efficiency by employing video-level labels to reduce the cost of manual fine-grained annotations (Zaheer et al., 2020; Zhong et al., 2019). The comparison among various VAD approaches is shown in Table 1.

In the weakly-supervised VAD task, there exist two fundamental problems. First, anomalous-labeled frames tend to be dominated by normal-labeled frames, as the videos are untrimmed and there is no strict length requirement for the anomalies in the video. Second, the anomaly may not necessarily stand out against normality. As a result, it occasionally becomes challenging to localize anomaly snippets. In order to combat the issues, Sultani et al. (2018); Tian et al. (2021); Wu et al. (2020); Zhang et al. (2019); Zhu & Newsam (2019) have attempted to tackle the problem in multiple instance learning (MIL) frameworks, which treat a video as a bag containing multiple instances, each instance being a video snippet. A video is labeled as anomalous if any of its snippets are anomalous, and normal if all of its snippets are normal. Following the MIL framework, anomalous-labeled videos belong to the positive bag and normal-labeled videos belong to the negative bag.

Furthermore, the existing approaches encode the extracted visual content by applying a backbone, *e.g.*, C3D (Ji et al., 2013), I3D (Carreira & Zisserman, 2017), 2Stream (Simonyan & Zisserman, 2014), which are pre-trained on action recognition tasks. Different from the action recognition problem, VAD depends on discriminative representations that clearly represent the events in a scene. Thus, those existing backbones, C3D, I3D, and 2Stream, are not suitable because of the domain gap (Liu & Ma, 2019). To address such limitation, we leverage the success of the recent "vision-language" works (Patashnik et al., 2021; Yang & Zou, 2022; Vo et al., 2022; Yamazaki et al., 2023), which have proved the effectiveness of feature representation learned via Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021). CLIP consists of two networks, a vision encoder and a text encoder, which are trained on 400 million text-image pairs collected from a variety of publicly available sources on the Internet. Given a set of words and an image, CLIP can estimate the semantic similarity between them. We thus leverage CLIP as a visual feature extractor. Furthermore, the existing MIL-based weakly-supervised VAD approaches are limited in dealing with an arbitrary number of abnormal snippets in an abnormal video. To address such an issue, we are inspired by the differentiable top-K operator (Cordonnier et al., 2021) and introduce a novel technique, termed top-κ function, that localizes κ snippets of interest in the video with differentiable hard attention in the similar MIL setting to demonstrate its effectiveness and applicability to the traditional, popular setting. Furthermore, we introduce the Temporal Self-Attention (TSA) Mechanism, which aims to generate the reweighed attention feature by measuring the abnormal degree of snippets. Our proposed CLIP-TSA follows the MIL framework and consists of three components corresponding to (i) Feature

Encoding by CLIP; (ii) Modeling snippet coherency in the temporal dimension with our Temporal Self-Attention and (iii) Localizing anomalous snippets with Difference Maximization Trainer. As the real-world anomalies are diverse, in order to show the applicability of our proposed method to multiple environments, we run experiments on three different datasets commonly used for the VAD evaluation: UCF-Crime (Sultani et al., 2018), ShanghaiTech Campus (Liu et al., 2018), and XD-Violence (Wu et al., 2020). In addition, we conduct an ablation study on the effectiveness of our proposed method. Throughout the paper, the term *abnormal* and *anomaly* will be used interchangeably.

Our contributions are summarized as follows:

- We propose a Temporal Self-Attention (TSA) mechanism that is applicable to the Weakly-Supervised VAD problems and acquires anomaly likelihood scores for video snippets.
- We leverage CLIP, which uses a ViT as a backbone for visual features, to introduce 1) novel usage of CLIP features and 2) novel type of contextual representation in analyzing videos consisting of abnormal actions.
- We empirically validate the usefulness of our proposed method by showing that, to the best of our knowledge, it achieves superior performance to *all* of the current SOTA methods benchmarked on UCF-Crime and ShanghaiTech Campus datasets under any type of supervision setting. As for the XD dataset, it beats the performance of all the SOTAs trained without auditory features for a fair comparison.

2 Related Work

2.1 Unsupervised VAD

Unsupervised anomaly detection approaches do not require labeled data during training. In such approaches, the usual patterns with only normal training samples are first encoded and distinctive encoded patterns are detected as anomalies. While the early anomaly detection methods (Antić & Ommer, 2011; Basharat et al., 2008; Li et al., 2013; Saligrama & Chen, 2012; Wu et al., 2010) mainly depend on the handcrafted features, the recent approaches primarily make use of the merits of deep neural networks (DNNs) (Doshi & Yilmaz, 2020; Hasan et al., 2016; Ionescu et al., 2019; Lu et al., 2013; Ramachandra et al., 2020; Wang & Cherian, 2019; Zaheer et al., 2022). In such approaches, reconstruction error is utilized to identify anomalies with the assumption that anomalous events are often reconstructed poorly. For example, Hasan et al. (2016) used autoencoders as feature extractors to model the subsequent frame and estimated abnormality by reconstruction error. Later, Wang & Cherian (2019) assumed that anomalous events will cause a big difference between past and future frames and proposed spatiotemporal autoencoder with combinations of CNNs and LSTMs (Hochreiter & Schmidhuber, 1997). With a similar assumption on reconstruction errors as an abnormality recognizer, Feng et al. (2021); Liu et al. (2018); Park et al. (2020) adopted generative networks to synthesize or predict future frames. Furthermore, Doshi & Yilmaz (2020) proposed a hybrid use of DNNs and statistical kNN (k nearest neighbor) decision approach for finding video anomalies. Siamese network was employed to detect anomaly (Ramachandra et al., 2020) by learning a distance function between a pair of video patches.

Historically, the performance of unsupervised anomaly detection problems generally lagged behind that of weakly-supervised anomaly detection by a large margin because the model in an unsupervised setting significantly lacks the prior knowledge of anomaly needed for differentiation between normality and anomaly.

2.2 Weakly-supervised VAD

Weakly-supervised VAD methods (Lv et al., 2021; Purwanto et al., 2021; Sapkota & Yu, 2022; Sultani et al., 2018; Thakare et al., 2022; Tian et al., 2021; Wu et al., 2020; Zaheer et al., 2020; Zhang et al., 2019; Zhong et al., 2019; Zhu & Newsam, 2019) rely on the video-level labels. In this setup, a normal-labeled video contains all normal events, whereas an anomaly-labeled video contains both normal and anomalous events without any temporal information about starting and ending of anomalous events. Weakly-supervised VAD problem has been generally regarded as an MIL problem (He et al., 2017; Huo et al., 2012; Sultani et al., 2018) as the videos are labeled at bag-level (*i.e.,* video-level), with the anomaly-labeled video regarded as a positive bag and the normal-labeled video regarded as a negative bag. Particularly since Sultani et al. (2018) proposed a weakly-supervised framework to detect anomalies on UCF-Crime, in which both normal and abnormal samples annotated at video-level are included in both train and test sets, this research in the weakly-supervised setting has grown and gained significant popularity. Since then, more weakly-labeled VAD datasets, primarily for use in a weakly-supervised setting, have been introduced (Liu et al., 2018;

Wu et al., 2020). In such approaches, the feature extractor can be trained or utilized by pre-trained models. While Zhong et al. (2019); Zhu & Newsam (2019) trained both the feature encoder and classifier simultaneously, Sultani et al. (2018); Tian et al. (2021); Zhang et al. (2019) utilized pre-trained models such as C3D (Ji et al., 2013), I3D (Carreira & Zisserman, 2017), 2Stream (Simonyan & Zisserman, 2014), and SlowFast (Feichtenhofer et al., 2019) as feature extractors and trained the classifier only.

