

MMVAD: A Vision-Language Model for Cross-Domain Video Anomaly Detection with Contrastive Learning and Scale-Adaptive Frame Segmentation

Debojyoti Biswas^{a,*}, Jelena Tesic^a

^aTexas State University, 601 University Dr, San Marcos 78666, TX, USA

ABSTRACT

Video Anomaly Detection (VAD) is crucial for public safety and detecting abnormalities in risk-prone zones. However, detecting anomalies from weakly labeled datasets has been very challenging for CCTV surveillance videos. The challenge is more intense when we involve high-altitude drone videos for VAD tasks. Very few works have been done on drone-captured VAD, and even the existing CCTV VAD methods suffer from several limitations that hinder their optimal performance. Previous VAD works mostly used single modal data, e.g., video data, which was insufficient to understand the context of complex scenes. Moreover, the existing multimodal systems use the traditional linear fusion method to capture multimodal feature interaction, which does not address the misalignment issue from different modalities. Next, the existing work relies on fixed-scale video segmentation, which fails to preserve the fine-grained local and global context knowledge. Also, it was found that the feature magnitude-based VAD does not correctly represent the anomalous events. To address these issues, we present a novel vision-language-based video anomaly detection for drone videos. We use adaptive long-short-term video segmentation (ALSVS) for local-global knowledge extraction. Next, we propose to use a shallow yet efficient attention-based feature fusion (AFF) technique for multimodal VAD (MMVAD) tasks. Finally, for the first time, we introduce feature anomaly learning based on a saliency-aware contrastive algorithm. We found contrastive anomaly feature learning is more robust than the magnitude-based loss calculation. We performed experiments on two of the latest drone VAD datasets (Drone-Anomaly and UIT Drone), as well as two CCTV VAD datasets (UCF crime and XD-Violence). Compared to the baseline and closest SOTA, we achieved at least a +3.8% and +3.3% increase in AUC, respectively, for the drone and CCTV datasets.

1. Introduction

The task of Video Anomaly Detection (VAD) is to detect anomalous activities from video scenes (See Figure 1). The task of VAD has been there for the last few years, mainly focusing on CCTV videos, detecting explosions, theft, illegal parking, etc [1, 2, 3]. There have been very few works on UAV videos due to a shortage of sufficient VAD datasets [4, 5]. However, due to the increasing demand for UAV images and videos for computer vision applications [6, 7], there is now a sufficient amount of UAV videos to explore VAD tasks. Bonetto et al. [8] investigated privacy-related issues using mini-drone videos. Five privacy protections have been assessed by the authors via a crowdsourcing evaluation. Apart from the fixed CCTV VAD research, Jiao et al. [9] explored the challenges of moving camera video anomaly detection. The authors focused on moving cameras with dynamic scenes, such as car dash-cams and cameras equipped on robots. The survey covers three major domains of application: security, urban transportation, and marine environments. Unmanned aerial vehicles (UAVs) are extensively used to inspect risky work zones and construction sites, as well as overhead traffic surveillance and threat detection. The primary motivation for using UAV images is

the low cost of surveillance devices and high geographical coverage due to High Geospatial Distance (GSD).

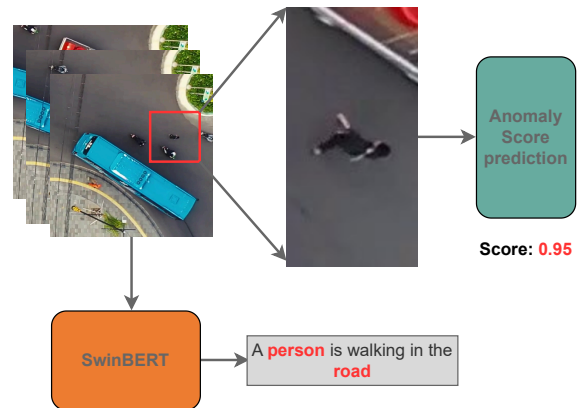



Figure 1: Multimodal VAD from the drone-captured scene. We use visual and text captions generated from scenes for anomaly classification and score prediction.

This paper investigates the effectiveness of weakly supervised VAD tasks where we have only video-level labels. The existing Unsupervised VAD task uses only regular videos to predict anomalous events. However, these unsupervised models suffer from sub-optimal results [10], reporting more false-positive cases as it is not possible to cover all regular events. On the other hand, the existing weakly supervised models mostly use single-domain video

*Debojyoti Biswas

 bishaldebojyoti@gmail.com (D. Biswas); jtesic@txstate.edu (J. Tesic)

ORCID(s): 0000-0002-8842-0207 (D. Biswas); 0000-0002-9972-9760 (J. Tesic)

data. Recent works reported that the single domain data is not sufficient for complex scene understating where we have complex backgrounds and a high number of object interactions [11, 12, 13]. Next, recent VAD models first extract video features using I3D/C3D networks [14, 15]. In the feature extraction process, all previous work relies on fixed-scale frame segmentation, where video snippet bags are created at fixed frame intervals [11, 16, 3]. The problem with a fixed frame rate is that all anomalous events are not the same in the temporal dimension; hence, as illustrated in Figure 2, the short anomalous events are not accurately captured with a long-term fixed segmentation rate. Many typical event frames will incur noise/ambiguity and lead to suboptimal results. Also, decreasing the frame segmentation will lead to incomplete scene information. Most of the previous VAD models perform traditional concatenation or multiplication on features to capture cross-modal interaction, ignoring the misalignment of unimodal video and text-domain features. An accurate fusion process is essential in order to aggregate rich semantic information. Finally, in the last few years, magnitude-based feature learning [3, 11] has been widely used for learning normal and abnormal scene features. However, the idea of calculating a single value to represent normality and abnormality is not always accurate [17, 1]. The high feature magnitude can come from the high number of objects, large-scale objects, or intense object movements in the frames.

To address the issues mentioned above, we propose multimodal video anomaly detection (MMVAD) (See Figure 3). Inspired by the work [11], we use text captions generated from video snippet bags. Since text features are semantically rich, we use text features as the second domain. Next, we create bags of frames called video snippets based on two scales. One is used for slow-forward, short-term scene capture, focused on fine-grained knowledge extraction. The other is fast-forward, relatively large-scale frame segmentation, focused on extracting robust context information for anomalous events. The text caption generated from videos often resides in different embedding spaces. The linear operation always incurs semantic and context aliasing phenomenon that adversely affects VAD performance in complex scenes. We introduce non-linear transformer-based attentive feature fusion (AFF). Figure 4 shows the designed feature fusion module where video snippets are used as patch tokens and text captions as CLS tokens. To address the issue of magnitude-based inaccurate learning, we propose saliency-aware feature enhancement, where we calculate the average power of every video snippet/bag and enhance the snippet frames features by injecting the overall energy. This way, we are able to propagate temporal information throughout the video snippet. Finally, we use the power-enhanced features for contrastive learning. We chose top K anomaly video snippets/bags for query and positive samples, and we chose the negative samples from top K regular video snippets/bags. We show that our method outperforms the current state-of-the-art (SOTA) methods by a large margin on the two latest UAV VAD datasets. The contributions are:

1. Use of adaptive long-short-term video segmentation (ALSVS) for local-global knowledge extraction.
2. Integration of a novel attention-based feature fusion (AFF) technique for video-text multimodal data.
3. Power-enhanced feature improvement for robust scene understanding.
4. Introducing contrastive learning instead of magnitude loss calculation for superior class separation on feature space.
5. Replacement of the binary cross entropy with the large-margin SVM classifier for the hard-example mining near the decision boundary.

Section 2 summarizes related work, and Section 3 introduces the proposed MMVAD method and describes the five different modules in the proposed pipeline. In Section 4, the proposed VAD model is evaluated using the latest cross-domain anomaly detection benchmarks over two high-altitude and two low-altitude VAD datasets. Finally, Section 5 summarizes the quantitative findings and outlines future works.

2. Related Works

2.1. Weakly Supervised Anomaly Detection

Leveraging the video-level labels from anomalous videos weakly supervised VAD techniques made significant improvements over the old-fashioned unsupervised VAD methods [18, 2, 1, 3, 16]. The earlier unsupervised works rejected using anomalous videos for training. Instead, they relied on the regular videos and used feature reconstruction error as the anomaly detection technique. However, this technique gives suboptimal results with false positive cases, as it is not possible to cover and define all normal scenes. The frame-level annotation in videos is very time-consuming and laborious. To trade off this issue, weakly supervised methods use video-level annotations. Sultani et al. [18] introduced weakly supervised VAD on surveillance video and outperformed previous unsupervised models. Since then, there have been many advancements in the method. The most widely used technique in weakly supervised VAD is multiple instance learning (MIL) [19, 18, 16]. However, there are a few shortcomings with this approach, which seeks more improvements. The major problem comes from the normal frames in the abnormal videos. The normal frame creates noise/ambiguity in the learning process. To fix this issue, several of the latest works use magnitude-based feature learning [3, 11, 20]. Although magnitude-based feature learning is not always accurate, the high magnitude value from the feature can be due to the high number of objects or intense object interaction in the scene [17, 1]. In this work, we propose to use the average power of the signal from each video snippet and inject it with each snippet feature to more correctly propagate temporal information. Finally, the power-enhanced feature will be used for contrastive learning.

