Text-Driven Traffic Anomaly Detection with Temporal High-Frequency Modeling in Driving Videos

Rongqin Liang, *Student Member, IEEE,* Yuanman Li, *Senior Member, IEEE,* Jiantao Zhou, *Senior Member, IEEE,* and Xia Li, *Member, IEEE*

Abstract—Traffic anomaly detection (TAD) in driving videos is critical for ensuring the safety of autonomous driving and advanced driver assistance systems. Previous single-stage TAD methods primarily rely on frame prediction, making them vulnerable to interference from dynamic backgrounds induced by the rapid movement of the dashboard camera. While two-stage TAD methods appear to be a natural solution to mitigate such interference by pre-extracting background-independent features (such as bounding boxes and optical flow) using perceptual algorithms, they are susceptible to the performance of first-stage perceptual algorithms and may result in error propagation. In this paper, we introduce TTHF, a novel single-stage method aligning video clips with text prompts, offering a new perspective on traffic anomaly detection. Unlike previous approaches, the supervised signal of our method is derived from languages rather than orthogonal one-hot vectors, providing a more comprehensive representation. Further, concerning visual representation, we propose to model the high frequency of driving videos in the temporal domain. This modeling captures the dynamic changes of driving scenes, enhances the perception of driving behavior, and significantly improves the detection of traffic anomalies. In addition, to better perceive various types of traffic anomalies, we carefully design an attentive anomaly focusing mechanism that visually and linguistically guides the model to adaptively focus on the visual context of interest, thereby facilitating the detection of traffic anomalies. It is shown that our proposed TTHF achieves promising performance, outperforming state-of-the-art competitors by +5.4% AUC on the DoTA dataset and achieving high generalization on the DADA dataset.

Index Terms—Traffic anomaly detection, multi-modality learning, high frequency, attention.

I. INTRODUCTION

TRAFFIC anomaly detection (TAD) in driving videos is a crucial component of automated driving systems [1], [2]
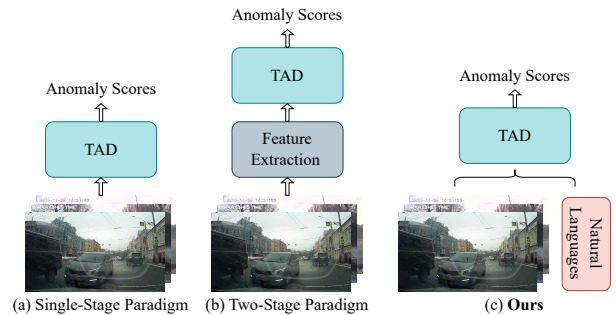
Fig. 1. Existing TAD approaches of single-stage paradigm (a) and two-stage paradigm (b) vs. the proposed TTHF framework (c). Existing single-stage approaches mainly rely on frame prediction, which is difficult to adapt to detecting traffic anomalies with a dynamic background, while the two-stage TAD approaches are vulnerable to the performance of the first-stage perceptual algorithms. The proposed TTHF framework is text-driven and focuses on capturing dynamic changes in driving scenes through modeling temporal high frequency to facilitate traffic anomaly detection.

and advanced driver assistance systems [3], [4]. It is designed to detect anomalous traffic behavior from the first-person driving perspective. Accurate detection of traffic anomalies helps improve road safety, shorten traffic recovery times, and reduce the number of regrettable daily traffic accidents.

Given the significance of traffic anomaly detection, scholars are actively involved in this field and have proposed constructive research [5]–[9]. We observe that these works on TAD can be mainly divided into the single-stage paradigm [6], [10], [11] and the two-stage paradigm [8], [9], [12]. As shown in Fig. 1, previous TAD methods mainly embrace a single-stage paradigm, exemplified by frame prediction [6] and reconstruction-based [11] TAD approaches. Nevertheless, these methods are subject to the dynamic backgrounds caused by the rapid movement of the dashboard camera and have limited accuracy in detecting traffic anomalies. To confront the challenges posed by dynamic backgrounds, researchers have advocated for TAD methods [8], [9], [12] that utilize a two-stage paradigm. These two-stage approaches first extract features such as optical flow, bounding boxes, or tracking IDs from video frames using existing visual perception algorithms, and then propose a TAD model for detecting traffic anomalies. While these approaches have laid the foundation for TAD in driving videos, they are susceptible to the performance

of the first-stage visual perception algorithm, which may cause error propagation, resulting in false detection or missing traffic anomalies. Therefore, in this paper, we strive to explore an effective single-stage paradigm-based approach for traffic anomaly detection in driving videos.