2.3 Vision-Language Pre-trained Models

Vision-language pre-trained model (VLPM) aims to learn the semantic correspondence between different modalities (*i.e.,* video and text) by pre-training the model on a large-scale dataset of video/image-text pairs. Specifically, the model mines the associations between objects or actions in the video and objects or actions in the text. Standard vision-language tasks include video captioning (Krishna et al., 2017; Pasunuru & Bansal, 2017; Vo et al., 2022; Yamazaki et al., 2022), text-to-video retrieval (Hendricks et al., 2018; Rohrbach et al., 2015), and video question answering (Girdhar & Ramanan, 2020; Lei et al., 2020b). Generally, VLPM can be divided into two categories: single-stream and dual-stream. The former uses a single transformer to model both image/video and text representations in a unified framework. Both image/video and text embeddings are concatenated into one feature. This category includes VisualBERT (Li et al., 2019), UNIMO (Li et al., 2020b), OSCAR (Li et al., 2020c) , UNICODER (Li et al., 2020a), and UNITER (Chen et al., 2020b). The latter one separately encodes image/video and text with a decoupled encoder. This category includes LXMERT (Tan & Bansal, 2019), ViLBERT (Lu et al., 2019), CLIP (Radford et al., 2021), and DeCLIP (Li et al., 2021). VisualBERT, ViLBERT, OSCAR, UNICODER, UNITER, and LXMERT use masked token tasks and are based on Language Modeling, whereas UNIMO, CLIP, and DeCLIP are trained on contrastive learning. Because of simplicity, flexibility, and low computation cost, we adopt the frozen self-supervised vision-language model CLIP, a dual-stream architecture and contrastive learning in this paper.

2.4 Attention Mechanism

Attention models have a long history. In 2015, Bahdanau et al. (2015) introduced one of the first soft attention models capable of attending to all the source words and attempted to solve the machine translation task without the traditional encoder-decoder models (*e.g.,* RNN, LSTM), which were common approach for the problem at the time (Cho et al., 2014; Sutskever et al., 2014). Shortly afterward, Xu et al. (2015) introduced a hard stochastic attention mechanism that is able to compute the relative importance of the source words with respect to the output words, combating the huge expense of computation required for soft attention. Because hard attention only places attention locally, the mechanism is generally computationally less expensive than the soft attention mechanism, which observes all hidden states (Luong et al., 2015). In general, while soft attention models are trainable end-to-end, hard attention models are not differentiable and require reinforcement learning (Xu et al., 2015). Today, many variations of attention mechanisms have been introduced. For example, Luong et al. (2015) proposed a local attention mechanism similar to hard attention, but is differentiable. In 2017, Vaswani et al. (2017) introduced a neural machine translation (NMT) architecture named Transformer that is designed with only fully connected layers and attention by leveraging the self-attention mechanism. Recently, Vo et al. (2021b; 2022) inherited the merits from both soft attention models and hard attention models and proposed adaptive attention models. Despite its original application in NMT, Transformer has been gaining great attraction, and its usage has expanded widely, including computer vision.

3 Proposed Method

3.1 Problem Setup

In weakly-supervised VAD, videos in the training set are only labeled at video-level. Let there be a set of weakly-labeled training videos $S = \{\mathcal{X}^{(k)}, y^{(k)}\}_{k=1}^{|S|}$, where a video $\mathcal{X}^{(k)} \in \mathbb{R}^{N_k \times W \times H}$ is a sequence of N_k frames that are W pixels wide and H pixels high, and $y^{(k)} = \{0, 1\}$ is the video-level label of video $\mathcal{X}^{(k)}$ in terms of anomaly (*i.e.,* 1 if the video contains anomaly; 0 otherwise).

Given a video $\mathcal{X}^{(k)} \in \mathbb{R}^{N_k \times W \times H}$ consisting of N_k frames, *i.e.,* $\mathcal{X}^{(k)} = \{x_j\}|_{j=1}^{N_k}$, we first divide $\mathcal{X}^{(k)}$ into a set of δ-frame snippets $\{s_i\}_{i=1}^{\lceil \frac{N_k}{\delta} \rceil}$. Feature representation of each snippet is extracted by applying a vision-language model into the middle frame. In this work, CLIP is chosen as a vision-language model; however, it can be substituted by any

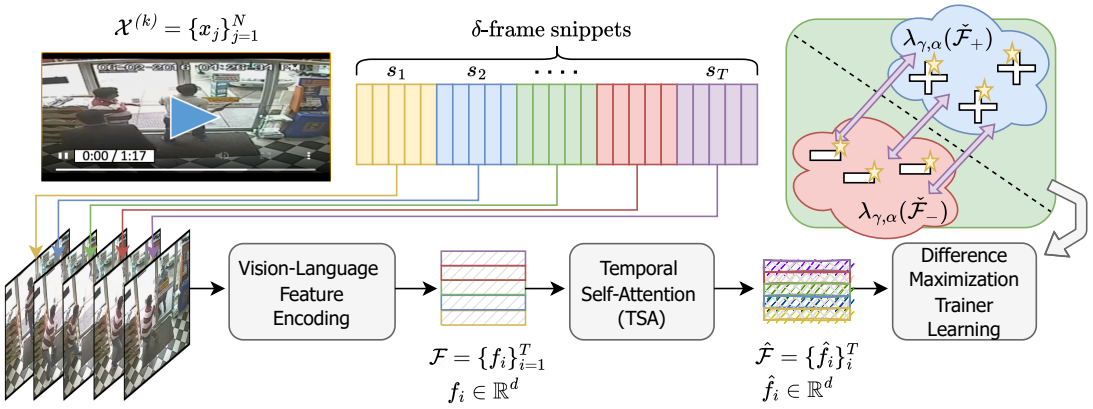

Figure 1: Overall flowchart of our proposed CLIP-TSA in train time. Given a video $\mathcal{X}$ consisting of N frames (*i.e.,* $\mathcal{X} = \{x_j\}_j^N$), we first divide into a set of δ-frame snippets $\{s_i\}_i^T$. Each δ-frame snippet s_i is represented by a vision-language feature $f_i \in \mathbb{R}^d$. Then, the features $\mathcal{F} = \{f_i\}_i^T$, where $f_i \in \mathbb{R}^d$, are resized into one uniform length T to allow batch training by following Eq. 1. Our proposed TSA is then applied onto the resized features to obtain anomaly attention feature $\hat{\mathcal{F}} = \{\hat{f}_i\}_i^T$, where $\hat{f}_i \in \mathbb{R}^d$. The anomaly attention feature $\hat{\mathcal{F}}$ is used for: 1) producing an anomaly likelihood score U using the score classifier C; 2) optimizing the model by employing the difference maximization trainer technique $\upsilon_{\gamma,\alpha}$ using the feature magnitude.

vision-language model as introduced in Section 2.3. Thus, each δ-frame snippet s_i is represented by a vision-language feature $f_i \in \mathbb{R}^d$ and the video $\mathcal{X}^{(k)}$ is represented by a set of video feature vectors $\mathcal{F}_k = \{f_i\}|_{i=1}^{T_k}$, where $\mathcal{F}_k \in \mathbb{R}^{T_k \times d}$ and T_k is the number snippets of $\mathcal{X}^{(k)}$.

CLIP-TSA is trained using a mini-batch; thus, it introduces an issue caused by the difference in video embedding feature length T between samples in the mini-batch. To address this issue, we normalize video feature length by following the approach introduced by (Sultani et al., 2018). Given two videos $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$, their corresponding sets of video feature vectors are $\mathcal{F}_1 = \{f_i\}|_{i=1}^{T_1}$ and $\mathcal{F}_2 = \{f_i\}|_{i=1}^{T_2}$, respectively, where $T_1 \neq T_2$. Following their paradigm, both $\mathcal{F}_1$ and $\mathcal{F}_2$ with size T_1 and T_2 are reshaped into the same size of T with Eq. 1, where $\lfloor g \rfloor = \lfloor \frac{T_1}{T} \rfloor$ and $\lfloor g \rfloor = \lfloor \frac{T_2}{T} \rfloor$ for videos $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$, respectively:

$$\mathcal{F} = \{f_{i'}\}|_{i'=1}^{T} = \frac{1}{\lfloor g \rfloor} \sum_{i=\lfloor g \times (i'-1) \rfloor}^{\lfloor g \times i' \rfloor} f_i \tag{1}$$

Using this technique, we can handle an arbitrary length of videos, allowing for training the features in batches. However, in test time, as the videos are evaluated one at a time, the features do not go through the normalization process in test time. In this paper, we assume that, in training time, the input features $\mathcal{F}$ come post-normalized into the uniform shape in temporal dimension T for batch training.