2.2. Multi-modal Anomaly Detection

Multimodal training has been an efficient way to learn rich semantic and context features for the last few years. Multimodal training proved to be more successful and achieved

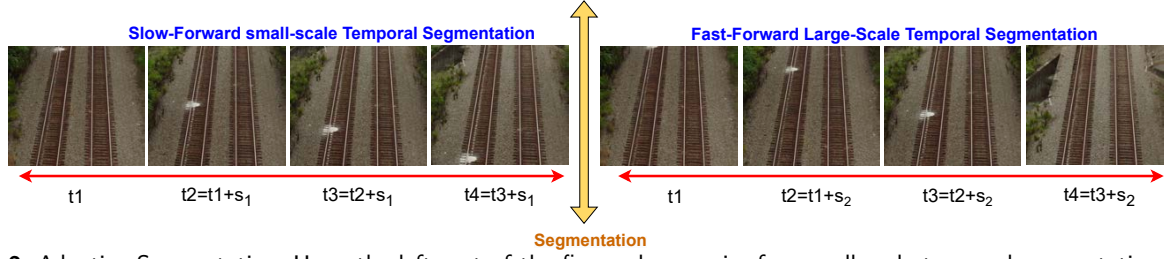


Figure 2: Adaptive Segmentation. Here, the left part of the figure shows noise-free small-scale temporal segmentation. On the other hand, the right part shows large-scale noisy segmentation, introducing normal scenes in the abnormal segmentation.

superior results than unimodal systems for diverse applications such as image classification, sign language interpretation, sentiment analysis, and many others [21, 22, 23, 24]. However, there has been very little work on multimodal VAD. Yang et al. [25] proposed the concept of Bottleneck Models using text sentences generated from GPT-3 for interpretable few-shot classification. Wu et al. [26] used an end-to-end trainable vision-language model for robust feature learning. On the other hand, Chen et al. [11] used text captions generated from video frames using a pre-trained SwinBERT [27] model for multimodal training. Although this work improves performance using the vision-language model, there are several drawbacks to the model. Firstly, similar to other works, this work uses fixed-scale frame segmentation for MIL. As we discussed in Section 1, fixed-scale frame segmentation incurs ambiguity and noise in the learning process. Also, this work follows linear old-fashioned fusion techniques such as concatenation and element-wise multiplication, ignoring the domain misalignment issue. To resolve these issues, we use multi-scale frame segmentation and attention-based feature fusion (AFF) for a more robust and efficient VAD.

2.3. Multimodal Feature Fusion

Multimodal training has been very popular due to its effectiveness in extracting rich context features. The application of multimodal training started with the most straightforward task, such as image classification. As it was found effective, later, it was incorporated into more complex tasks such as sentiment analysis, scene-graph generation, visual question answering, sign language understating, and video anomaly detection. In one of the first multimodal works [28, 29], the authors used the multiple kernel learning (MKL) framework method for fusing labeled and unlabeled images. Jaafar et al. [13] used several deep neural network layers to fusion audio, video, and text data for aggression detection in surveillance videos. Transformer-based fusion [30, 31, 22] was used to fuse hyperspectral image (HSI) and light detection and ranging (LiDAR) for remote sensing image classification. For the visual question-answering task, several works [32, 33, 34] used multimodal bilinear pooling and Hadamard product with co-attention. For the VAD task, Chen et al. [11] used linear operations such as concatenation, product, and addition of visual and text features. In their work, the concatenation method achieved the highest performance

gain. However, we found very little work on vision-language anomaly detection. Also, the existing multimodal works do not address the domain misalignment and aliasing issues coming from different modalities. To resolve the issues, we introduce attention-based feature fusion (AFF).

2.4. Magnitude Based Feature Anomaly Learning

Tian et al. [3], for the first time, proposed Robust Temporal Feature Magnitude learning (RTFM) to learn the degree of anomaly event from video scenes. They used temporal information with MIL to increase true positive anomaly prediction. RTFM also uses dilated convolution and self-attention mechanisms to capture long and short-term temporal interaction. They used the L2-norm on video snippet frames and later averaged the magnitudes from all frames to get a single feature magnitude for the video snippet. Finally, top-k video snippets from the anomaly and normal videos are kept for loss calculation and feature learning. This work got significant attention from researchers, and later, several other VAD models [11, 20] use a magnitude-based VAD mechanism for optimal performance. However, Chen et al. [1] argue about the correctness of using feature magnitude/saliency as the degree of anomaly detection criteria. This work proposes to use a Magnitude-Contrastive Glance-and-Focus Network (MGFN) for anomaly detection. Inspired by this work, we deal with the VAD task as a positive and negative sample learning problem. The anomalous samples are counted as positive cases, and normal samples as negative cases. It was evident that contrastive learning is very successful for complex cross-domain representation learning [7]. Hence, we decided to use contrastive learning to increase the separability between the positive and negative samples.

2.5. Contrastive Learning for Representation Learning

Hadsell et al. [35] first introduced the idea of representation learning in a contrastive fashion for the dimensionality reduction task. With the improvement of the InfoNCE loss [36], contrastive learning became more popular for clustering similar features together and dissimilar features at a distance in feature space. The idea of contrastive learning is very simple yet effective. In recent years, contrastive learning has been used in several challenging tasks such as domain adaptation [7], video anomaly detection[1], and

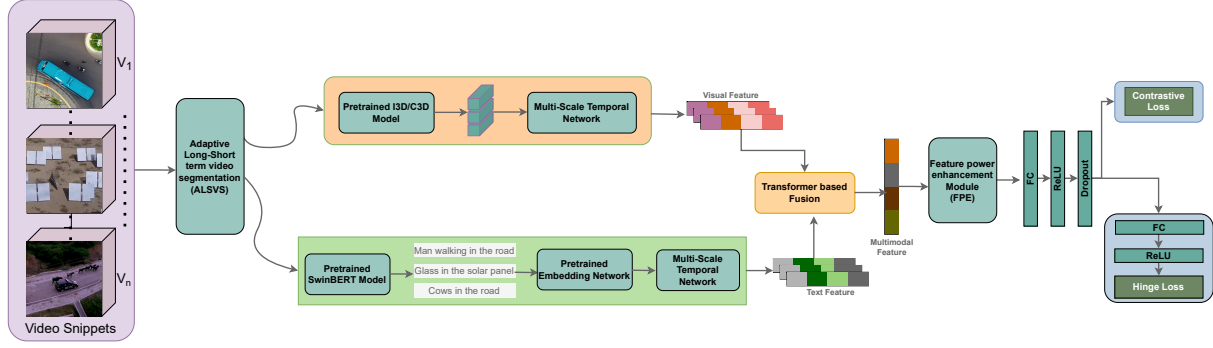


Figure 3: MMVAD: The Proposed Architecture. The MMVAD includes several modules, such as ALSVS, AFF, PEMF, and contrastive learning. Broadly, we have two pipelines: 1) Text modality and 2) Image modality. We use the AFF module to fuse both modalities and create multimodal features for anomaly classification and score prediction.

video-text alignments [37]. Contrastive learning is able to maintain effective separative characteristics for different domain feature embeddings. A similarity function such as Euclidean distance or cosine similarity is often used to calculate similarity among feature vectors. In this work, our goal is to use contrastive learning to group normal cases and abnormal cases at distant places in feature space. We decided to use contrastive learning based on its success in representing learning in earlier works [7, 1]. Motivated by these works, we use a custom InfoNCE loss for the VAD task. We present contrastive learning that is less susceptible to subtle samples from drone datasets by using multiple positive and negative examples.

3. Methodology

In this work, we use [11] as our baseline model and integrate our improvement over the architecture. Our proposed MMVAD architecture is illustrated in Figure 3. For each of the datasets, we have separate normal videos (V_n) and abnormal videos (V_a). The normal videos are denoted from V_n^i to V_n^N , and abnormal videos as V_a^i to V_a^N . Here, i denotes i th video out of N videos. We use the MIL approach as the core training architecture. So, the first step is to create a T number of snippets for each video. To do this, we use adaptive frame segmentation (ALSVS) as illustrated in Figure 2 and pass the snippets in the I3D network for feature extraction. The video features are denoted with $F_{vis} \in \mathbb{R}^{d_{vis}}$. Next, we use SwinBERT to extract video captions from ALSVS output in text format with a dimension of $F_{txt} \in \mathbb{R}^{d_{txt}}$. We take advantage of multi-scale temporal networks (MTN) in both branches to better capture the temporal dependencies. The visual and text features are then fused using an attention-based fusion technique, as shown in Figure 4. Next, we calculate the average power-enhanced magnitude for each snippet and pick Top- k snippets from normal and abnormal videos for contrastive feature learning. We use the SVM classifier and contrastive learning to separate normal and abnormal features in feature space. For frame-level predictions and evaluation, we propagate the snippet-level predictions to each frame and compare them with the ground truth (GT).

The next subsections will explain each of our improvements in detail.

3.1. Adaptive Long-Short-Term Video Segmentation (ALSVS)

The core of our architecture is Multiple Instance Learning (MIL) based on video snippets. The previous works [3, 11, 1] rely on fixed-stride/scale video segmentation for creating video snippets. Since an anomaly event can range from a very short period (e.g., Explosion, shootout) to a more extended period (e.g., Protest, stealing) of time, fixed-scale video segmentation is not an accurate way to create video snippets. Similar issues were discussed in some earlier works related to sign-language recognition [38], where the author used multi-scale stride instead of fixed single stride for slow-fast frame segmentation. Using multi-scale frame segmentation, they were able to create better video snippets and achieve optimal results. Next, Gopalakrishnan et al. [39] used a slow-fast network for human action recognition, which is a similar task to VAD. In this work, the authors use a slow pathway that captures spatial semantics with a lower frame rate and another fast pathway with a high frame rate for capturing motion data and information dynamics. In both of the works, the authors address the fixed-scale video segmentation issue suboptimally. The strides used in both of the works are fixed and do not change with the video lengths; hence, does not provide complete noise-free segmentation. As shown in Figure 2, improper segmentation can lead to ambiguity/noise in the learning process. Although the baseline [11] work has a multi-scale temporal network; it provides suboptimal performance due to inefficient snippet creation.