Recently, large-scale visual language pre-training models [13]–[15] have achieved remarkable results by utilizing language knowledge to assist with visual tasks. Among them, CLIP [13] stands out for its exceptional transferability through the alignment of image-text semantics and has demonstrated outstanding capabilities across various computer vision tasks such as object detection [16], semantic segmentation [17], and video retrieval [18]. The success of image-text alignment techniques can be attributed to their ability to map the natural languages associated with an image into high-dimensional non-orthogonal vectors. This is in contrast to traditional supervised methods that map predefined labels to low-dimensional one-hot vectors. Compared to the low-dimensional one-hot vectors, these high-dimensional vectors offer more comprehensive representations to guide the network training. Motivated by this, we endeavor to investigate a language-guided approach for detecting traffic anomalies in driving videos. Intuitively, the transition of CLIP from image-text alignment to video-text alignment primarily involves the consideration of modeling temporal dimensions. Despite the exploration of various methods [19]–[22] for temporal modeling, encompassing various techniques such as *Average Pooling*, *Conv1D*, *LSTM*, *Transformer*, the existing approaches predominantly concentrate on aggregating visual context along the temporal dimension. In the context of traffic anomaly detection for driving videos, we emphasize that beyond the visual context, characterizing dynamic changes in the driving scene along the temporal dimension proves advantageous in determining abnormal driving behavior. For instance, traffic events such as vehicle collisions or loss of control often result in significant and rapid alterations in the driving scene. Therefore, *how to effectively characterize the dynamic changes of driving scenes holds paramount importance for traffic anomaly detection in driving videos*.

Additionally, considering that different types of traffic anomalies exhibit unique characteristics, a straightforward encoding of the entire driving scene may diminish the discriminability of driving events and impede the detection of diverse traffic anomalies. For instance, traffic anomalies involving the ego-vehicle are often accompanied by global jittering of the dashboard camera, while anomalies involving non-ego vehicles often lead to local anomalies in the driving scene. Consequently, *how to better perceive various types of traffic anomalies proves crucial for traffic anomaly detection*.

In this work, we propose a novel traffic anomaly detection approach: **T**ext-Driven Traffic Anomaly Detection with **T**emporal **H**igh-**F**requency Modeling (TTHF), as shown in Fig. 2. To represent driving videos comprehensively, our fundamental idea is to not only capture the spatial visual context but also emphasize the depiction of dynamic changes in the driving scenes, thereby enhancing the visual representation of driving videos. Specifically, we initially leverage the pre-trained visual encoder of CLIP, endowed with rich

prior knowledge of visual language semantics, to encode the visual context of driving videos. Then, to capture the dynamic changes in driving scenes, we innovatively introduce temporal high-frequency modeling (THFM) to obtain temporal high frequency representations of driving videos along the temporal dimension. Subsequently, the visual context and temporal high-frequency representations are fused to enhance the overall visual representation of driving videos. To better perceive various types of traffic anomalies, we propose an attentive anomaly focusing mechanism (AAFM) to guide the model to adaptively focus both visually and linguistically on the visual context of interest, thereby facilitating the detection of traffic anomalies.

It is shown that our proposed TTHF model exhibits promising performance on the DoTA dataset [9], outperforming state-of-the-art competitors by +5.4% AUC. Furthermore, without any fine tuning, the AUC performance of TTHF on the DADA dataset [23] demonstrates its generalization capability. The main contributions of our work can be summarized as follows:

1) We introduce a simple yet effective single-stage traffic anomaly detection method that aligns the visual semantics of driving videos with matched textual semantics to identify traffic anomalies. In contrast to previous TAD methods, the supervised signals in our approach are derived from text, offering a more comprehensive representation in high-dimensional space.
2) We emphasize the modeling of high frequency in the temporal domain for driving videos. In contrast to previous approaches that solely aggregate visual context along the temporal dimension, we place additional emphasis on modeling high frequency in the temporal domain. This enables us to characterize dynamic changes in the driving scene over time, thereby significantly enhancing the performance of traffic anomaly detection.
3) We further propose an attentive anomaly focusing mechanism to enhance the perception of various traffic anomalies. Our proposed mechanism guides the model both visually and linguistically to adaptively focus on the visual contexts of interest, facilitating the detection of traffic anomalies.
4) Comprehensive experimental results on public benchmark datasets demonstrate the superiority and robustness of the proposed method. Compared to existing state-of-the-art methods, the proposed TTHF improves AUC by +5.4% on the DoTA dataset and also achieves state-of-the-art AUC on the DADA dataset without any fine-tuning.

The remainder of this paper is organized as follows. Section II gives a brief review of related works. Section III details our proposed TTHF for traffic anomaly detection in driving videos. Extensive experimental results are presented in Section IV, and we finally draw a conclusion in Section V.