Our proposed anomaly detection CLIP-TSA's pipeline is portrayed in Figure 1 with three main components *i.e.,* (i) Feature Encoding, (ii) Temporal Self Attention (TSA), and (iii) Difference Maximization, which are elaborated in the following sections.

3.2 Feature Encoding

CLIP (Radford et al., 2021) is an image-text matching model, and it has recently attained remarkable achievements in various computer vision tasks such as image classification (Cheng et al., 2021), image-text retrieval (Dzabraev et al., 2021), and image generation (Patashnik et al., 2021). Originally, CLIP is trained to match an image with its corresponding natural language descriptions. CLIP consists of two independent encoders respectively for visual and textual features encoding. Given a batch of images and texts, CLIP aims to align their feature in the embedding space with a contrastive loss during the training process. CLIP is comprehensively trained on 400 million image-text pairs

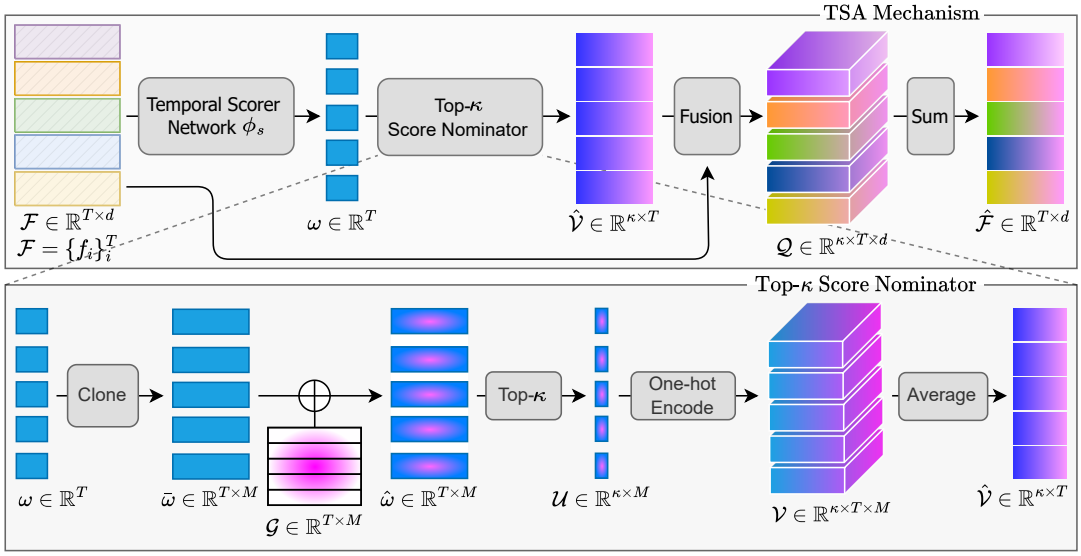

Figure 2: Top: Illustration of our proposed TSA mechanism to model coherency between snippets. The TSA mechanism consists of three components corresponding to (i) temporal scorer network ϕ_s to obtain a relevance score vector ω; (ii) top-κ score nominator to extract the κ most relevant snippets from a video; (iii) fusion process to combine information to provide the model output. The TSA mechanism takes the vision language encoding feature $\mathcal{F} \in \mathbb{R}^{T\times d}$ of T post-normalized snippets (Eq. 1) of a given video as its input and returns the reweighed attention feature $\hat{\mathcal{F}} \in \mathbb{R}^{T\times d}$ as its output. Bottom: Details of top-κ score nominator network, which takes the score vector ω as its input and returns a stack of soft one-hot vectors $\hat{\mathcal{V}}$ as its output.

collected from the Internet. In this work, we leverage CLIP as a feature extractor to obtain a vision-language scene feature. Specifically, we choose the middle frame I_i that represents each snippet s_i. We first encode frame I_i with the pre-trained Vision Transformer (Dosovitskiy et al., 2021) to extract visual feature I_i^f. We then project feature I_i^f onto the visual projection matrix L, which was pre-trained by CLIP to obtain the image embedding $f_i = L \cdot I_i^f$. Thus, the embedding feature $\mathcal{F}_k$ of video $\mathcal{X}$, which consists of T_k snippets $\mathcal{X} = \{s_i\}|_{i=1}^{T_k}$, is defined in Eq. 2b. Finally, we apply the video normalization as in Eq. 1 into the embedding feature to obtain the final embedding feature $\mathcal{F}$ as in Eq. 2c.

$$f_i = L \cdot I_i^f \quad \text{where } f_i \in \mathbb{R}^d \tag{2a}$$
$$\mathcal{F}_k = \{f_i\}|_{i=1}^{T_k} \quad \text{where } \mathcal{F}_k \in \mathbb{R}^{T_k\times d} \tag{2b}$$
$$\mathcal{F} = Norm(\mathcal{F}_k) \quad \text{where } \mathcal{F} \in \mathbb{R}^{T\times d} \tag{2c}$$

3.3 Temporal Self-Attention (TSA)

Our proposed TSA mechanism aims to model the coherency between snippets of a video and select the top-κ most relevant snippets. It contains three modules *i.e.*, (i) temporal scorer network, (ii) top-κ score nominator, and (iii) fusion network, as visualized in Figure 2 and mathematically explained in Algorithm 1.

In TSA, the vision language feature $\mathcal{F} \in \mathbb{R}^{T\times d}$ (from 3.2 Feature Encoding) is first converted into a score vector $\omega \in \mathbb{R}^{T\times 1}$ through a *temporal scorer network* ϕ_s, *i.e.*, $\omega = \phi_s(\mathcal{F})$. This network is meant to be shallow; thus, we choose a multi-layer perceptron (MLP) of 3 layers in this paper. The scores, each of which is representing the snippet s_i, are then passed into the *top-κ score nominator* to extract the κ most relevant snippets from the video. The top-κ score nominator is implemented by the following two steps. First, the scores $\omega \in \mathbb{R}^{T\times 1}$ are cloned M times and the cloned score $\bar{\omega} \in \mathbb{R}^{T\times M}$ is obtained; M represents the number of independent samples of score vector ω to generate for the empirical mean, which is to be used later for computing the expectation with noise-perturbed features.

Throughout the paper, we set M to be 100. Second, Gaussian noise $\mathcal{G} \in \mathbb{R}^{T \times M}$ is applied to the stack of M clones by the following Eq. 3 to produce $\hat{\omega} \in \mathbb{R}^{T \times M}$:

$$\hat{\omega} = \mathcal{G} \oplus \bar{\omega} \qquad \text{where } \oplus \text{ is an element-wise addition} \tag{3}$$

From the Gaussian-perturbed scores $\hat{\omega} \in \mathbb{R}^{T \times M}$, the indices of top-$\kappa$ snippets are selected based on the score magnitude independently across its M dimension to represent the most relevant snippets and are later one-hot encoded into a matrix $\mathcal{V} = \{V_i\}_{i=1}^{M}$, with each $V_i \in \mathbb{R}^{\kappa \times T}$ containing a set of one-hot vectors. More specifically, we guide the network to place the attention on κ magnitudes with the highest values because the Difference Maximization Trainer (See 3.4) trains the anomalous snippets to have a high value and the normal snippets to have a low value. The matrix $\mathcal{V}$ is then averaged across its M dimension to produce a stack of soft one-hot vectors $\hat{\mathcal{V}} \in \mathbb{R}^{\kappa \times T}$. Through the soft one-hot encoding mechanism, the higher amount of attention, or weight, is placed near and at the indices of top-κ scores (*e.g.*, [0, 0, 1, 0] $\rightarrow$ [0, 0.03, 0.95, 0.02]). The top-κ score nominator can be summarized by the pseudocode in Algorithm 2.