We use T number of snippets for each video. To create short-term snippets, we use a lower stride S_1 , and for long-term snippets, we use a higher stride S_2 . The values of both strides are not fixed; they adapt based on the duration/number of frames in the video. The goal is to make $N = T/2 * S_1 + T/2 * S_2$, where N is the total number of frames in the video. As discussed earlier, S_1 is the small stride, and we use the formula $S_2 = \beta * S_1$ to define the large stride. The value of β is explored by experiments, and we found that $\beta = 1.33$ gives optimal

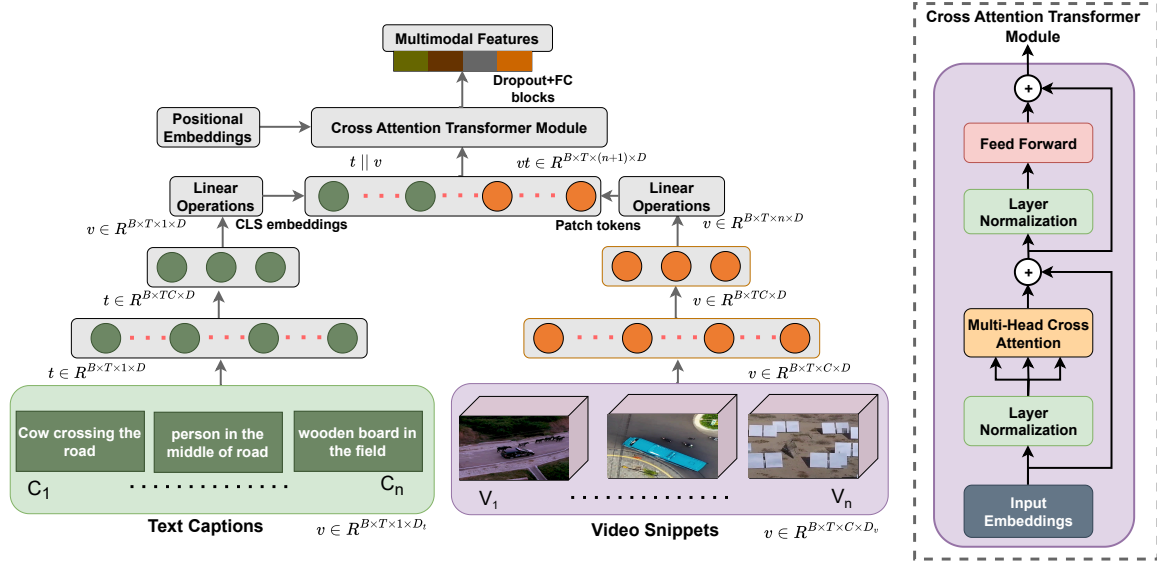


Figure 4: Attention-based Multimodal feature fusion (AFF) module. Here, text captions are passed to create CLS embeddings, and the video snippets are passed to create patch tokens. Next, we use the traditional multi-head attention mechanism for rich multimodal feature generation.

and stable results. The value of S_1 is found by applying a binary search algorithm on the total frame number N . Our method is faster than most of the existing methods because we were able to reduce the time complexity of the frame segmentation to $O(\log N)$. For example, if we have 10,000 frames in a video, we will need only 13 comparisons to find a suitable value for S_1 . Most of the VAD datasets are weakly labeled, which means we have annotations on the video level. Stacking fixed-sized snippets makes the segmentation inefficient and erroneous. Our idea is to stack the snippets by grouping them in a short-term followed by a long-term snippet ($t1_{short} || t2_{long} || t3_{short} || t4_{long} \dots tN$) to get optimal accuracy in frame segmentation. This way, some snippets will achieve a high anomaly score due to pure anomaly frames in the snippet, and some will achieve a very low anomaly score due to the majority of normal frames in the snippet. In section 3.3, we discuss that we pick only top-K snippets based on the anomaly score for final classification, and by this, we can reject the low-scored normal snippets from an anomaly video for further consideration.

3.2. Attention-based Feature Fusion (AFF)

For visual feature extraction, we use the I3D/C3D network, and we denote with $F_{vis} \in \mathbb{R}^{d_{vis}}$. On the other hand, for text feature extraction, we use the swinBERT pre-trained model and denote with $F_{txt} \in \mathbb{R}^{d_{txt}}$. Here, the dimension of d_{vis} and d_{txt} is $B \times T \times M \times D_v$ and $B \times T \times 1 \times D_t$, respectively, where B is the batch size, T and M are respectively number of snippets per video and number of crops per frames, and C represents the vector length. There is an MTN module in both the visual and text feature extraction branches. The goal of MTN is to extract important short and long-range temporal dependencies. MTN architecture was first introduced by [3] using a 3-layer pyramid dilated

convolutions (PDC) [40] block and a non-local block (NLB) [41]. We follow the same architecture but adapt it according to our visual and text feature dimensions. The output from the MTN block is $B \times M \times T \times D_v$ and $B \times M \times T \times D_t$, respectively, for visual and text features.

The first step in our attention-based feature fusion is tokenizing video snippets and text captions. First, we change the formation of the visual feature matrix to $B \times T \times M \times D_{dim}$ for the linear operation. Our goal is to extract n key patches/crops out of $M=10$ to reduce later computation. Next, we create a CLS token from each snippet text feature and add the CLS token in each snippet of the video. To extract key features from video and CLS token from text, we use two weight parameters, W_a and W_b . The operations for visual token (see Eq. 1) creation are given as follows:

$$\begin{aligned} F_{vis}^{out} &= MTN(F_{vis}) \\ F_{wa}^{vis} &= Softmax(T(F_{vis} \cdot W_a^{vis})) \\ F_{wb}^{vis} &= F_{vis}^{out} \cdot W_b^{vis} \\ F_{token}^{vis} &= F_{wa}^{vis} \cdot F_{wb}^{vis} \end{aligned} \quad (1)$$

We calculate the patch token for each snippet at a time. Here, the $T(\cdot)$ denotes matrix transpose operation, $W_a^{vis} \in \mathbb{R}^{D_v \times n}$, $W_b^{vis} \in \mathbb{R}^{D_v \times D_v}$ and the $F_{token}^{vis} \in \mathbb{R}^{n \times D_v}$. Next, CLS tokens are derived from the text features as outlined in Equation 2.

$$\begin{aligned} F_{txt}^{out} &= MTN(F_{txt}) \\ F_{wa}^{txt} &= Softmax(T(F_{txt} \cdot W_a^{txt})) \\ F_{wb}^{txt} &= F_{txt}^{out} \cdot W_b^{txt} \\ F_{token}^{CLS} &= F_{wa}^{txt} \cdot F_{wb}^{txt} \end{aligned} \quad (2)$$

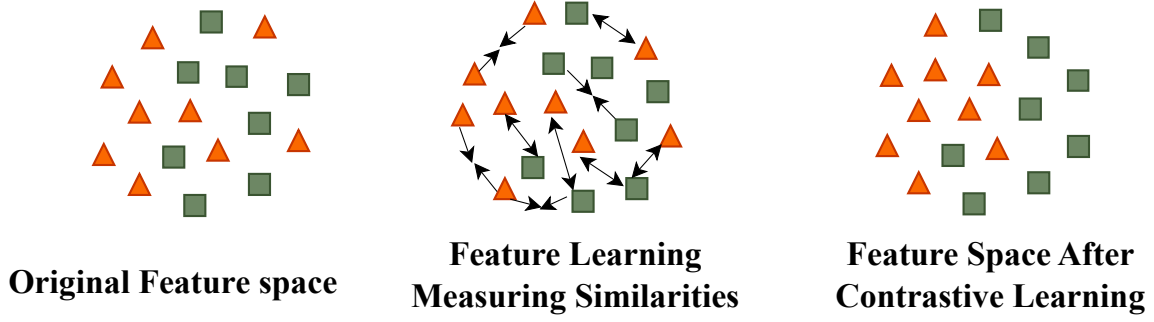


Figure 5: Normal and Abnormal Feature Learning using Contrastive Learning Algorithm. Here, different colored shapes represent features from different groups.

In Eq. 2, the $T(\cdot)$ denotes matrix transpose operation, $W_a^{txt} \in \mathbb{R}^{D_v \times 1}$, $W_b^{txt} \in \mathbb{R}^{D_t \times D_v}$ and the $F_{token}^{CLS} \in \mathbb{R}^{1 \times D_v}$. Here, we choose the W_b^{txt} dimension $D_t \times D_v$ so that we can match the visual token dimension for further matrix operation. Now, we stack the CLS token for a video snippet with its visual token (see Eq. 3). We have n visual token for each video snippet, so the total number of tokens for each video snippet becomes $(n+1)$. Our operation can be written as Equation 3.

$$F_{token}^{fused} = F_{token}^{vis} \parallel F_{token}^{CLS} \quad (3)$$

Figure 3 illustrates why we have to retain the positional information for each token inside each snippet. We add a trainable positional embedding with each token embedding to achieve this goal (see Eq. 4).

$$\begin{aligned} F_{pos}^{fused} &= F_{token}^{vis} \oplus P_{emb} \parallel F_{token}^{CLS} \oplus P_{emb} \\ &= F_{pos}^{vis} \parallel F_{pos}^{CLS} \end{aligned} \quad (4)$$

After creating the positional Vision-Text embedding (F_{pos}^{fused}), we pass it into a traditional cross-attention transformer encoder to train for abstract representation of visual and textual rich semantic information throughout the $(n+1)$ tokens. We can see the architecture of the cross-attention transformer encoder from Figure 4. The abstraction is carried out X times inside the attention encoder. The dimension of the transformer encoder output is $F_{encoder}^{out} \in \mathbb{R}^{B \times T \times (n+1) \times D_v}$. The encoder block is followed by a mean operation (Eq. 5) in the $(n+1)$ token dimension to summarize the overall trends or representations for a snippet. The mean operation collapses the 3rd and 4th dimensions, and the output becomes $F_{out} \in \mathbb{R}^{B \times T \times D_v}$.

$$F_{out} = Dropout(Mean(F_{encoder}^{out}, axis = (2, 3))) \quad (5)$$

The succeeding dropout layer ensures that the network does not overfit due to K times attention abstraction. The F_{out} is the final multimodal vision-text feature, which is later used for classification prediction and contrastive learning.