II. RELATED WORKS

A. Traffic Anomaly Detection (TAD) in Driving Videos

Traffic anomaly detection (TAD) in driving videos aims to identify abnormal traffic events from the perspective of driving, such as collisions with other vehicles or obstacles,

being out of control, and so on. Such events can be classified into two categories: ego-involved anomalies (*i.e.*, traffic events involving the ego-vehicle) and non-ego anomalies (*i.e.*, traffic events involving observed objects but not the ego-vehicle). A closely related topic to TAD in driving videos is anomaly detection in surveillance videos (VAD), which involves identifying abnormal events such as fights, assaults, thefts, arson, and so forth from a surveillance viewpoint. In recent years, various VAD methods [24]–[29] have been proposed for surveillance videos, which have greatly contributed to the development of this field. However, in contrast to the static background in surveillance videos, the background in driving videos is dynamically changing due to the fast movement of the ego vehicle, which makes the VAD methods prone to failure in the TAD task [9], [12]. Recently, Wang et al. [30] proposed a method for detecting crowd flow anomalies by comparing anomalous samples with normal samples that were estimated based on prototypes. However, crowd flow anomaly detection methods are difficult to apply to the TAD task due to the differences in tasks and the data processed. In this paper, we work on the task of traffic anomaly detection in driving videos to provide a new solution for this community.

Early TAD methods [5], [31] mainly extracted features in a handcrafted manner and utilized a Bayesian model for classification. However, these methods are sensitive to well-designed features and generally lack robustness in dealing with a wide variety of traffic scenarios. With the advances of deep neural networks in computer vision, researchers have proposed deep learning-based approaches for TAD, laying the foundation for this task. Based on our observations, the existing TAD methods can be basically classified into single-stage paradigm [6], [10], [11] and two-stage paradigm [12], [32]–[34].

Previous single-stage paradigm-based TAD approaches mainly comprise frame reconstruction-based and frame prediction-based TAD approaches [6], [10], [11]. These methods used reconstruction or prediction errors of video frames to evaluate traffic anomalies. For instance, Liu et al. [6] predicted video frames of normal traffic events through appearance and motion constraints, thereby helping to identify traffic anomalies that do not conform to expectations. Unfortunately, these methods tend to detect ego-involved anomalies (*e.g.*, out of control) and perform poorly on non-ego traffic anomalies. This is primarily attributed to ego-involved anomalies causing significant shaking of the dashboard camera, leading to substantial global errors in frame reconstruction or prediction. Such errors undoubtedly facilitate anomaly detection. However, the methods based on frame reconstruction or prediction have difficulty distinguishing the local errors caused by the traffic anomalies of other road participants because of the interference of the dynamic background from the fast-moving ego-vehicle. This impairs their ability to detect traffic anomalies.

In recent years, to address the challenges posed by dynamic backgrounds, researchers have proposed applying a two-stage paradigm to the traffic anomaly detection task. In this paradigm, the perception algorithm is initially applied to extract visual features in the first stage. Then, the TAD model utilizes these features to detect traffic anomalies. For instance,

Yao et al. [9], [32] applied Mask-RCNN [35], FlowNet [36], DeepSort [37], and ORBSLAM [38] algorithms to extract bounding boxes (bboxes), optical flow, tracking ids, and ego motion, respectively. Then, they used these visual features to predict the future locations of objects over a short horizon and detected traffic anomalies based on the deviation of the predicted location. Along this line, Fang et al. [12] used optical flow and bboxes as visual features. They attempted to collaborate on frame prediction and future object localization tasks [39] to detect traffic anomalies by analyzing inconsistencies in predicted frames, object locations, and the spatial relation structure of the scene. Zhou et al. [8] obtained bboxes of objects in the scene from potentially abnormal frames as visual features. They then encoded the spatial relationships of the detected objects to determine the abnormality of these frames. Despite the success of the two-stage paradigm TAD methods, they rely on the perception algorithms in the first stage, which may cause error propagation and lead to missed or false detection of traffic anomalies. Different from existing TAD methods, we propose a text-driven single-stage traffic anomaly detection approach that provides a promising solution for this task.