Afterwards, the stack of perturbed soft one-hot vectors $\hat{\mathcal{V}} \in \mathbb{R}^{\kappa \times T}$ is transformed into $\tilde{\mathcal{V}} \in \mathbb{R}^{\kappa \times T \times d}$ by making d clones of $\hat{\mathcal{V}}$, and the set of input feature vectors $\mathcal{F} \in \mathbb{R}^{T \times d}$ is transformed into $\tilde{\mathcal{F}} \in \mathbb{R}^{\kappa \times T \times d}$ by making κ clones of $\mathcal{F}$. Next, the matrices $\tilde{\mathcal{V}}$ and $\tilde{\mathcal{F}}$, which carry the reweighed information of snippets and represent the input video features, respectively, are fused together to create a perturbed feature $\mathcal{Q} \in \mathbb{R}^{\kappa \times T \times d}$ that represents the reweighed feature magnitudes of snippets based on the previous computations as follows:

$$\mathcal{Q} = \tilde{\mathcal{V}} \otimes \tilde{\mathcal{F}} \qquad \text{where } \otimes \text{ is an element-wise multiplication} \tag{4}$$

Then, each stack of perturbed feature vectors $Q \in \mathbb{R}^{\kappa \times d}$ within the perturbed feature $\mathcal{Q} = \{Q_i\}_{i=1}^{T}$ is independently summed up across its dimension κ to combine the magnitude information of Q_i into one vector $\hat{f}_i \in \mathbb{R}^d$. This step is akin to the process of reversing the previous one-hot encoding procedure by reducing the one dimension previously expanded for one-hot encoding. The reweighed feature vector, $\hat{f}_i \in \mathbb{R}^d$, which collectively forms $\hat{\mathcal{F}} = \{\hat{f}_i\}|_{i=1}^{T}$, is collectively obtained as the model output from the TSA mechanism σ to represent an anomaly attention feature $\hat{\mathcal{F}} \in \mathbb{R}^{T \times d}$. The pipeline of TSA is described by the pseudocode in Algorithm 1 and illustrated in Figure 2.

Algorithm 1: TSA mechanism σ to produce anomaly attention features $\hat{\mathcal{F}}$

Data: Feature $\mathcal{F} \in \mathbb{R}^{T \times d}$,
Top snippet count $\kappa \in \mathbb{R}^1$
Result: Anomaly attention feature $\hat{\mathcal{F}}$
$\omega \leftarrow \phi_s(\mathcal{F})$ // $\mathbb{R}^{T \times 1}$
$\hat{\mathcal{V}} \leftarrow$ Top-κ Score(M, κ, ω) // Alg.2, $\mathbb{R}^{\kappa \times T}$
$\tilde{\mathcal{V}} \leftarrow$ Make d clones of $\hat{\mathcal{V}}$ // $\mathbb{R}^{\kappa \times T \times d}$
$\tilde{\mathcal{F}} \leftarrow$ Make κ clones of $\mathcal{F}$ // $\mathbb{R}^{\kappa \times T \times d}$
$\mathcal{Q} \leftarrow \tilde{\mathcal{V}} \otimes \tilde{\mathcal{F}}$ // $\mathbb{R}^{\kappa \times T \times d}$
$\hat{\mathcal{F}} \leftarrow$ summation of $\mathcal{Q}$ across dim κ // $\mathbb{R}^{T \times d}$
return $\hat{\mathcal{F}}$ // dim:dimension

Algorithm 2: Top-κ Score function

Data: Sample count M,
Top snippet count κ,
Score vector ω
Result: A stack of soft one-hot vectors $\hat{\mathcal{V}}$
set $\bar{\omega}$ to M clones of ω // $\mathbb{R}^{T \times M}$
set $\mathcal{G}$ to Gaussian noise // $\mathbb{R}^{T \times M}$
$\hat{\omega} \leftarrow \mathcal{G} \oplus \bar{\omega}$ // $\mathbb{R}^{T \times M}$
$\mathcal{U} \leftarrow$ indices of top-κ scores
across dim M in $\hat{\omega}$ // $\mathbb{R}^{\kappa \times M}$
$\mathcal{V} \leftarrow$ one-hot encode κ in $\mathcal{U}$ // $\mathbb{R}^{\kappa \times T \times M}$
$\hat{\mathcal{V}} \leftarrow$ average of $\mathcal{V}$ across dim M // $\mathbb{R}^{\kappa \times T}$
return $\hat{\mathcal{V}}$ // dim:dimension

3.4 Difference Maximization Trainer Learning

Our weakly-supervised VAD model, CLIP-TSA, is set up as an MIL framework, in which the positive bag represents anomaly and the negative bag denotes normality. Following the paradigm, a video, treated as a bag, is labeled a positive bag if it contains at least one snippet of anomaly, while it is labeled a negative bag otherwise. Given a mini-batch of $2 * B$ videos $\{\mathcal{X}^{(k)}\}|_{k=1}^{2*B}$, each video $\mathcal{X}^{(k)}$ is represented by $\mathcal{F}_k = \{f_i\}|_{i=1}^{T}$ obtained by TSA (Section 3.3). Let the input mini-batch be represented by $\mathcal{Z} = \{\mathcal{F}_k\}|_{k}^{2*B} \in \mathbb{R}^{2*B \times T \times d}$, where B, T, and d denote the user-input batch size, normalized time snippet count, and feature dimension, respectively. The actual batch size is dependent on the user-input batch size, following the equation of $2 * B$, because the first half, $\mathcal{Z}_{-} \in \mathbb{R}^{B \times T \times d}$, is loaded with a set of normal bags, and the second half, $\mathcal{Z}_{+} \in \mathbb{R}^{B \times T \times d}$, is loaded with a set of abnormal bags in order within the mini-batch.

After the mini-batch undergoes the phase of TSA, it outputs a set of reweighed normal attention features $\hat{\mathcal{Z}}_- = \{\hat{\mathcal{F}}_k\}|_{k=1}^{B}$ and a set of reweighed anomaly attention features $\hat{\mathcal{Z}}_+ = \{\hat{\mathcal{F}}_k\}|_{k=B}^{2*B}$. The reweighed attention features $\hat{\mathcal{Z}}$ are then passed into a convolutional network module J composed of dilated convolutions (Yu & Koltun, 2016) and non-local block (Wang et al., 2018) to model the long- and short-term relationship between snippets based on the reweighed magnitudes. The resulting stack of convoluted attention features $\check{\mathcal{Z}} = \{\check{\mathcal{F}}_k\}|_{k}^{2*B}$, where $\check{\mathcal{Z}} \in \mathbb{R}^{2*B\times T\times d}$, is then passed into a shallow MLP-based score classifier network C that converts the features into a set of scores $U \in \mathbb{R}^{2*B\times T\times 1}$ to determine the binary anomaly state of feature snippets. The set of scores U is saved as part of a group of returned variables, for use in loss.

Next, each convoluted attention feature $\{\check{\mathcal{F}}_k\}|_{k=1}^{2*B}$ of the batch $\check{\mathcal{Z}}$ undergoes Difference Maximization Trainer (DMT). Leveraging the top-α instance separation idea employed by Li & Vasconcelos (2015); Sultani et al. (2018), we use DMT, represented by $\upsilon_{\gamma,\alpha}$, in this problem to maximize the separation, or difference, between top instances of two contrasting bags, $\check{\mathcal{Z}}_-$ and $\check{\mathcal{Z}}_+$, by first picking out the top-α snippets from each convoluted attention feature $\check{\mathcal{F}}_k$ based on the feature magnitude. This produces a top-α subset $\dot{\mathcal{F}}_k \in \mathbb{R}^{\alpha\times d}$ for each convoluted attention feature $\check{\mathcal{F}}_k \in \mathbb{R}^{T\times d}$. Second, $\dot{\mathcal{F}}_k$ is averaged out across top-α snippets to create one feature vector $\ddot{\mathcal{F}}_k \in \mathbb{R}^d$ that represents the bag. The procedure is explained by Eq. 5 below:

$$\lambda_{\gamma,\alpha}(\check{\mathcal{F}}) = \ddot{\mathcal{F}} = \max_{\Omega_\alpha(\check{\mathcal{F}})\subseteq\{\check{f}_i\}_{i=1}^{T}} \frac{1}{\alpha} \sum_{\check{f}_i\in\Omega(\check{\mathcal{F}})} \check{f}_i \tag{5}$$

In the equation, λ is parameterized by γ, which denotes its dependency on the ability of the convolutional network module J (*i.e.,* representation of $\ddot{\mathcal{F}}$ depends on the top-α positive instances selected with respect to J). In addition, α in Eq. 5 denotes the size of Ω, where Ω represents a subset of α snippets from $\check{\mathcal{F}}$. Each representative vector $\ddot{\mathcal{F}}$ is then normalized to produce $\ddot{F} \in \mathbb{R}^1$.

$$\upsilon_{\gamma,\alpha}(\check{\mathcal{F}}_+, \check{\mathcal{F}}_-) = ||\lambda_{\gamma,\alpha}(\check{\mathcal{F}}_+)|| - ||\lambda_{\gamma,\alpha}(\check{\mathcal{F}}_-)|| \tag{6}$$

The separability is computed as in Eq. 6, where $\check{\mathcal{F}}_- = \{\check{f}_{-,i}\}|_{i}^{T}$ represents a negative bag and $\check{\mathcal{F}}_+ = \{\check{f}_{+,i}\}|_{i}^{T}$ represents a positive bag. More specifically, we leverage the theorem below to maximize the separability of the top-α instances (feature snippets) from each contrasting bag.