3.3. Power-Enhanced Multimodal Feature

As discussed in the section 1, the approach used in previous works [11, 3] for addressing a snippet as an anomaly

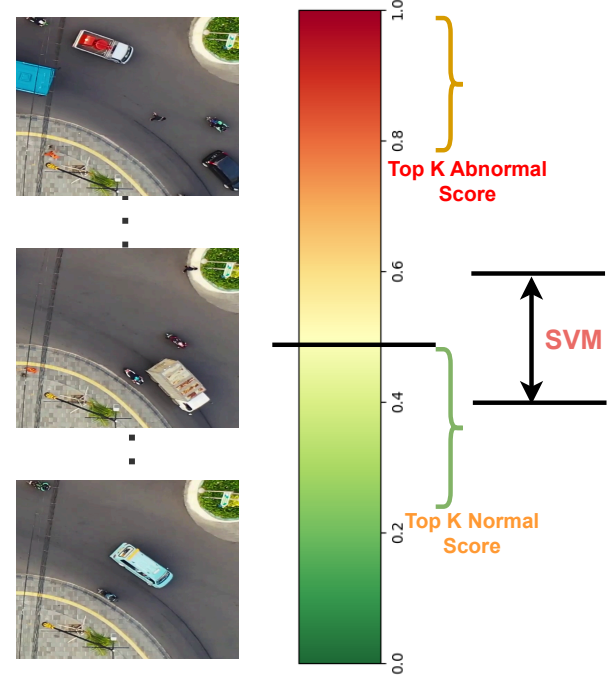


Figure 6: Large Margin SVM Classifier for better separating the data points near the decision boundary. This particularly helps in the case of high-altitude drone scenes where the difference between normal and abnormal scenes is very subtle.

is not entirely accurate. In those works, the authors used the mean magnitude (L2 norm) of a snippet to rank the snippets in a video. The assumption is that if we have a high feature intensity value, it must be due to some anomalous activities in the frames. However, it was argued earlier in previous works [7, 1] that such a notion does not always hold in every situation. The high intensity can be due to the number of objects, intense movement, or interactions. Inspired by [1], we decided to inject the average RMS power of a video into its snippets. We reject the idea of using L2 magnitude as it contributes a larger value compared to the RMS and leads to gradient explosion and overfitting. Our goal is to propagate the key saliency information in a video to its snippets (see Eq. 6) and then calculate the L2 norm to rank the snippets (see Eq. 7). The operations are as follows:

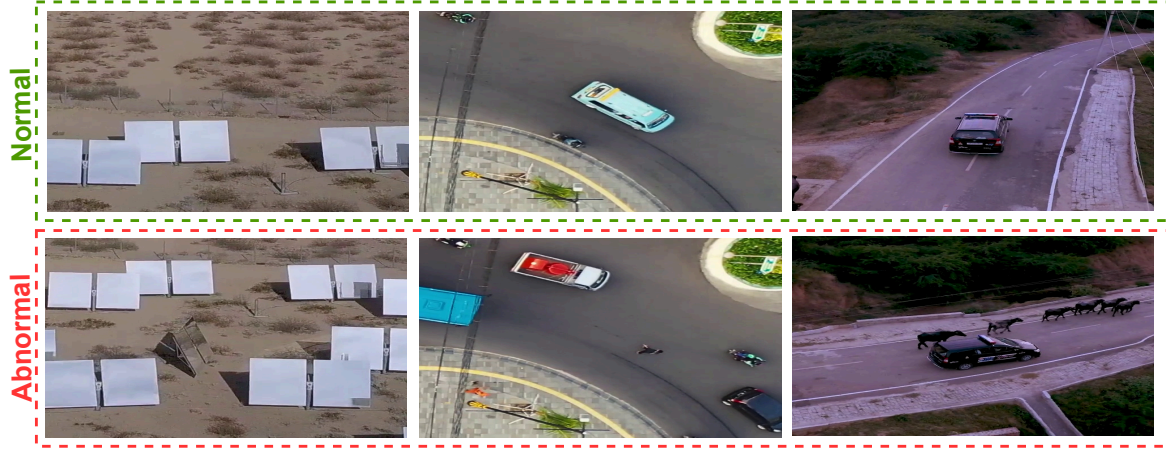


Figure 7: Normal and Abnormal scene samples from Drone-Anomaly dataset

$$\begin{aligned} \text{RMS}_{b,t} &= \sqrt{\frac{1}{D_v} \sum_{d=1}^{D_v} (F_{b,t,d}^{\text{out}})^2} \\ \text{Avg_RMS}_b &= \frac{1}{T} \sum_{t=1}^T \text{RMS}_{b,t} \\ F^{\text{RMS}} &= \text{Avg_RMS}_{B,T} \oplus F_{\text{out}} \end{aligned} \quad (6)$$

$$F_{\text{topK}} = \text{Max}_K(\|F_{i \in b,t}^{\text{RMS}}\|_2) \quad (7)$$

We rank the video snippets and select only *top K* from abnormal and normal videos. We select *top K* snippets from the abnormal videos because those videos also include normal frames. To correctly learn anomaly features, we need accurate noise-free abnormal cases. On the other hand, *top K* snippets in the normal video represent the normal snippets that share some characteristics of anomalous events and are close to the decision boundary. This fashion of snippet selection gives us a maximum margin near the decision boundary.

3.4. Contrastive Learning

We separate the features from abnormal and normal scenes in different places of feature space. To do that, we use contrastive learning. Contrastive learning has been very successful in feature clustering in feature space in past years [7]. As shown in Figure 5, contrastive learning use similarity functions such as: cosine-similarity to attract similar features together and keeps apart dissimilar feature at distant spaces. Contrastive learning uses Query (Q), Positive (P), and Negative (N) samples to separate features in feature space. Query and positive samples are drawn from the same class, and negative cases are drawn from different classes. Here, in our work, we select a random sample from *top K* anomaly snippets as a query, and N_{pos} out of the remaining *top (K-1)* samples are selected as positive cases. Apart from traditional

contrastive learning, we use more than one positive case for robust feature learning. Using more than one positive case helps to cluster features where intra-class discrepancy is an issue. Next, we select N_{neg} negative samples from the regular videos. The general formula for contrastive learning is in Eq. 8 as:

$$\text{InfoNCE} = -\log \frac{\sum_{i=1}^{N_{\text{pos}}} \exp(\text{sim}(Q, P_i^+)/\tau)}{\sum_{k=1}^{N_{\text{neg}}} \exp(\text{sim}(Q, N_k^-)/\tau)} \quad (8)$$

The steps to calculate contrastive learning are outlined in the sequence below:

$$\begin{aligned} Q &= F_{i,\text{topK}}^a, P^+ = F_{j,\text{topK}}^a \text{ where } i \neq j \text{ and } j \in N_{\text{pos}} \\ N^- &= F_{k,\text{topK}}^n \text{ where } k \in N_{\text{neg}} \\ \text{Logits}^+ &= Q.T(P^+) \\ \text{Logits}_{1 \times 1}^+ &= \sum_{i=1}^{N_{\text{pos}}} \text{Logits}_{\in 1 \times N_{\text{pos}}}^+ \\ \text{Logits}_{\in 1 \times N_{\text{neg}}}^- &= Q.T(N^-) \\ \text{Logits} &= \text{Logits}^+ \parallel \text{Logits}^- \end{aligned} \quad (9)$$

$$\text{Loss}_{\text{cons}} = \text{Mean}(\text{Softmax}(\text{Logits})) \quad (10)$$

Here, the dimension of $Q \in 1 \times D_v$, $P^+ \in N_{\text{pos}} \times D_v$ and $N^- \in N_{\text{neg}} \times D_v$. After calculating the positive and negative logits, we stack them together and pass them into *Softmax* function for probability calculation. We reduce the softmax output by calculating the mean value to get the final loss.

3.5. Large-Margin SVM Classification

In the baseline architecture, Binary Cross Entropy (BCE) was used as the loss function. However, in our experimental

datasets, there are plenty of samples where the difference between normal and abnormal events is very subtle. We can verify such a scenario from Figure 6, where we can see a single person's position in the street can impact the decision. To increase the margin around the decision boundary, we decided to use SVM as the classifier. Similar to contrastive learning, we only pick top K normal and abnormal snippets for loss calculation. We pass $F_{i,topK}^a$ and $F_{i,topK}^n$ to a linear unit followed by ReLU activation. Next, we use the hinge loss criterion for total loss calculation. The following are the steps:

$$\begin{aligned} S_a &= ReLU(MLP(F_{i,topK}^a)) \\ S_n &= ReLU(MLP(F_{i,topK}^n)) \\ S &= Sigmoid(S_a || S_n) \end{aligned} \quad (11)$$

$$HingeLoss(S, y) = \frac{1}{2 * K} \sum_{i=1}^{2*K} \max(0, \text{margin} - y_i \cdot S_i) \quad (12)$$

Above, From Eq. 11, we get the probability scores for each of the top K snippets; in total, we have $2 * k$ scores from abnormal and normal videos. Next, we pass the scores and segment labels into the *Hingeloss* (Eq. 12) and calculate the classification loss. The overall loss can be calculated below Eq. 13 where W_c and W_h are weight parameters for contrastive and hinge loss, respectively.

$$Loss = W_c * Loss_{cons} + W_h * HingeLoss(S, y) \quad (13)$$

4. Experiments

4.1. Datasets

We use two publicly available drone and CCTV datasets to verify the model's effectiveness. Our goal is to prove the cross-domain performance in low—and high-altitude situations. The datasets we consider for experiments are Drone-Anomaly, UIT-ADrone, UCF Crime, and XD-Violence. Below are the dataset's short descriptions:

Drone Anomaly: Drone datasets are comparatively more challenging for anomaly detection. The main reason is the altitude of the frame capture and the very subtle difference between normal and abnormal scenes. The Drone-anomaly dataset [4] is collected from online sources such as YouTube and Pixels. The dataset contains seven real-world scenes, such as highways, crossroads, bike roundabouts, vehicle roundabouts, railway inspections, solar panel inspections, and farmland inspections. Figure 7 illustrates the samples of normal and abnormal scenes. In total, we have 59 videos in the dataset, among which 37 are training and 22 are testing videos. The dataset is designed for weakly-supervised anomaly detection with only video-level ground truth annotation for each video. The dataset has 87,488 video

frames (51,635 for training and 35,853 for testing) with a frame rate of 30 frames/s.