B. Vision-Text Multi-Modality Learning

Recently, there has been a gradual focus on vision-text multi-modal learning. Among them, contrastive language-image pre-training methods have achieved remarkable results in many computer vision tasks such as image classification [13], [14], object detection [16], [40], semantic segmentation [17], [41] and image retrieval [42], [43]. At present, CLIP [13] has become a mainstream visual learning method, which connects visual signals and language semantics by comparing large-scale image-language pairs. Essentially, compared to traditional supervised methods that convert labels into orthogonal one-hot vectors, CLIP provides richer and more comprehensive supervision information by collecting large-scale image-text pairs from web data and mapping the text into high-dimensional supervision signals (usually non-orthogonal). Following this idea, many scholars have applied CLIP to various tasks in the video domain, including video action recognition [19], [44], video retrieval [18], [20], [45], video recognition [46], [47], and so on. For example, ActionCLIP [19] modeled the video action detection task as a video-text matching problem in a multi-modal learning framework and strengthened the video representation with more semantic language supervision to enable the model to perform zero-shot action recognition. More recently, Wu et al. [48] proposed a vision-language model for anomaly detection in surveillance videos. However, as mentioned earlier, traffic anomaly detection faces the problem of dynamic changes in the driving scene, which often makes VAD methods prone to fail in TAD tasks. To the best of our knowledge, there is no effective approach to model traffic anomaly detection task from the perspective of vision-text multi-modal learning. In this paper, we preliminarily explore an effective text-driven method for traffic anomaly detection, which we hope can provide a new perspective on this task.
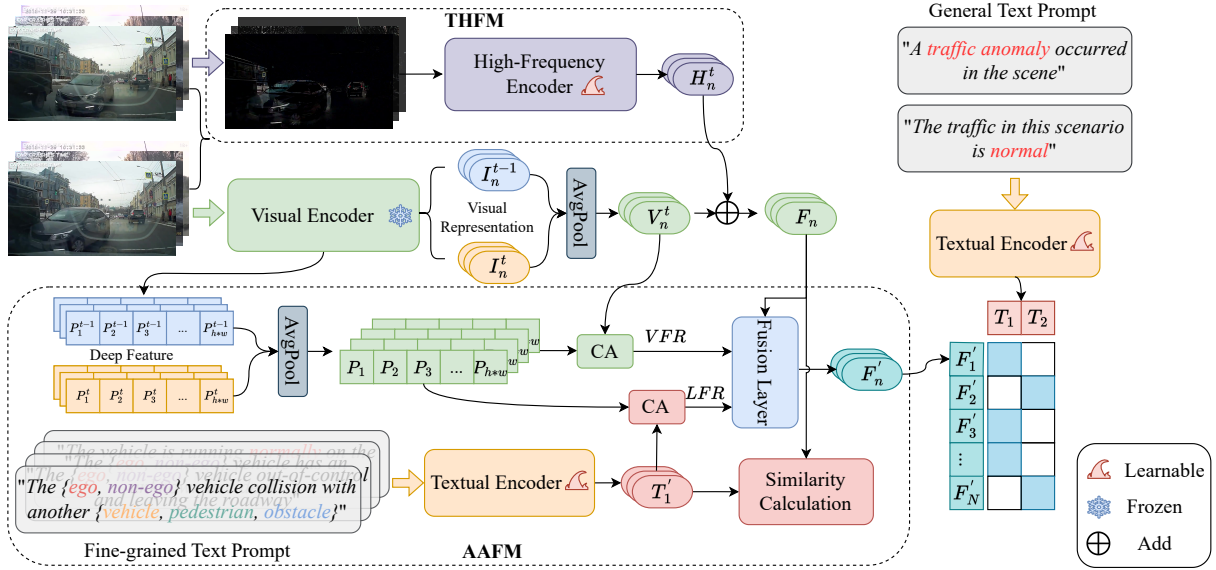
Fig. 2. Overview of our proposed TTHF. It is a CLIP-like framework for traffic anomaly detection. In this framework, we first apply a visual encoder to extract visual representations of driving video clips. Then, we propose Temporal High-Frequency Modeling (THFM) to characterize the dynamic changes of driving scenes and thus construct a more comprehensive representation of driving videos. Finally, we introduce an attentive anomaly focusing mechanism (AAFM) to enhance the perception of various types of traffic anomalies. Besides, for brevity, we denote the cross-attention as CA, the visually focused representation as VFR, and the linguistically focused representation as LFR.

III. THE PROPOSED APPROACH: TTHF

In this section, we mainly introduce the proposed TTHF framework. First, we describe the overall framework of TTHF. Then, we explain two key modules in TTHF, *i.e.*, temporal High-Frequency Modeling (THFM) and attentive anomaly focusing mechanism (AAFM). Moreover, we describe the contrastive learning strategy for cross-modal learning of video-text pairs, and finally show how to perform traffic anomaly detection in our TTHF.

A. Overview of Our TTHF Framework

The overall framework of TTHF is illustrated in Fig. 2. It presents a CLIP-like two-stream framework for traffic anomaly detection. For the visual context representation, considerable research [49]–[51] has demonstrated that CLIP possesses a robust foundation of vision-language prior knowledge. Leveraging this acquired semantic knowledge for anomaly detection in driving videos facilitates the perception and comprehension of driving behavior. Therefore, we advocate applying the pre-trained visual encoder of CLIP to extract visual representations from driving video clips of two consecutive frames. After obtaining the frame representations, we employ *Average Pooling* along the temporal dimension as in previous works [19]–[21] to aggregate these representations to characterize the visual context of the video clip. For the text representation, we first describe normal and abnormal traffic events as text prompts (*i.e.*, a_1 and a_2 in Table I), and then apply the pretrained textual encoder in CLIP to extract text representations.