Theorem 1 (Li & Vasconcelos, 2015; Tian et al., 2021): Expected Separability. *Let* $\mathbb{E}[\|\check{f}_+\|_2] \geq \mathbb{E}[\|\check{f}_-\|_2]$*, where* $\check{\mathcal{F}}_+$ *has* $\epsilon \in [1, T]$ *abnormal samples and* $(T - \epsilon)$ *normal samples,* $\check{\mathcal{F}}_-$ *has T normal samples, and* $T = |\check{\mathcal{F}}_+| = |\check{\mathcal{F}}_-|$*. Let* $\Upsilon_{\gamma,\alpha}(\cdot)$ *be the random variable from which the separability scores* $\upsilon_{\gamma,\alpha}(.)$ *of Eq. 6 are drawn.*

1. *If* $0 < \alpha < \epsilon$*, then*
$$0 \leq \mathbb{E}[\Upsilon_{\gamma,\alpha}(\check{\mathcal{F}}_+, \check{\mathcal{F}}_-)] \leq \mathbb{E}[\Upsilon_{\gamma,\alpha+1}(\check{\mathcal{F}}_+, \check{\mathcal{F}}_-)] \tag{7}$$
2. *For a finite* ϵ, then
$$\lim_{\alpha\to\infty} \mathbb{E}[\Upsilon_{\gamma,\alpha}(\check{\mathcal{F}}_+, \check{\mathcal{F}}_-)] = 0 \tag{8}$$

In simple terms, the theorem in our setting conveys that, as the number of samples in the top-α snippets of the abnormal video increases – but no greater than ϵ – the separability between the two contrasting bags may be maximized. However, if it exceeds the number, it becomes difficult as the number of negative (normal) samples starts to dominate in both negative and positive bags.

Afterward, to compute the loss, a batch of normalized representative features $\{\ddot{\mathcal{F}}_{normal}\}|_{k=1}^{B}$ and $\{\ddot{\mathcal{F}}_{abnormal}\}|_{k=B}^{2*B}$ are then measured for margins between each other. A batch of margins is then averaged out and used as part of the net loss together with the score-based binary cross-entropy loss computed using the score set U.

3.5 Inference

In test time, the video feature vectors $\mathcal{F}$ that have been extracted with CLIP do not undergo the normalization process to be reshaped into the common size of T because each feature is evaluated at a time. When $\mathcal{F}_k \in \mathbb{R}^{T_k\times d}$ is input into the model in test time, the feature $\mathcal{F}_k$ undergoes the proposed TSA process to produce the reweighed attention features $\hat{\mathcal{F}}_k$. They are then passed into the convolutional network module J, followed by the MLP-based score classifier network C, to acquire a set of scores $U \in \mathbb{R}^{T_k\times 1}$. Each score $\{u_i\}|_{i}^{T_k}$ within this set of scores U represents the anomaly

likelihood of the snippet at the corresponding index and carries a value between 0 and 1. Each score u_i is rounded to produce a set of binary scores $U' = \{u'_i\}|_i^{T_k}$. When the binary score u' is 1, the snippet at the corresponding index is deemed to be anomalous; whereas, when the score is 0, the snippet at the corresponding index is assumed to be normal. Lastly, each binary score in U' is repeated δ times, preserving the original order, to reproduce a vector $\hat{U} = \{\hat{u}_i\}|_i^{\delta * T_k}$ with the common *frame* length as the video $\mathcal{X}^{(k)}$, for use in evaluation against the ground truth labels as in Eq. 9 below. The remainder frames $N_k - \delta * T_k$ are either discarded or padded with the final label of the video.

$$\hat{u}_{[\delta * i:\delta * (i+1)]} = u'_i \tag{9}$$

4 Experimental Results

4.1 Datasets and Metrics

UCF-Crime Dataset (Sultani et al., 2018) contains 1,900 untrimmed video clips encompassing 13 different anomalies and normal activities. The types of anomalies in the videos include abuse, arrest, arson, assault, burglary, explosion, fighting, road accident, robbery, shooting, shoplifting, stealing, and vandalism. Each of the real-world surveillance videos, totaling 128 hours in length, has been weakly annotated at video-level as anomalous or normal. The dataset comes pre-split into a train set of 800/810 normal/anomalous videos; a test set of 150/140 normal/anomalous videos.

ShanghaiTech Campus Dataset (Liu et al., 2018) contains 317,398 frames of video clips encompassing the scenes of multiple areas in ShanghaiTech Campus. The dataset cumulatively covers 13 scenes, in which 300,308 frames represent normal events and the remaining 17,090 frames comprise 130 distinct anomalous events. The dataset is split into a train set of 330 videos (274,515 frames) and a test set of 107 videos (42,883 frames), captured at 480×856 pixels. The train set contains only normal videos, while the test set contains a mix of normal and anomalous videos, where the anomalies in the test set are annotated at pixel-level.

XD-Violence Dataset (Wu et al., 2020) contains 217 hours of 4,754 untrimmed videos encompassing six different anomalies and normal activities. The anomalous actions in the dataset include abuse, car accident, explosion, fighting, riot, and shooting. The train set contains video-level annotations, while the test set contains frame-level annotations (*i.e.,* rough from-and-to frame locations of each anomaly, not to exceed three, in a video). The dataset is split into a train set of 3,954 videos, where 1,905 of them are anomalous, and a test set of 800 videos, where 500 are anomalous.

Metrics: Similar to other work (Hasan et al., 2016; Sultani et al., 2018; Tian et al., 2021; Wu & Liu, 2021; Wu et al., 2020; Zhong et al., 2019), UCF-Crime and Shanghai datasets are evaluated using AUC@ROC and XD-Violence dataset is evaluated using AUC@PR. AUC@ROC refers to the area under the receiver operating characteristics curve, whereas AUC@PR refers to the area under the precision-recall curve.

4.2 Implementation Details

In training time, we follow Sultani et al. (2018); Tian et al. (2021) and divide each video in the batch into 32 video snippets, (*i.e.,* T is set as 32 in train time), using Eq. 1. For all datasets, we follow the aforementioned steps to preprocess videos with the snippet length set to $\delta = 16$. The scorer network θ_s in Section 3.3 is defined as an MLP of three layers of 512, 256, and 1 units. The hidden layer is followed by a ReLU activation function, and the final layer is followed by a sigmoid function to produce a value between 0 and 1. To extract the linguistic scene elements features of the scene, we employ CLIP (Radford et al., 2021) that was pre-trained on a large-scale dataset of 400M image-text pairs crawled from the Internet. Thus, d is set as 512 for all experiments. We set M as 100 for Gaussian noise in Eq. 3. In addition, we choose 0.7 (70%), 0.7 (70%), and 0.9 (90%) for r in UCF-Crime, ShanghaiTech Campus, and XD-Violence datasets, respectively, for the best performance, where r denotes the number of snippets in a feature to place attention onto using TSA in a proportionate, relative figure rather to later compute κ in a hard number:

$$\kappa = \lfloor T \times r \rfloor \tag{10}$$

Our CLIP-TSA is trained in an end-to-end manner and implemented using PyTorch. We use the Adam optimizer (Kingma & Ba, 2015) with a weight decay of 0.005 and a batch size of 16 for 4,000 (UCF-Crime), 35,000 (ShanghaiTech Campus), and 4,000 (XD-Violence) epochs. The learning rate is set to 0.001 for all datasets.