UIT-ADrone: UIT-ADrone is our second drone dataset [5] for the VAD task. This dataset is more focused on three real-life traffic scene roundabouts in Ho Chi Minh City in Vietnam. The dataset has 51 video recordings totaling 6.5h runtime. The traffic recordings are from several complex scenarios that pose challenges, such as complex scenes, occlusion, high density, lighting conditions, and small objects. The abnormal activities in the dataset are divided into ten categories. The categories are as follows: crossing the road in the wrong lane, walking under the street, driving in the wrong roundabout, illegally driving on the sidewalk, illegal left turn/turn right, illegally parking in the street, carrying oversized loads, sidewalk parking, driving opposite-directions, and accidents via motorcycles. The actions in the dataset are very subtle and can be confirmed from Figure 8. We can see that one single person *walking under the street* can change the scene categorization completely. Also, there can be more than abnormal events in a single frame due to the huge diversity of situations that arise on real-world roundabouts. The dataset has only normal videos for training and both normal and abnormal videos for testing. This dataset is supervised and annotated at the frame level. There are 592 training videos and 905 testing videos.

UCF Crime: UCF Crime [18] is the most popular benchmark dataset for anomaly detection. The dataset is collected from online sources such as YouTube and Live-Leak using different text search prompts. There are, in total, 13 different anomalous scenarios recorded in the videos. The anomaly events are Abuse, Arrest, Arson, Assault, Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism. This dataset includes anomalous events, which are important concerns for public safety. The dataset has 1900 videos with a duration of 128hr, where 950 are regular videos and 950 are anomalous videos. The training and testing set contains 1610 and 290 videos, respectively. The normal and abnormal event samples from the dataset are illustrated in Figure 9.

XD-Violence: This dataset [42] shares similar characteristics to the UCF Crime dataset. Both are focused on different crime scene anomaly events. The videos are collected from multiple sources, mostly YouTube and movies, including cartoons, sports, games, music, etc. There are 4754 videos, with 2405 violent videos and 2349 non-violent videos. The dataset is divided into two parts: 3954 videos in the training and 800 videos in the testing set. The annotation in the dataset is at the video level and is suitable for weakly supervised training. The anomaly scene samples from the dataset are presented in Figure 10.

4.2. Implementation Details

We use Python as a programming language and PyTorch as the deep learning framework to implement the whole project. We used the GitHub repository from the latest work, TEVAD [11], as our baseline code structure. At first, to capture temporal information, we created video snippets/bags of



Figure 8: Normal and Abnormal scene samples from UIT-ADrone dataset



Figure 9: Normal and Abnormal scene samples from UCF Crime dataset

frames and performed MIL. Throughout all experiments, we create $T=32$ snippets out of each video. Next, we extracted the video features using the I3D/C3D model with num. of crop $M=10$ and saved them as feature vectors of length 2048 in the NumPy array. However, we have replaced the feature extractor ResNet50 with CSP DarkNet53 because of its superior performance for small objects [7]. Next, we use the SwinBERT pre-trained model [27] to generate captions from the videos. The text sentence captions are then converted into numerical vectors of length 768 and saved as a NumPy array. We use a transformer-based fusion method to combine the visual and text feature vectors. During the multimodal feature fusion, we selected $n=4$ crops out of 10 visual crops of each frame to reduce computational cost. From the abnormal and normal video snippets, we have picked the top $K=5$ snippets out of $T=32$. Those top K snippets were later used for the binary classification task and contrastive learning. For binary classification, we use the PyTorch implementation of *MultiMarginLoss* as *Hingeloss* with a margin of 10. For contrastive learning, we have used our implementation of InfoNCE loss as discussed in section

Table 1
System Specifications.

System Unit	Specification
CPU	Intel® Core™ i9-11900K
CPU Core	3.50GHzx16
RAM	167GB 128-bit LPDDR4
GPU	GP102 TITAN Xp
GPU Memory	12GB

3.4 with $N_{pos}=4$ and $N_{neg}=8$ in all experiments. The configuration of the workstation used for all experiments is presented in Table 1, and the list of all hyper-parameters is presented in Table 2.

4.3. State of the Art

We compare our proposed model with several SOTA VAD models. We include a few models that focus on low altitude and a few that focus on high altitude to improve



Figure 10: Normal and Abnormal scene samples from XD-Violence dataset

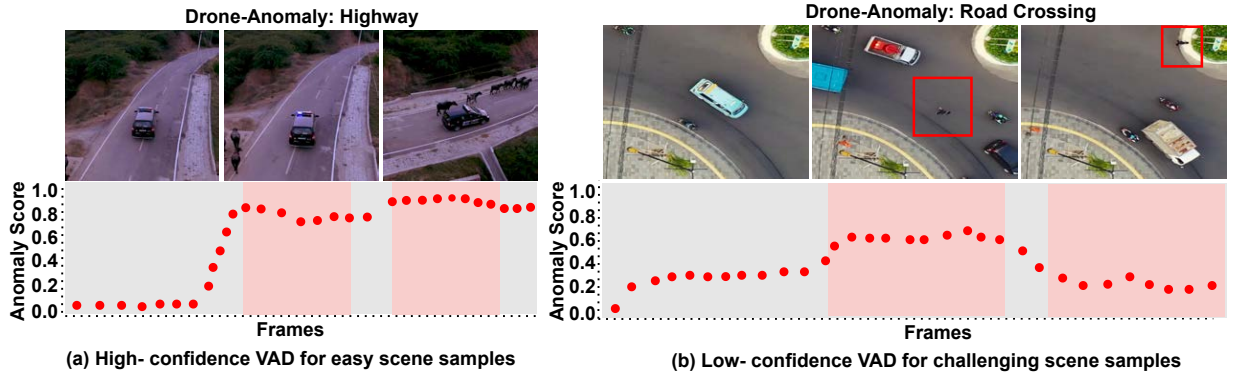


Figure 11: Frame-level anomaly detection with anomaly score using MMVAD model for Drone-Anomaly dataset.

Table 2

List of all hyper-parameters with their values.

Hyper-parameter	Value
Batchsize (B)	8 (8 norm + 8 abnormal)
Num. of snippets (T)	32
Num. of crops (M)	10
Num. of selected crops (n)	4
Top K feature	5
N_{pos} in CL	4
N_{neg} in CL	8
τ in CL	0.01
W_h in SVM	0.7
W_c in CL	0.3
Margin in SVM	10

the qualitative comparison. The brief descriptions of each of them are given below:

RTFM: This work applies multiple instance learning (MIL) for VAD by creating video snippets from video

frames. The focus of this work is to correctly detect abnormal cases that are very subtle from normal scenes. The work introduces a novel method named Robust Temporal Feature Magnitude Learning (RTFM), which trains a feature magnitude learning function to recognize the positive and negative instances effectively. RTFM performs satisfactorily on low-altitude CCTV video datasets.

TEVAD: This is our baseline model from Chen et al. [11]. This is the first proposed model to introduce multimodal VAD architecture. This work shows the efficacy of using a vision-text model for extracting rich context information from complex scenes. The work uses magnitude-based feature learning following the previous work RTFM [3]. TEVAD shows improved performance over RTFM on low-altitude CCTV video datasets.

MGFN [1] argues about the correctness of magnitude-based feature learning in the RTFM paper. It was observed that the feature magnitude sometimes can be misleading, as a high-magnitude value does not always relate to anomaly events. To resolve this issue, this work proposes a Magnitude-Contrastive Glance-and-Focus Network (MGFN) for anomaly detection. Our application of power-enhanced feature learning is inspired by them, but we apply

Table 3

Frame-level AUC comparison between State-of-the-art and proposed model on Drone-Anomaly dataset.

Supervision	Method	Feature	AUC
Weakly-Supervised	TEVAD [11]	I3D-RGB	67.35
	RTFM [3]	I3D-RGB	65.56
	MGFN [1]	VideoSwin-RGB	68.92
	MSL [43]	VideoSwin-RGB	67.11
Unsupervised	ANDT [4]	ViT	60.35
	ASTT [44]	ViT	67.80
Weakly-Supervised	MMVAD*	I3D-RGB (RNet50)	70.78
	MMVAD	I3D-RGB (CSPDNet53)	71.22

it for multimodal domains more accurately and robustly. MGFN outperforms other SOTA on CCTV video datasets. **MSL** [43] works quite similarly to RTFM, but it fixed the early-noise issue due to inferior feature quality. They introduce the Multi-Sequence Learning (MSL) method and a hinge-based MSL ranking loss that uses a sequence composed of multiple snippets as an optimization unit. They also integrate a transformer module that learns video-level and snippet-level anomaly probability.

ANDT [4] focuses on high-altitude drone anomaly detection. This work also uses MIL and a transformer head as an encoder for video snippets. Next, they predict the upcoming frame from previous information. ANDT learns normal scenes in the training phase in an unsupervised manner and predicts an event as an anomaly with unpredictable temporal dynamics in the test phase.