Intuitively, after extracting the visual and textual representations of driving video clips, we can directly leverage contrastive learning to align them for traffic anomaly detection. However, in our task, solely modeling the visual representation

from visual context is insufficient to capture the dynamic changes in the driving scene. Therefore, we introduce temporal high-frequency modeling (THFM) to characterize the dynamic changes and provide a more comprehensive representation of the driving video clips. Additionally, to better perceive various types of traffic anomalies, we further propose an attentive anomaly focusing mechanism (AAFM) to adaptively focus on the visual context of interest in the driving scene, thereby facilitating the detection of traffic anomalies. In the following sections, we will introduce these two key modules in detail.

B. Temporal High-Frequency Modeling (THFM)

Video-text alignment diverges from image-text alignment by necessitating consideration of temporal characteristics. Numerous methods [19]–[21] have effectively employed CLIP in addressing downstream tasks within the video domain. The modeling strategies adopted in these approaches for the temporal domain encompass various techniques such as *Average Pooling*, *Conv1D*, *LSTM*, and *Transformer*. These strategies primarily emphasize aggregating visual context from distinct video frames along the temporal dimension. Nevertheless, for the anomaly detection task in driving videos, we contend that not only the visual context but also the temporal dynamic changes in the driving scene hold significant importance in modeling driving behavior. For instance, a collision or loss of vehicle control often induces substantial changes in the driving scene within a brief timeframe. Therefore, in our work, we propose to model the visual representation of driving videos in two aspects, *i.e.*, the visual context of video frames in the spatial domain and the dynamic changes of driving scenes in the temporal domain. Considering the fact that the high frequency of the driving video in the temporal domain reflects the dynamic changes of the driving scene. To clarify,
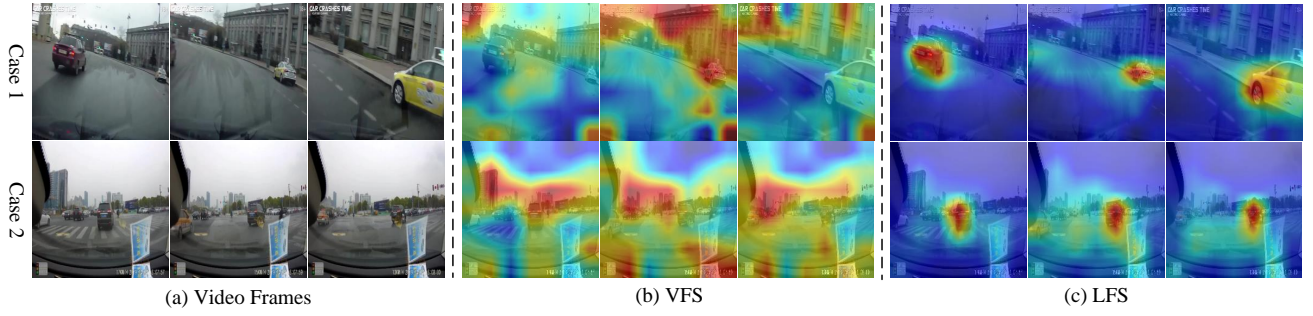
(a) Video Frames   (b) VFS   (c) LFS

Fig. 3. An illustration of the AAFM. The original video frames are displayed in column (a). In column (b), we visualize the attention of the visual representation to the deep features of a video clip under the visually focused strategy (VFS). In column (c), we visualize the attention of the soft text representation to the deep features of a video clip under the linguistically focused strategy (LFS). We present two types of traffic anomaly scenarios. Specifically, case 1 illustrates an instance where the ego-vehicle experiences loss of control while executing a turn. In case 2, the driving vehicle observes a collision between the car turning ahead and the motorcycle traveling straight on the right.

we present several cases in Fig. 4 for illustration. Based on
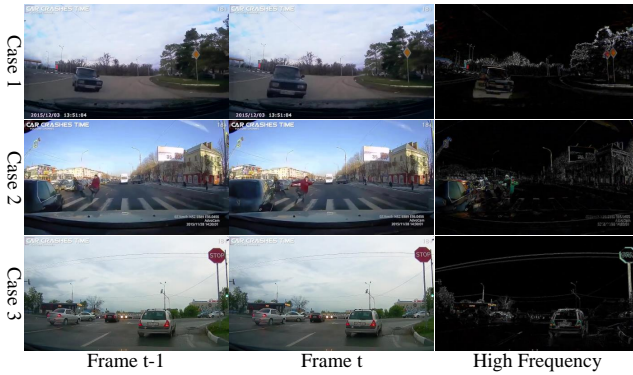

Frame t-1   Frame t   High Frequency

Fig. 4. An illustration of the high frequency. We show 3 cases as examples. The first and second columns correspond to the original consecutive video frames, and the last column is the high-frequency component extracted along the temporal dimension.

the above observations, we introduce the Temporal High Frequency Modeling (THFM) to enhance the visual representation of the driving video within the temporal-spatial domain.