Table 2: Performance comparisons (AUC@ROC) between the SOTA methods and our method on UCF-Crime dataset (Sultani et al., 2018). They are grouped into the unsupervised, supervised, and weakly-supervised methods in order.

Sup.	Method	Venue	Feature	AUC@ROC ↑
Un-	Hasan et al. (2016)	CVPR'16	-	50.60
	Lu et al. (2013)	ICCV'13	C3D	65.51
	BODS (Wang & Cherian, 2019)	ICCV'19	I3D	68.26
	GODS (Wang & Cherian, 2019)	ICCV'19	I3D	70.46
	GCL (Zaheer et al., 2022)	CVPR'22	ResNext	71.04
Fully-	Liu & Ma (2019)	MM'19	NLN	82.0
Weakly-	GCL_{WS} (Zaheer et al., 2022)	CVPR'21	ResNext	79.84
	GCN (Zhong et al., 2019)	CVPR'19	TSN	82.12
	WSAL (Lv et al., 2021)	TIP'21	TSN	85.38
	Purwanto et al. (2021)	ICCV'21	TRN	85.00
	Thakare et al. (2022)	ExpSys'22	C3D+I3D	84.48
	Sultani et al. (2018)	CVPR'18	C3D	75.41
	Zhang et al. (2019)	ICIP'19		78.70
	GCN (Zhong et al., 2019)	CVPR'19		81.08
	CLAWS (Zaheer et al., 2020)	ECCV'20		83.03
	RTFM (Tian et al., 2021)	ICCV'21		83.28
	Ours: CLIP-TSA†			**83.94**
	Sultani et al. (2018)	CVPR'18	I3D	77.92
	Wu et al. (2020)	ECCV'20		82.44
	DAM (Majhi et al., 2021)	AVSS'21		82.67
	BN-SVP (Sapkota & Yu, 2022)	CVPR'22		83.39
	RTFM (Tian et al., 2021)	ICCV'21		84.30
	Wu & Liu (2021)	TIP'21		84.89
	Ours: CLIP-TSA‡			**84.66**
	Ours: CLIP-TSA		CLIP	**87.58**

4.3 Performance Comparison

Besides CLIP-TSA, which is conducted on vision-language feature CLIP and our temporal attention mechanism TSA, we also test CLIP-TSA on other common features, *i.e.,* C3D (Ji et al., 2013) and I3D (Carreira & Zisserman, 2017), to fairly compare CLIP-TSA with other existing approaches. Thus, we report the performance of the CLIP-TSA and its variants in this section as follows:

- **CLIP-TSA**†: replacement of CLIP feature with C3D (Ji et al., 2013) feature (TSA Preserved)
- **CLIP-TSA**‡: replacement of CLIP feature with I3D (Carreira & Zisserman, 2017) feature (TSA Preserved)
- **CLIP-TSA**: utilization of the CLIP feature and the temporal attention mechanism TSA

Table 2 shows the frame-level AUC@ROC results of SOTA models that we have found to the best of our ability on the UCF-Crime dataset. Based on the table, first, it is apparent that unsupervised methods generally provide an inferior performance. Second, it can be observed that the performance of our method, CLIP-TSA, stands out against other SOTA methods by a large margin in any type of supervision setting. Compared to the current best-performing model, *i.e.,* Lv et al. (2021), our CLIP-TSA holds 2.2% better performance when evaluated with the same metric. Furthermore, on the same feature, our CLIP-TSA† yields better performance than the current SOTA on C3D by 0.66%, and CLIP-TSA‡ obtains very competitive scores on I3D.

Similarly, Table 3 shows the frame-level AUC@ROC results of SOTA models on the ShanghaiTech Campus dataset. In the table, it can be seen that our model outperforms all of the previous SOTA methods reported in the table. Empirically, it shows that, on the same feature, CLIP-TSA† beats BN-SVP (Sapkota & Yu, 2022), the current SOTA on C3D, by 1.19%, and CLIP-TSA‡ outperforms Wu & Liu (2021), the current SOTA on I3D, by 0.5%. Furthermore,

Table 3: Performance comparisons (AUC@ROC) between the SOTA methods and our method on ShanghaiTech Campus dataset (Liu et al., 2018). The first group is unsupervised methods, and the rest are weakly-supervised methods. Sultani et al. (2018)* is retrained with I3D features as it was previously not evaluated on ShanghaiTech Campus.

Sup.	Method	Venue	Feature	AUC@ROC ↑
Un-	Hasan et al. (2016)	CVPR'16	-	60.85
	Gao et al. (2021)	ICCV'19	-	71.20
	Yu et al. (2020)	MM'20	-	74.48
	GCL$_{PT}$ (Zaheer et al., 2022)	CVPR'21	ResNext	78.93
Weakly-	GCN (Zhong et al., 2019)	CVPR'19	TSN	84.44
	GCL$_{WS}$ (Zaheer et al., 2022)	CVPR'21	ResNext	86.21
	Purwanto et al. (2021)	ICCV'21	TRN	96.85
	GCN (Zhong et al., 2019)	CVPR'19	C3D	76.44
	Zhang et al. (2019)	ICIP'19		82.50
	CLAWS (Zaheer et al., 2020)	ECCV'20		89.67
	RTFM (Tian et al., 2021)	ICCV'21		91.57
	BN-SVP (Sapkota & Yu, 2022)	CVPR'22		96.00
	Ours: CLIP-TSA†			**97.19**
	Sultani et al. (2018)*	CVPR'18	I3D	85.33
	DAM (Majhi et al., 2021)	AVSS'21		88.22
	AR-Net (Wan et al., 2020)	ICME'20		91.24
	RTFM (Tian et al., 2021)	ICCV'21		97.21
	Wu & Liu (2021)	TIP'21		97.48
	Ours: CLIP-TSA‡			**97.98**
	Ours: CLIP-TSA		CLIP	**98.32**

Table 4: Performance comparisons (AUC@PR) between the SOTA methods and our method on XD-Violence dataset (Wu et al., 2020). The first group is an unsupervised method, and the other group is weakly-supervised methods. V and A represent visual and audio features, respectively.

Sup.	Modality	Method	Venue	Feature	AUC@PR ↑
Un-	–	OCSVM (Schölkopf et al., 1999)	NeurIPS'00	–	27.25
		Hasan et al. (2016)	CVPR'16	–	30.77
Weakly-	Vision & Audio	Wu & Liu (2021)	TIP'21	I3D(V) + VGGish(A)	75.90
		Wu et al. (2020)	ECCV'20	I3D(V) + VGGish(A)	78.64
		Pang et al. (2021)	ICASSP'21	I3D(V) + VGGish(A)	81.69
		MACIL-SD (Yu et al., 2022)	MM'22	I3D(V) + VGGish(A)	83.40
		DDL (Pu & Wu, 2022)	ICCECE'22	I3D(V) + VGGish(A)	83.54
	Vision	Sultani et al. (2018)	CVPR'18	C3D(V)	73.20
		RTFM (Tian et al., 2021)	ICCV'21	C3D(V)	75.89
		Ours: CLIP-TSA†		C3D(V)	**77.66**
		RTFM (Tian et al., 2021)	ICCV'21	I3D(V)	77.81
		Ours: CLIP-TSA‡		I3D(V)	**78.19**
		Ours: CLIP-TSA		CLIP(V)	**82.19**

CLIP-TSA yields superior performance to that of Wu & Liu (2021), the current best-performing model, by 0.84% with the end-to-end training scheme.

Lastly, Table 4 shows the frame-level AUC@PR results of SOTA models on the XD-Violence dataset, which is the most recently released dataset of the three. From the table, it can be seen that ours outperforms all SOTA models on various visual features as well as some models that leveraged both visual and auditory features. More specifically, it has left a remarkable margin of 1.77%, 0.38%, and 4.38% on C3D, I3D, and CLIP, respectively.

Our hypothesis for relatively small performance improvement on the ShanghaiTech Campus dataset compared to UCF-Crime and XD-Violence is that optimality has already been achieved in the ShanghaiTech Campus dataset as it is already yielding very high, near-100% scores by SOTA models. As a result, we believe that it is much more difficult to pull up its score in comparison to the remaining two datasets, with problems potentially being noise or subjective, frame-level human label errors.