ASTT [44] focuses on high-altitude drone video benchmarks and several challenging scenarios: small sizes, multi-scale objects, complex backgrounds of great variations, and high overlap between objects, as related to traffic anomaly detection. They propose a sequencing method similar to ANDT, which uses a transformer-encoder to predict future abnormal scenes from high-altitude videos in an unsupervised way.

4.4. Performance Comparison

Performance Metrics: To prove the effectiveness of our model with other comparing methods, we adopt Area Under the Curve (AUC) as an evaluation metric for the Drone-Anomaly, UIT-ADrone, UCF-Crime and utilize frame-level Average Precision (AP) as the metrics for the XD-Violence. Note that Higher AUC and AP indicate better performance for the model. By following the previous VAD works [11, 3, 43, 1], it was evident that the AUC and AP are more suitable metrics for anomaly detection evaluation. We reject using the F1-score, as it was found that [45] the F1-score is sensitive for anomaly detection and biased towards the number of abnormal samples in the test set. From the performance comparison Tables and detection illustration from the Figures, we see that the use of AUC and AP gives more stable results across each dataset.

Table 4

Frame-level AUC comparison between State-of-the-art and proposed model on UIT-ADrone dataset.

Supervision	Method	Feature	AUC
Weakly-Supervised	TEVAD [11]	I3D-RGB	63.46
	RTFM [3]	I3D-RGB	62.88
	MGFN [1]	VideoSwin-RGB	64.25
	MSL [43]	VideoSwin-RGB	63.66
Unsupervised	ANDT [4]	ViT	60.50
	ASTT [44]	ViT	65.45
Weakly-Supervised	MMVAD*	I3D-RGB (RNet50)	67.40
	MMVAD	I3D-RGB (CSPDNet53)	69.56

Performance of Drone-Anomaly Dataset: We present the quantitative performance of our model for the drone anomaly dataset in Table 3. In Table 3, we use six latest SOTA models to validate the superior performance of our model. We only have video-level labels for the drone anomaly video, so we use AUC as the performance metric. In Table 3, we see that the baseline TEVAD achieved 67.35% with multimodal RTFM loss and outperformed the single-modal RTFM model by nearly 2%. The optimal performance from the existing model was found from MGFN by rejecting the idea of using the RTFM loss function, which achieved 68.92% of AUC and became our closest competitor. We present the result of our model based on two feature extractors; in one, we show the outcome from visual features extracted from I3D-RGB with ResNet50, and in the other, with CSP-DarkNet53 as the core extractor. With **MMVAD***, we achieved 70.78% AUC and beat the closest competitor by 1.86%. On the other hand, with **MMVAD**, we achieve the optimal AUC of 71.22% and beat the MGFN with a margin of 2.3%. The qualitative result is presented in Figure 11, where we show the video frames with their Ground Truth (GT) and the anomaly score probability. Figure 11 (a) shows the scene detection with high confidence due to a clear anomaly case in the frames. However, in Figure 11 (b), we present some challenging cases where the model fails. This low-confidence detection is because the difference between the normal and abnormal scenes is very subtle, and the object that makes the difference is very small. However, our model was able to detect the second frame with a reasonable average score of 0.7. Figure 11 (a) and (b) explain and present some factors that make the drone VAD task more challenging than the CCTV VAD tasks.

Performance of UIT-ADrone Dataset: This is our second drone dataset and the most challenging drone VAD dataset publicly available. We use the same set of SOTA methods for comparisons. The baseline TEVAD achieves an AUC of 63.46%, whereas the RTFM achieves 62.88%. In this dataset, the previous SOTA winner MGFN achieves 64.25% of AUC and falls slightly behind the SOTA winner ASTT model. The ASTT method is designed for drone VAD

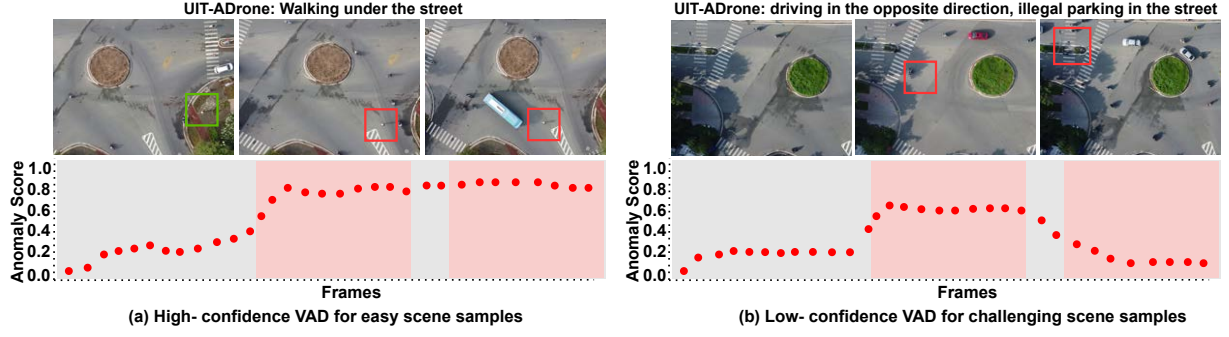


Figure 12: Frame-level anomaly detection with anomaly score using MMVAD model for UIT-ADrone dataset.

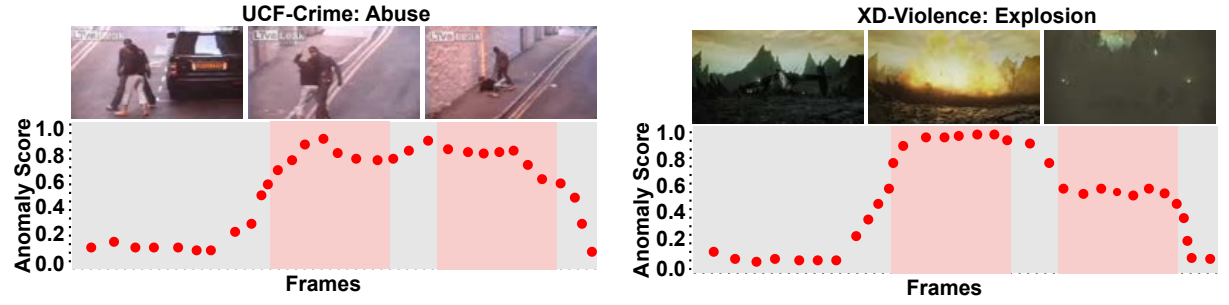


Figure 13: Frame-level anomaly detection with anomaly score using MMVAD model for UCF-Crime and XD-Violence datasets.

Table 5

Frame-level AUC comparison between State-of-the-art and proposed model on UCF-Crime dataset.

Supervision	Method	Feature	AUC
Weakly-Supervised	TEVAD [11]	I3D-RGB	84.90
	RTFM [3]	I3D-RGB	84.31
	MGFN [1]	VideoSwin-RGB	86.67
	MSL [43]	VideoSwin-RGB	85.62
Unsupervised	ANDT [4]	ViT	84.63
	ASTT [44]	ViT	85.20
Weakly-Supervised	MMVAD*	I3D-RGB (RNet50)	87.78
	MMVAD	I3D-RGB (CSPDNet53)	88.20

tasks and achieves 65.45% AUC. Our proposed **MMVAD*** and **MMVAD** model achieves AUC of 67.40% and 69.56%, respectively, and outperforms the ASTT model by 3.15% and 5.31%. From Figure 12, we can verify the qualitative performance of our model for easy and complex scenes. Figure 12 (a) shows the abnormality detection for "Walking under the street"; it is clear from the visualization that the pivotal object is very unclear to detect and small. Also, Figure 12 (b) presents the challenging case due to a small object, a show of the critical object, occupation, etc. All of these issues make the UIT-ADrone a very challenging VAD

dataset. However, our model performs satisfactorily on easy and semi-hard abnormal scenes.

Performance of UCF-Crime Dataset: To prove the cross-domain effectiveness of our model, we also use low-altitude CCTV videos for VAD experiments. Table 5 demonstrates the performance of the MMVAD model on the largest UCF-Crime benchmark dataset. Our model outperforms all the other SOTA methods by a minimum margin of 1.53% AUC. Compared to the drone videos, the performance of each SOTA model is entirely satisfactory in CCTV videos. The baseline TEVAD method achieves 84.90% of AUC with multimodal RTFM loss, outperforming the original RTFM model very closely. The MSL and ASTT model performs relatively the same, with an AUC of 85.62% and 85.20%, respectively. The highest SOTA result from existing models comes from the MGFN model with 86.67% AUC. However, our **MMVAD** outperforms all existing methods with an AUC of 88.20% by using multimodal data and contrastive loss instead of the previous RTFM loss. Figure 13 (a) illustrates the frame-level anomaly detection from the UCF-Crime dataset for *Abuse* scene. The detection confidence is very high in the dataset because of the less complex scene than in the drone datasets. We were able to detect crime-scene anomalies from UCF-crime with an average confidence of over 80% due to the robust complex-scene capture characteristics of our **MMVAD** model.

Performance of XD-Violence Dataset: Among the CCTV benchmark datasets, XD-Violence proved to be more

Table 6

Frame-level AUC comparison between State-of-the-art and proposed model on XD-Violence dataset.