Our fundamental idea involves utilizing the high frequency presented in the temporal domain of the driving video to characterize dynamic changes. Specifically, we first extract the high frequency of the driving video clip in the temporal dimension, which is formulated as:

$$I_n^{hp} = HP(frame_n^t, frame_n^{t-1}), \quad (1)$$

where $HP(\cdot)$ is the difference operation to extract high frequency I_n^{hp} along the temporal dimension from two consecutive frames $t-1$ and t of the n-th driving video clip. Further, we encode I_n^{hp} to the high-frequency representation by

$$H_n^t = F_{hf}(I_n^{hp}; \xi_{hf}), \quad (2)$$

where $F_{hf}(\cdot)$ represents the high-frequency encoder, sharing the same architecture as the visual encoder (i.e., ResNet50 unless specified otherwise). The resultant high-frequency representation is denoted as H_n^t. Finally, to obtain the visual representation of the driving video clip in the spatio-temporal domain, we fuse the spatial visual context representation

with the temporal high-frequency representation H_n^t, which is expressed as follows:

$$
\begin{aligned}
I_n^t, I_n^{t-1} &= F_{ve}(frame_n^t, frame_n^{t-1}; \xi_{ve}), \\
V_n^t &= AvgPool(I_n^t, I_n^{t-1}), \\
F_n &= V_n^t + H_n^t,
\end{aligned}
\quad (3)
$$

where F_{ve} is the visual encoder with frozen pre-trained parameter ξ_{ve}, I_n^t and I_n^{t-1} represent visual representations of frame t and $t-1$, respectively, and V_n^t denotes the spatial visual context representation after *Average Pooling*. Here, $F_n \in \mathbb{R}^{1 \times C}$ is the fused visual representation, where C denotes the feature dimension. The fused visual representation F_n not only models the visual context of driving video clips, but also characterizes the dynamic changes in the temporal dimension, which is beneficial for perception and understanding driving behaviors.

C. Attentive Anomaly Focusing Mechanism

Different types of traffic anomalies tend to exhibit distinct characteristics. For instance, anomalies involving the ego vehicle are often accompanied by global jitter from the dashboard camera, whereas anomalies involving non-ego vehicles typically cause anomalies in local regions of the driving scene. Blindly encoding the entire driving scene may reduce the discriminability of driving events and impede the ability to detect various types of traffic anomalies. Therefore, adaptively focusing on the visual context of interest is critical to perceiving different types of traffic anomalies.

In our work, we propose an attentive anomaly focusing mechanism (AAFM). The fundamental idea is to decouple the visual context visually and linguistically, to guide the model to adaptively focus on the visual content of interest. Specifically, we carefully design two focusing strategies: the *visually focused strategy (VFS)* and the *linguistically focused strategy (LFS)*. The former utilizes visual representations with global context to concentrate on the most semantically relevant visual context, while the latter adaptively focuses on visual contexts that are most relevant to text prompts through the guidance of language.

1) Visually Focused strategy (VFS): In fact, the spatial visual representation inherently captures the global context. Utilizing the attention of visual representation towards the

deep features of various regions in the driving scene enables a focus on the most semantically relevant visual content. Specifically, as shown in Fig. 2, we focus on and weight the deep features of interest by using cross-attention (CA) on the spatial visual context representation V_n^t and deep features of the video clip, which can be written as:

$$VFR_n = softmax\left(\frac{Q\left(V_n^t\right) \cdot K^\top\left(P\right)}{c}\right) \cdot V\left(P\right), \quad (4)$$

where Q, K and V are linear transformation, $P \in \mathbb{R}^{h*w \times C}$ is the deep feature map of the video clip, (h, w) represents the size of the feature map, and c is the scaling factor which refers to the rooted square of feature dimension. Note that, for transformer-based visual encoders, V_n^t is represented by the class token, and P is represented by the patch tokens. $VFR_n \in \mathbb{R}^{1 \times C}$ denotes the visually focused representation of the n-th video clip. Since the spatial visual representation encodes global context, focusing on its most relevant visual content helps guide the model to perceive the semantics of the driving scene. As shown in Fig. 3 (b), our VFS can adaptively focus on the crucial scene semantics in the driving scene. Such attention helps to detect traffic anomalies involving the ego-vehicle, especially the loss of control of the ego vehicle (case 1 in Fig. 3).

2) linguistically focused strategy (LFS): Intuitively, the fine-grained text prompts clearly define the subjects, objects, and traffic types involved in the traffic events. In contrast to general text prompts (as listed in a_1 and a_2 in Table I), utilizing fine-grained text prompts helps guide the model to focus on relevant visual contexts, thereby improving the comprehension of various traffic anomalies. Therefore, to facilitate the model's adaptive perception of relevant visual context, we further design a linguistically focused strategy. The core idea is to utilize the carefully designed fine-grained text prompts (as listed in b_1 to b_4 in Table I) to guide the model to adaptively focus on the visual context of interest, thereby enhancing the understanding of traffic anomalies.