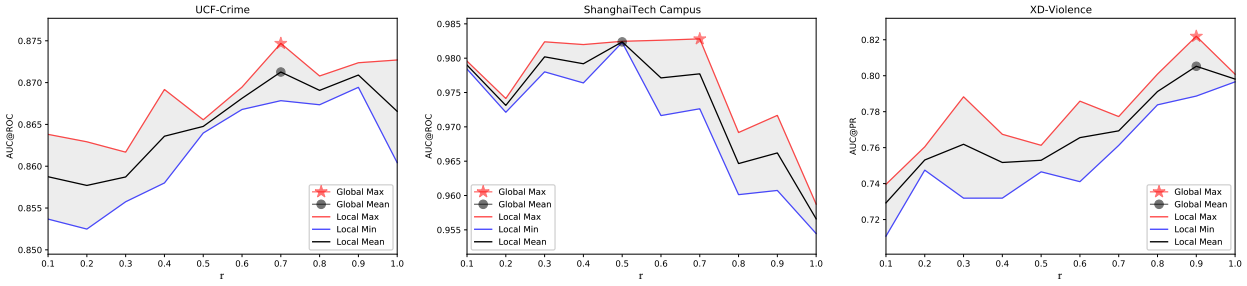

Figure 3: r-adjusted performance comparison chart. Local best denotes the best score yielded when tests were run with r set to the corresponding value in the x-axis. Global best denotes the best score achieved across all r. The experiments were conducted at 10 different intervals of 0.1, starting from $r = 0.1$.

4.4 Ablation Study

In this section, we conduct two ablations studies to analyze hyperparameter r in Eq. 10 and the effectiveness of the proposed TSA mechanism.

Hyperparameter Study First, we run the experiments with our CLIP-TSA model under the same setting as in Section 4.2, but using various r. More specifically, r will be set to steps of 0.1 from 0.1 to 1.0 (*i.e.,* $\{0.1, ..., 0.9, 1.0\}$).

To ensure the reported performance is generalized enough, we run the model five times each. The performance of the model at each r for each dataset is shown in Figure 3. According to the figure, the value of r where the model performs at optimal level differs for each. For example, UCF-Crime yields 87.6% AUC@ROC when r is set to 0.7, ShanghaiTech Campus obtains 98.3% AUC@ROC at $r \in [0.3, 0.7]$, and XD-Violence gets 82.2 AUC@PR at $r = 0.9$.

To understand why the optimal value of r is changing from dataset to dataset, we collect two pieces of information: 1) data distribution of UCF-Crime (1,900 videos with 13 types of anomalies), ShanghaiTech Campus (317,398 videos with 130 anomaly events), and XD-Violence (4,754 videos with 6 types of anomalies) datasets; 2) frame-level anomaly-to-all ratio ($\frac{\#ofAnomaly}{\#ofAnomaly+\#ofNormal}$) of their test sets, which are 0.1819, 0.4247, and 0.4977, respectively. The ShanghaiTech Campus is a big-scale data with a comparatively large number of anomaly events, so we assume that the model is able to be generalized enough, and the values of $r \in [0.3, 0.7]$ are around the anomalous-to-all ratio. However, UCF-Crime and XD-Violence have imbalanced anomaly category distributions as shown in Figure 4. Furthermore, the important features for making the correct decision are not limited to anomalous snippets, but also include some normal snippets as well, especially as part of computation for loss, in which both the magnitudes for top anomalous snippets and normal snippets are factored in. Thus, the best values of r in UCF-Crime and XD-Violence datasets are not aligned around anomalous-to-all ratios.

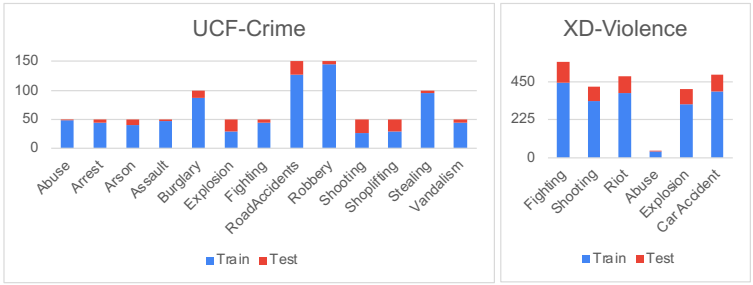

Figure 4: Per-anomaly class distribution of UCF-Crime (left) and XD-Violence (right) datasets on train and test sets.

Effectiveness of TSA mechanism In order to investigate the effectiveness of TSA, we conduct the experiments in two cases with and without TSA on the same vision-language feature on UCF-Crime, ShanghaiTech Campus, and XD-Violence datasets. Both experiments are sharing the same configuration settings with five different seeds, five different batch sizes (4, 8, 16, 32, 64), and three different learning rates (0.01, 0.001, 0.0001).

Table 5 reports the best performance of both CLIP-TSA and baseline model (w/o TSA) on three separate datasets. From the table, it can be observed that CLIP-TSA outperforms the baseline on all three when it is compared to the same dataset. Moreover, comparing Table 5 with Tables 2-4, we observe that the best-performing baseline model for

Table 5: Ablation study of TSA on UCF-Crime, ShanghaiTech Campus, and XD-Violence Datasets, using the corresponding metric for each. The best score is **bolded**, the runner-up is underlined, and the improved score after the TSA is *italicized.*

Feature	**TSA**	**UCF-Crime** (AUC@ROC ↑)	**ShanghaiTech** (AUC@ROC ↑)	**XD-Violence** (AUC@PR ↑)
C3D	✗	82.59	96.73	76.84
C3D	✓	*83.94*	*97.19*	*77.66*
I3D	✗	83.25	96.39	77.74
I3D	✓	*84.66*	*97.98*	*78.19*
ViT	✗	86.29	98.18	80.43
ViT	✓	***87.58***	***98.32***	***82.19***

each dataset is shown to yield a higher score than the SOTA models for the respective dataset. That implies the strength and efficiency of vision-language in VAD.

5 Conclusion

This paper presents CLIP-TSA, an effective end-to-end weakly-supervised VAD framework. Specifically, we proposed the novel TSA mechanism that maximizes attention on a subset of features while minimizing attention on noise and showed its applicability to the weakly-supervised VAD problem. We also applied TSA to CLIP-extracted features to demonstrate its efficacy in Visual Language features and exploited visual language features in the weakly-supervised VAD problem. We also empirically validate the excellence of our model on the three popular VAD datasets by comparing ours against the SOTAs.

Future investigations might aim for better techniques to incorporate both temporal and spatial information as well as handle imbalanced data with less annotation. Techniques for attention such as (Li et al., 2022) and self-supervised learning (Caron et al., 2021; Chen et al., 2020a) are also potential extensions for performance improvement.

References

Borislav Antić and Björn Ommer. Video parsing for abnormality detection. In *2011 International conference on computer vision*, pp. 2415–2422. IEEE, 2011.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Arslan Basharat, Alexei Gritai, and Mubarak Shah. Learning object motion patterns for anomaly detection and improved object detection. In *2008 IEEE conference on computer vision and pattern recognition*, pp. 1–8. IEEE, 2008.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pp. 6299–6308, 2017.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3): 1–58, 2009.

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

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020b.

Ruizhe Cheng, Bichen Wu, Peizhao Zhang, Peter Vajda, and Joseph E Gonzalez. Data-efficient language-supervised zero-shot learning with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3119–3124, 2021.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, October 2014.

Jean-Baptiste Cordonnier, Aravindh Mahendran, Alexey Dosovitskiy, Dirk Weissenborn, Jakob Uszkoreit, and Thomas Unterthiner. Differentiable patch selection for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2351–2360, June 2021.

Zihang Dai, Zhilin Yang, et al. Transformer-XL: Attentive language models beyond a fixed-length context. In *ACL*, pp. 2978–2988, 2019.

Keval Doshi and Yasin Yilmaz. Any-shot sequential anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 934–935, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021.

Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3354–3363, 2021.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pp. 6201–6210. IEEE, 2019.

Xinyang Feng, Dongjin Song, Yuncong Chen, Zhengzhang Chen, Jingchao Ni, and Haifeng Chen. *Convolutional Transformer Based Dual Discriminator Generative Adversarial Networks for Video Anomaly Detection*, pp. 5546–5554. ACMMM, New York, NY, USA, 2021.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pp. 214–229. Springer, 2020.