Supervision	Method	Feature	AP
Weakly-Supervised	TEVAD [11]	I3D-RGB	79.84
	RTFM [3]	I3D-RGB	77.80
	MGFN [1]	VideoSwin-RGB	80.11
	MSL [43]	VideoSwin-RGB	78.58
Unsupervised	ANDT [4]	ViT	77.62
	ASTT [44]	ViT	78.45
Weakly-Supervised	MMVAD*	I3D-RGB (RNet50)	80.84
	MMVAD	I3D-RGB (CSPDNet53)	81.98

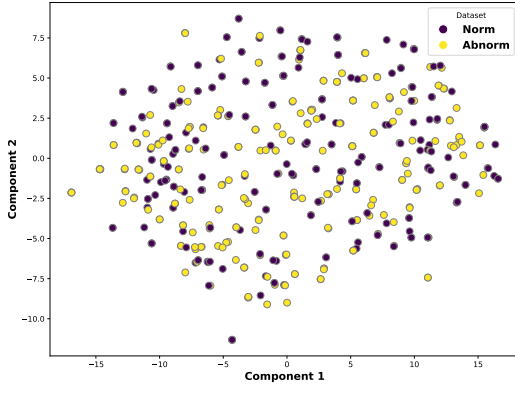
challenging than existing others. It was observed from existing SOTA methods that the transformer-based models are more successful than deep CNNs. The improved performance due to the transformer can be proved by the outcome of MGFN, ASTT, and MSL methods. We have frame-level annotation for the XD-Violence, so we use AP as the performance metric. Throughout all experiments, we found the baseline TEVAD model gives sub-optimal results due to the linear-fusion technique and incorrect feature learning method. The TEVAD achieves an AP of 79.84% for the XD-violence dataset, whereas single-modal RTFM provides 77.80%. The AP from transformer-based MGFN, MSL, and ASTT are respectively 80.11%, 78.58%, and 78.45%, and MGFN achieves the highest SOTA result among the existing methods. Our proposed **MMVAD** also performs quite satisfactorily compared to the existing SOTA methods. We gain 81.98% of AP with a +2.14% gain compared to the baseline TEVAD method. We use an *Explosion* scene from the XD-Violence dataset to illustrate the qualitative performance. The normal scene/ True Negative scene from the first frame was detected with very high confidence. The abnormal scene/ True Positive *Explosion* scene is also very accurately detected with $\sim 100\%$ confidence. However, we found some low-confidence detection when we have unclear scenes such as *smoke after the explosion*. We believe this sub-optimal detection is due to the similarity of the situation with the typical background scene.

4.5. Ablation Study

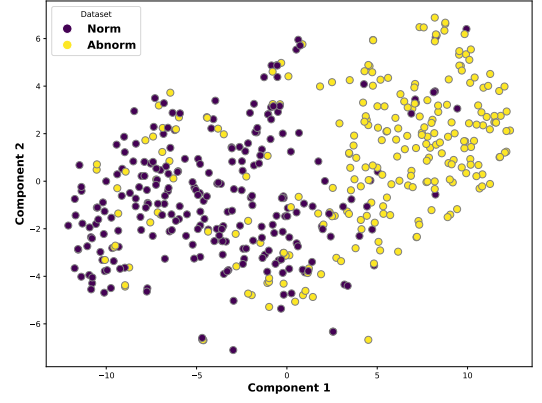
In this section, we answer several questions regarding the effectiveness of different modules in our MMVAD architecture. The first question is: Does ALSVS help to reduce noise due to improper snippet creation? From Figure 2, we noticed a fixed scale stride for frame segmentation can lead to erroneous snippets. The snippets are the core input for the whole architecture, and overall architecture's success relies on good snippet creation. The creation of noisy snippets would lead to suboptimal performance. It is evident from Table 7 that ALSVS indeed has some effect of increased AUC and AP for the experimental datasets. While training with the Baseline+ALSVS module, we received a performance gain

of +1.53%, +0.3%, +0.85%, and +0.36%, respectively, for DA, UIT, UCF, and XD datasets compared to the baseline model. The second question is: is attention-based fusion more effective than linear fusion? We answer the question by comparing the baseline results with the baseline+AFF performance. In Table 7, we show that attention-based fusion can achieve up to +2.11% increase in AUC and AP in cross-domain situations. We also compare the performance of different linear fusions with the AFF method in Table 9 to answer the question: Does linear feature fusion have performance issues because of domain misalignment or aliasing effect? A quantitative answer to this question is presented in Table 9; we see the concatenation method for feature fusion is at least $\sim 1.5\%$ lower in AUC and AP compared to the AFF technique. The other two methods of linear fusion are addition and product, which are also not optimal in performance. The elementwise addition in the multimodal feature proved to be the least effective method for the VAD task. In contrast, the element-wise product shows a very close performance with the concatenation method. The goal of the AFF is to propagate the textual information of each snippet in the critical visual snippet embeddings without making any elementwise operation in visual embeddings. By this, we avoid any unwanted misalignment issues during the fusion process and gain optimal performance.

The effective use of Power enhance multimodal feature (PEMF) is also proved in Table 7; we see it also contributes toward improved performance across datasets. We notice an average of +0.7% performance gain by using the PEMF module. However, the effectiveness of PEMF is not limited to this module itself because we use the power-enhanced topK features for later classification and contrastive learning. So, the effectiveness of this module propagates to the later part of the pipeline. Now, the fourth question is: should we use CL instead of magnitude-based learning for VAD? As shown in Figure 14 and 15, we observed promising results when using CL instead of RTFM. We were able to increase the AUC and AP of each dataset by nearly +2.0% compared to the baseline RTFM. In this work, we use customized CL with multiple positive and negative cases [7] for more robust learning. As presented in Table 8, we tried several combinations of positive and negative cases for CL and achieved the best performance with 8 negative and 4 positive cases. We also observed that increasing the positive case more than the negative cases slightly decreases the performance. The effect of CL is presented in Figure 14 and 15, where we prove the effectiveness of CL on two high-altitude VAD datasets. Here, we plot nearly 500 video snippet features from AFF on the test sets. The comparison shows CL gives much better clustering than the previous magnitude-based learning for UAV datasets. The use of CL helped us gain optimal performance on the drone datasets where the difference between normal and abnormal scenes is very subtle. From Table 3 and 4, we observed that magnitude-based learning suffers when dealing with drone datasets as the scene difference is very subtle and cannot maintain clear separability for hard examples.

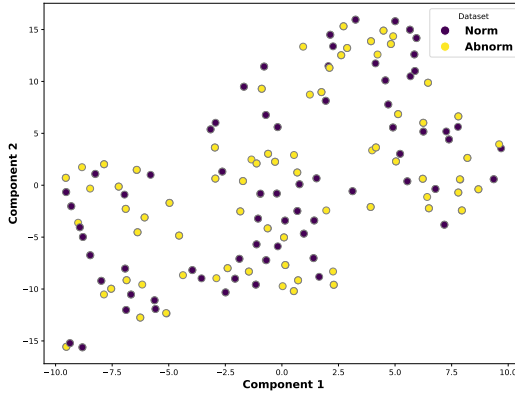


(a) Yellow circles represent the Anomaly frames, and blue circles represent the Normal frames before using contrastive learning.

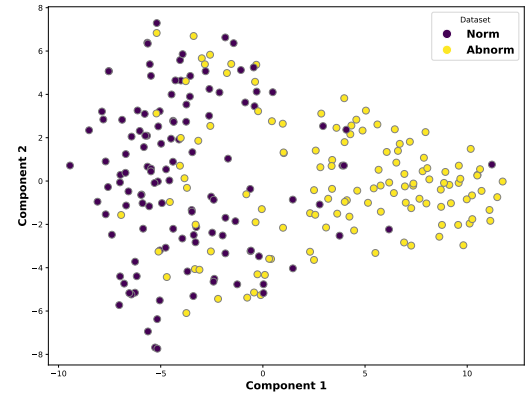


(b) Yellow circles represent the Anomaly frames, and blue circles represent the Normal frames after using contrastive learning.

Figure 14: Illustration of the effect of contrastive learning on UIT-ADrone test-set. Here, we plot 500 snippet features from the UITADrone test-set. For point labeling, we use the frame-level GT. It is evident from the figure that the features are much better grouped after using contrastive learning.



(a) Yellow circles represent the Anomaly frames, and blue circles represent the Normal frames before using contrastive learning.



(b) Yellow circles represent the Anomaly frames, and blue circles represent the Normal frames after using contrastive learning.

Figure 15: Illustration of the effect of contrastive learning on Drone-Anomaly test-set. Here, due to the small test-set size, we plot 300 snippet features from the Drone Anomaly dataset. For point labeling, we use the frame-level GT. It is evident from the figure that the features are much better grouped after using contrastive learning.

Table 7

Ablation study for different modules of our **MMVAD** method. Here, ALSVS= Adaptive Long-Short-Term Video Segmentation, AFF= Attention-based Feature Fusion, PEMF= Power-Enhanced Multimodal Feature, and CL= Contrastive Learning.

Method	ALSVS AFF PEMF CL SVM					Drone-Anomaly UIT-ADrone UCF-Crime XD-Violence			
						AUC	AUC	AUC	AP
Baseline w/(CSPDNet53)						67.35	63.46	84.90	79.84
w/ALSVS	✓					68.88	63.76	85.75	80.20
w/AFF		✓				69.46	64.65	86.38	80.58
w/PEMF			✓			68.30	64.11	85.25	80.22
w/CL				✓		69.53	65.95	86.70	80.69
w/SVM					✓	68.10	64.80	85.37	80.40
MMVAD $W_c, W_h = 0.5, 0.5$	✓	✓	✓	✓	✓	69.10	66.35	87.08	79.58
MMVAD $W_c, W_h = 0.3, 0.7$	✓	✓	✓	✓	✓	71.22	69.56	88.20	81.98

Table 8

Ablation study for AUC from **MMVAD** model with various negative and positive case values on contrastive learning.

# neg	# pos	Drone- Anomaly AUC	UIT- Adrone AUC	UCF- Crime AUC	XD- Violence AP
1	1	68.80	66.75	86.25	80.80
4	2	68.84	69.56	87.69	81.36
2	4	70.03	68.74	87.14	81.45
8	4	71.22	68.82	88.20	81.98

Table 9

Ablation study for AUC from **MMVAD** model with various fusion techniques.