Specifically, first, we categorize traffic events into four groups based on their types. Second, we further categorize each type of traffic event according to the different subjects (*i.e.*, ego or non-ego vehicle) and objects (*i.e.*, vehicle, pedestrian, or obstacle) involved. Finally, we define a total of 11 types of fine-grained text prompts, as summarized in Table I from $b1$ to $b4$. Note that the DoTA dataset used in our experiments is annotated with 9 types of traffic anomalies, as shown in Table II, with each anomaly encompassing both ego-involved and non-ego traffic anomalies. With the defined fine-grained text prompts, we apply the textual encoder in CLIP to extract the fine-grained text representation as follows:

$$T_m^{'} = F_{te}(t_m; \xi_{te}), \quad (5)$$

where F_{te} is the textual encoder with parameter ξ_{te}, t_m ($m \in [1, 11] \cap \mathbb{Z}$) denotes the m-th fine-grained text prompt, and $T_m^{'}$ represents the corresponding text representation. As we can see, the fine-grained text prompts describe the subjects and objects involved in a traffic event in a video frame, as well as the event type, which helps to focus on the visual regions in the driving scene where the traffic event occurred.

Therefore, we further propose to leverage the similarity of the fine-grained text representation with each deep feature of the video clip to focus on the most relevant visual context of the text prompt. Note that in the driving scenario, we do not have direct access to realistic text prompt that match the driving video. To solve this problem, we leverage the similarity between the visual representation F_n and fine-grained text representations to weight the text representations, and obtain the soft text representation as follows:

$$A_n^m = \frac{F_n \cdot (T_m^{'})^\top}{\|F_n\| \cdot \|T_m\|},$$
$$T_{soft} = \sum_{m=1}^{11} A_n^m T_m^{'}, \quad (6)$$

where A_n^m is the cosine similarity between the n-th visual representation F_n and the m-th fine-grained text representation $T_m^{'} \in \mathbb{R}^{1 \times C}$. After obtaining the soft text representation $T_{soft} \in \mathbb{R}^{1 \times C}$, similar to Section III-C1, we can further focus on the most semantically relevant visual context of the text description based on the cross-attention (CA) on the soft text representation T_{soft} and deep features P, which is denoted as:

$$LFR_n = softmax\left(\frac{Q\left(T_{soft}\right) \cdot K^\top\left(P\right)}{c}\right) \cdot V\left(P\right). \quad (7)$$

$LFR_n \in \mathbb{R}^{1 \times C}$ represents the linguistically focused representation of the n-th video clip, which focuses on the visual context that is most relevant to the soft text representation T_{soft}. Moreover, Fig. 3(c) shows that our LFS can indeed adaptively concentrate on road participants potentially linked to anomalies. This capability is crucial for identifying local anomalies in driving scenarios arising from non-ego vehicles (case 2 in Fig. 3).

Finally, we enhance the visual representation F_n of driving videos by fusing it with visually and linguistically focused representations. Formally, it can be expressed as:

$$F_n^{'} = F_{fusion}([VFR_n, LFR_n, F_n]; \xi_f), \quad (8)$$

where F_{fusion} is the fusion layer composed of multi-layer perceptrons with parameter ξ_f. $F_n^{'}$ is an enhanced visual representation that not only adaptively focuses on the visual contexts of interest but also more comprehensively characterizes the driving video clip in the spatio-temporal domain. Moreover, such representations facilitate the alignment of visual representations with general text prompts, thus improving the detection of traffic anomalies.

D. Contrastive Learning Strategy and Inference Process

In this section, we introduce the contrastive learning strategy of the proposed TTHF framework for cross-modal learning and present how to perform traffic anomaly detection.

Suppose that, there are N video clips in the batch, we denote:

$$F = \{F_1, F_2, ..., F_N\},$$
$$F^{'} = \{F_1^{'}, F_2^{'}, ..., F_N^{'}\}, \quad (9)$$

TABLE I
SUMMARY OF WELL-DESIGNED TEXT PROMPTS.

General Text Prompt	a_1:	"A traffic anomaly occurred in the scene."
	a_2:	"The traffic in this scenario is normal."
Fine-grained Text Prompt	b_1:	"The {ego, non-ego} vehicle collision with another {vechicle, pedestrian, obstacle}."
	b_2:	"The {ego, non-ego} vehicle out-of-control and leaving the roadway."
	b_3:	"The {ego, non-ego} vehicle has an unknown accident."
	b_4:	"The vehicle is running normally on the road."

TABLE II
TRAFFIC ANOMALY CATEGORY IN THE DoTA DATASET.