Xiaoyu Gao, Xiaoyong Zhao, and Lei Wang. Memory augmented variational auto-encoder for anomaly detection. In *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, pp. 728–732. IEEE, 2021.

Rohit Girdhar and Deva Ramanan. CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning. In *ICLR*, 2020.

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*, pp. 1705–1714, 2019.

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*, pp. 733–742, 2016.

Chengkun He, Jie Shao, and Jiayu Sun. An anomaly-introduced learning method for abnormal event detection. *Multimedia Tools and Applications*, 77:29573–29588, 2017.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. *Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Jing Huo, Yang Gao, Wanqi Yang, and Hujun Yin. Abnormal event detection via multi-instance dictionary learning. In *International conference on intelligent data engineering and automated learning*, pp. 76–83. Springer, 2012.

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*, pp. 7842–7851, 2019.

S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE TPAMI*, 35(1):221–231, 2013. doi: 10.1109/TPAMI.2012.59.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017.

Jie Lei, Liwei Wang, et al. MART: Memory-augmented recurrent transformer for coherent video paragraph captioning. In *ACL*, pp. 2603–2614, 2020a. doi: 10.18653/v1/2020.acl-main.233.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. *Empirical Methods in Natural Language Processing (EMNLP)*, 2020b.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11336–11344, 2020a.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *Association for Computational Linguistics*, 2020b.

Weixin Li and Nuno Vasconcelos. Multiple instance learning for soft bags via top instances. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pp. 121–137. Springer, 2020c.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *International Conference on Learning Representations*, 2021.

Yehao Li, Ting Yao, Yingwei Pan, and Tao Mei. Contextual transformer networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Kun Liu and Huadong Ma. Exploring background-bias for anomaly detection in surveillance videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1490–1499, 2019.

W. Liu, D. Lian W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pp. 2720–2727, 2013.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, 2015.

Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *IEEE Transactions on Image Processing*, 30:4505–4515, 2021. doi: 10.1109/TIP.2021.3072863.

Snehashis Majhi, Srijan Das, and François Brémond. Dam: Dissimilarity attention module for weakly-supervised video anomaly detection. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–8, 2021. doi: 10.1109/AVSS52988.2021.9663810.

Wen-Feng Pang, Qian-Hua He, Yong-jian Hu, and Yan-Xiong Li. Violence detection in videos based on fusing visual and audio information. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2260–2264, 2021. doi: 10.1109/ICASSP39728.2021.9413686.

Preksha Pareek and Ankit Thakkar. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, 54(3):2259–2322, 2021.

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*, pp. 14372–14381, 2020.

Ramakanth Pasunuru and Mohit Bansal. Multi-task video captioning with video and entailment generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1273–1283, Vancouver, Canada, 2017. doi: 10.18653/v1/P17-1117.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. doi: 10.1109/ICCV48922.2021.00209.

Yujiang Pu and Xiaoyu Wu. Audio-guided attention network for weakly supervised violence detection. In *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pp. 219–223, 2022. doi: 10.1109/ICCECE54139.2022.9712793.

Didik Purwanto, Yie-Tarng Chen, and Wen-Hsien Fang. Dance with self-attention: A new look of conditional random fields on anomaly detection in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 173–183, October 2021.

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*, pp. 8748–8763. PMLR, 2021.

Bharathkumar Ramachandra, Michael Jones, and Ranga Vatsavai. Learning a distance function with a siamese network to localize anomalies in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2598–2607, 2020.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3202–3212, 2015.

Venkatesh Saligrama and Zhu Chen. Video anomaly detection based on local statistical aggregates. In *2012 IEEE Conference on computer vision and pattern recognition*, pp. 2112–2119. IEEE, 2012.

Hitesh Sapkota and Qi Yu. Bayesian nonparametric submodular video partition for robust anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3212–3221, June 2022.

Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pp. 568–576, Cambridge, MA, USA, 2014. MIT Press.

Cees GM Snoek, Marcel Worring, et al. Concept-based video retrieval. *Foundations and Trends® in Information Retrieval*, 2(4): 215–322, 2009.

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*, pp. 6479–6488, 2018.

Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*, 2022.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Kamalakar Vijay Thakare, Nitin Sharma, Debi Prosad Dogra, Heeseung Choi, and Ig-Jae Kim. A multi-stream deep neural network with late fuzzy fusion for real-world anomaly detection. *Expert Systems with Applications*, 201:117030, 2022.

Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4975–4986, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, 2017.

Khoa Vo, Hyekang Joo, Kashu Yamazaki, Sang Truong, Kris Kitani, Minh-Triet Tran, and Ngan Le. Aei: Actors-environment interaction with adaptive attention for temporal action proposals generation. In *32nd British Machine Vision Conference 2021, BMVC 2021, Virtual Event, UK, November 22-25, 2021*, 2021a.

Khoa Vo, Hyekang Joo, Kashu Yamazaki, Sang Truong, Kris Kitani, Minh-Triet Tran, and Ngan Le. AEI: actors-environment interaction with adaptive attention for temporal action proposals generation. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, pp. 111. BMVA Press, 2021b.

Khoa Vo, Sang Truong, Kashu Yamazaki, Bhiksha Raj, Minh-Triet Tran, and Ngan Le. Aoe-net: Entities interactions modeling with adaptive attention mechanism for temporal action proposals generation. *International Journal of Computer Vision*, pp. 1–22, 2022.

Duc-Quang Vu, Ngan Le, and Jia-Ching Wang. Teaching yourself: A self-knowledge distillation approach to action recognition. *IEEE Access*, 9:105711–105723, 2021a.

Duc Quang Vu, Ngan TH Le, and Jia-Ching Wang. Self-supervised learning via multi-transformation classification for action recognition. *arXiv preprint arXiv:2102.10378*, 2021b.

Duc-Quang Vu, Ngan TH Le, and Jia-Ching Wang. (2+ 1) d distilled shufflenet: A lightweight unsupervised distillation network for human action recognition. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 3197–3203. IEEE Computer Society, 2022.

Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2020. doi: 10.1109/ICME46284.2020.9102722.

Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8201–8211, 2019.

Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5079–5088, 2021.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Michael Wray, Hazel Doughty, and Dima Damen. On semantic similarity in video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3650–3660, 2021.

Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, 30:3513–3527, 2021. doi: 10.1109/TIP.2021.3062192.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *ECCV 2020*, pp. 322–339, 2020.

Shandong Wu, Brian E Moore, and Mubarak Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 2054–2060. IEEE, 2010.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2048–2057, 2015.

Mengmeng Xu, Chen Zhao, David S. Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *CVPR*, June 2020.

Kashu Yamazaki, Sang Truong, Khoa Vo, Michael Kidd, Chase Rainwater, Khoa Luu, and Ngan Le. Vlcap: Vision-language with contrastive learning for coherent video paragraph captioning. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 3656–3661. IEEE, 2022.

Kashu Yamazaki, Khoa Vo, Sang Truong, Bhiksha Raj, and Ngan Le. VLTinT: Visual-Linguistic Transformer-in-Transformer for Coherent Video Paragraph Captioning. *arXiv e-prints*, November 2022.

Kashu Yamazaki, Khoa Vo, Sang Truong, Bhiksha Raj, and Ngan Le. Vltint: Visual-linguistic transformer-in-transformer for coherent video paragraph captioning. *AAAI*, 2023.

B. Yang and Y. Zou. CLIP Meets Video Captioners: Attribute-Aware Representation Learning Promotes Accurate Captioning. *PRCV*, 2022.

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 583–591, 2020.

Jiashuo Yu, Jinyu Liu, Ying Cheng, Rui Feng, and Yuejie Zhang. Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection. *MM*, 2022.

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*, pp. 14744–14754, 2022.

Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *ECCV 2020*, pp. 358–376, 2020.

Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *ICCV*, pp. 7094–7103, 2019.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, volume 13664 of *LNCS*, pp. 492–510, 2022.

Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4030–4034. IEEE, 2019.

Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. 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*, pp. 1237–1246, 2019.

Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *The British Machine Vision Conference (BMVC)*, 2019.