Method	Fusion Technique	Drone- Anomaly AUC	UIT- Adrone AUC	UCF- Crime AUC	XD- Violence AP
MMVAD	Concat	68.86	68.70	86.85	80.01
	Add	67.89	67.05	85.30	78.63
	Product	68.16	67.39	86.50	78.75
	AFF	71.22	69.56	88.20	81.98

The SVM comes into play when we need significant margin separation near the decision boundaries. As the drone videos are full of confusing samples, we decided to use SVM instead of BCE as the classifier. The performance of SVM is presented in Table 7; we see that SVM increases the AUC and AP by 0.75%, 1.34%, 0.47%, and 0.56%, respectively, for the Drone, UIT, UCF, and XD datasets. Finally, We put weights W_c and W_h in CL and SVM loss, respectively, to control the effect of each learning process. As the features at the beginning are not very accurate, putting more weight on CL shows an adverse effect on overall performance. Hence, we put more focus on binary classification ($W_h = 0.7$) and less weight on contrastive learning ($W_c = 0.3$) and reach the optimal performance gain.

5. Conclusion

In this work, we proposed a multimodal video anomaly detection model (MMVAD) for cross-domain datasets. Our goal was to propose a model that can perform satisfactorily in both high-altitude drone and low-altitude CCTV datasets. From the dataset sample analysis, we found that the drone videos are far more challenging for VAD than the CCTV videos. To reach our goal, we proposed several novel modules, such as ALSVS, AFF, and PEMF. We argued that using fixed-scale strides for frame segmentation is noisy and may lead to inferior learning. To solve this, we used multi-scale adaptive segmentation. Next, to reduce the misalignment issues from existing linear fusion methods, we propose attention-based feature fusion (AFF). Our other main argument was to reject the idea of using magnitude-based RTFM loss for feature learning. The idea of high

feature magnitude representing anomalous activities is not always correct. To resolve that issue, we use power-enhanced features that propagate the overall snippet power of a video throughout each snippet and amplify the signal. Next, we use the amplified signals for binary classification and contrastive learning. Finally, through the ablation study, we show the effectiveness of each module. The contrastive learning helped us to significantly increase the performance for each dataset. We gained at least +1.75% AUC/AP for all datasets with the SVM and CL modules. The outcome of the MMVAD model is 71.22%, 69.56%, 88.20%, and 81.98%, respectively, for Drone, UIT, UCF, and XD datasets, which is roughly ~3.5% higher than the baseline TEVAD performance.

Funding

The NVIDIA RTX 6000 GPU used for this research was donated by NVIDIA Corporation. NAVAIR SBIR N68335-18-C-0199 partially supports this work. This article's views, opinions, and findings are those of the authors. They should not be interpreted as representing the official views or policies of the Department of Defense or the US government.

References

- [1] Y. Chen, Z. Liu, B. Zhang, W. Fok, X. Qi, Y.-C. Wu, Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 387–395.
- [2] H. Karim, K. Doshi, Y. Yilmaz, Real-time weakly supervised video anomaly detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 6848–6856.
- [3] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, G. Carneiro, Weakly-supervised video anomaly detection with robust temporal feature magnitude learning, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 4975–4986.
- [4] P. Jin, L. Mou, G.-S. Xia, X. X. Zhu, Anomaly detection in aerial videos with transformers, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–13.
- [5] T. M. Tran, T. N. Vu, T. V. Nguyen, K. Nguyen, Uit-adrone: A novel drone dataset for traffic anomaly detection, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2023).
- [6] J. Yan, J. Liu, D. Liang, Y. Wang, J. Li, L. Wang, Semantic segmentation of land cover in urban areas by fusing multi-source satellite image time series, IEEE Transactions on Geoscience and Remote Sensing (2023).
- [7] D. Biswas, J. Tešić, Unsupervised domain adaptation with debiased contrastive learning and support-set guided pseudo labeling for remote sensing images, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2024).
- [8] M. Bonetto, P. Korshunov, G. Ramponi, T. Ebrahimi, Privacy in mini-drone based video surveillance, in: 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG), Vol. 4, IEEE, 2015, pp. 1–6.
- [9] R. Jiao, Y. Wan, F. Poiesi, Y. Wang, Survey on video anomaly detection in dynamic scenes with moving cameras, Artificial Intelligence Review 56 (Suppl 3) (2023) 3515–3570.
- [10] B. Ramachandra, M. J. Jones, R. R. Vatsavai, A survey of single-scene video anomaly detection, IEEE transactions on pattern analysis and machine intelligence 44 (5) (2020) 2293–2312.
- [11] W. Chen, K. T. Ma, Z. J. Yew, M. Hur, D. A.-A. Khoo, Tevad: Improved video anomaly detection with captions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5548–5558.

- [12] P. H. Seo, A. Nagrani, A. Arnab, C. Schmid, End-to-end generative pretraining for multimodal video captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17959–17968.
- [13] N. Jaafar, Z. Lachiri, Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance, *Expert Systems with Applications* 211 (2023) 118523.
- [14] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.
- [15] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.
- [16] J.-C. Feng, F.-T. Hong, W.-S. Zheng, Mist: Multiple instance self-training framework for video anomaly detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 14009–14018.
- [17] D. Biswas, J. Tešić, Domain adaptation with contrastive learning for object detection in satellite imagery, *IEEE Transactions on Geoscience and Remote Sensing* (2024).
- [18] W. Sultani, C. Chen, M. Shah, Real-world anomaly detection in surveillance videos, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6479–6488.
- [19] J. Zhang, L. Qing, J. Miao, Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 4030–4034.
- [20] H. Lv, Z. Yue, Q. Sun, B. Luo, Z. Cui, H. Zhang, Unbiased multiple instance learning for weakly supervised video anomaly detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 8022–8031.
- [21] Y. Hatae, Q. Yang, M. F. Fadrimiratno, Y. Li, T. Matsukawa, E. Suzuki, Detecting anomalous regions from an image based on deep captioning, in: VISIGRAPP (5: VISAPP), 2020, pp. 326–335.
- [22] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, J. Chanussot, Multimodal fusion transformer for remote sensing image classification, *IEEE Transactions on Geoscience and Remote Sensing* (2023).
- [23] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, T. Duerig, Scaling up visual and vision-language representation learning with noisy text supervision, in: International conference on machine learning, PMLR, 2021, pp. 4904–4916.
- [24] J. Du, J. Jin, J. Zhuang, C. Zhang, Hierarchical graph contrastive learning of local and global presentation for multimodal sentiment analysis, *Scientific Reports* 14 (1) (2024) 5335.
- [25] Y. Yang, A. Panagopoulou, S. Zhou, D. Jin, C. Callison-Burch, M. Yatskar, Language in a bottle: Language model guided concept bottlenecks for interpretable image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19187–19197.
- [26] P. Wu, X. Zhou, G. Pang, L. Zhou, Q. Yan, P. Wang, Y. Zhang, Vadclip: Adapting vision-language models for weakly supervised video anomaly detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 6074–6082.
- [27] K. Lin, L. Li, C.-C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, L. Wang, Swinbert: End-to-end transformers with sparse attention for video captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17949–17958.
- [28] M. Guillaumin, J. Verbeek, C. Schmid, Multimodal semi-supervised learning for image classification, in: 2010 IEEE Computer society conference on computer vision and pattern recognition, IEEE, 2010, pp. 902–909.
- [29] Y. Pang, Z. Ma, Y. Yuan, X. Li, K. Wang, Multimodal learning for multi-label image classification, in: 2011 18th IEEE International Conference on Image Processing, IEEE, 2011, pp. 1797–1800.
- [30] X. Wu, D. Hong, J. Chanussot, Convolutional neural networks for multimodal remote sensing data classification, *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021) 1–10.
- [31] S. K. Roy, A. Deria, C. Shah, J. M. Haut, Q. Du, A. Plaza, Spectral-spatial morphological attention transformer for hyperspectral image classification, *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023) 1–15.
- [32] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, *arXiv preprint arXiv:1606.01847* (2016).
- [33] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, B.-T. Zhang, Hadamard product for low-rank bilinear pooling, *arXiv preprint arXiv:1610.04325* (2016).
- [34] Z. Yu, J. Yu, J. Fan, D. Tao, Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 1821–1830.
- [35] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), Vol. 2, IEEE, 2006, pp. 1735–1742.
- [36] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, *arXiv preprint arXiv:1807.03748* (2018).
- [37] D. Li, J. Li, H. Li, J. C. Niebles, S. C. Hoi, Align and prompt: Video-and-language pre-training with entity prompts, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4953–4963.
- [38] D. Li, C. Xu, X. Yu, K. Zhang, B. Swift, H. Suominen, H. Li, Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation, *Advances in Neural Information Processing Systems* 33 (2020) 12034–12045.
- [39] T. Gopalakrishnan, N. Wason, R. J. Krishna, N. Krishnaraj, Comparative analysis of fine-tuning i3d and slowfast networks for action recognition in surveillance videos, *Engineering Proceedings* 59 (1) (2024) 203.
- [40] C. Liu, X. Xu, Y. Zhang, Temporal attention network for action proposal, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 2281–2285.
- [41] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.
- [42] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, Z. Yang, 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, Springer, 2020, pp. 322–339.
- [43] S. Li, F. Liu, L. Jiao, Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 1395–1403.
- [44] T. M. Tran, D. C. Bui, T. V. Nguyen, K. Nguyen, Transformer-based spatio-temporal unsupervised traffic anomaly detection in aerial videos, *IEEE Transactions on Circuits and Systems for Video Technology* (2024).
- [45] D. Fourure, M. U. Javaid, N. Posocco, S. Tihon, Anomaly detection: how to artificially increase your f1-score with a biased evaluation protocol, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2021, pp. 3–18.