Label	Anomaly Category
ST	Collision with another vehicle that starts, stops, or is stationary
AH	Collision with another vehicle moving ahead or waiting
LA	Collision with another vehicle moving laterally in the same direction
OC	Collision with another oncoming vehicle
TC	Collision with another vehicle that turns into or crosses a road
VP	Collision between vehicle and pedestrian
VO	Collision with an obstacle in the roadway
OO	Out-of-control and leaving the roadway to the left or right
UK	Unknown

where F is the visual representation of N video clips and F' represents the enhanced visual representation. For text prompts, we denote:

$$T = \{T_1, T_2, ..., T_N\},$$
$$T' = \{T'_1, T'_2, ..., T'_N\}, \quad (10)$$

where T means the matched general text representation of N video clips and T' is the matched fine-grained text representation. Note that T_n and T'_n denote the high-dimensional representations of one of the D predefined text prompts. In our case, $D = 2$ for general text prompts and $D = 11$ for fine-grained text prompts. To better understand abstract concepts of traffic anomalies, we first perform contrastive learning to align visual representations F with fine-grained text representations T'. Formally, the objective loss along the visual axis can be expressed as:

$$\mathcal{L}_{vf} = \frac{1}{N} \sum_{i=1}^{N} -\log \frac{\exp(F_i \cdot (T'_i)^\top / \tau)}{\sum_{j \in [1,D]} \exp(F_i \cdot (T'_j)^\top / \tau)}. \quad (11)$$

For the j-th trained text representation T_j, it may actually match more than one visual representation. Symmetrically, we can calculate the loss along the text axis by:

$$\mathcal{L}_{tf} = \frac{1}{D} \sum_{j=1}^{D} -\log \frac{\sum_{k \in [1,N], T'_k = T'_j} \exp\left(T'_k \cdot F_k^\top / \tau\right)}{\sum_{i \in [1,N]} \exp\left(T'_j \cdot F_i^\top / \tau\right)}, \quad (12)$$

where τ is a learned temperature parameter [13]. Similarly, we further apply contrastive learning to align the enhanced visual representations with the general text representations. The calculations along the visual and textual axis are as follows:

$$\mathcal{L}_{vg} = \frac{1}{N} \sum_{i=1}^{N} -\log \frac{\exp(F'_i \cdot (T_i)^\top / \tau)}{\sum_{j \in [1,D]} \exp(F'_i \cdot (T_j)^\top / \tau)},$$
$$\mathcal{L}_{tg} = \frac{1}{D} \sum_{j=1}^{D} -\log \frac{\sum_{k \in [1,N], T_k = T_j} \exp\left(T_k \cdot (F'_k)^\top / \tau\right)}{\sum_{i \in [1,N]} \exp\left(T_j \cdot (F'_i)^\top / \tau\right)}. \quad (13)$$

The overall loss then becomes:

$$\mathcal{L} = ((\mathcal{L}_{vf} + \mathcal{L}_{tf}) + (\mathcal{L}_{vg} + \mathcal{L}_{tg}))/2. \quad (14)$$

The inference procedure is similar to the training procedure. For the i-th testing driving video clip, our TTHF first extracts the visual representation F_i and the enhanced visual representation F'_i. For text prompts, the text encoder constructs 11

fine-grained text representations $T' = \{T'_1, T'_2, ..., T'_{11}\}$ and 2 general text representations $T = \{T_1, T_2\}$. We then compute the cosine similarity between F_i and T' and between F'_i and T, respectively. Finally, we calculate the anomaly score for the i-th driving video clip as:

$$Score_i = 1 - (S_f^{11} + S_g^2)/2, \quad (15)$$

where S_f^{11} represents the cosine similarity after softmax between F_i and T'_{11}, and S_g^2 denotes the cosine similarity after softmax between F'_i and T_2. By taking the complement of the average over the prompts corresponding to normal traffic at different levels, we can obtain the final anomaly score $Score_i$.

IV. EXPERIMENTS AND DISCUSSIONS

In this section, we evaluate the performance of our proposed method, which is performed on a platform with one NVIDIA 3090 GPU. All experiments were implemented using the PyTorch framework. Our source code and trained models will be publicly available upon acceptance.

A. Implementation Details

In the experiments, we resize the driving video frames to 224×224 and take every two consecutive frames as the input video clip. Except where noted otherwise, in all experimental settings, we adopt ResNet-50 [52] for the visual and high-frequency encoders and Text Transformer [53] for the textual encoder. All of them are initialized with the parameters of CLIP's pre-trained model. Note that during the training phase, we freeze the pre-trained parameters of the visual encoder to prevent the model from overfitting to a specific dataset (*e.g.*, DoTA) while enhancing the generalization of the visual representation. Besides, we optimize loss functions using the Adam algorithm with batch size 128, learning rate 5e-6, weight decay 1e-4, and train the framework for 10 epochs. During inference, we evaluate the traffic anomaly score by taking the complement of the similarity score of normal traffic prompts on both fine-grained and general text prompts.

B. Dataset and Metrics

1) Dataset: For the sake of fairness, we evaluate our method on two challenging datasets, namely, DoTA [9] and DADA-2000 [23], following prior works [8], [9], [12]. DoTA is the first traffic anomaly video dataset that provides detailed

TABLE III
THE AUC ↑ (%) OF DIFFERENT APPROACHES ON THE DoTA DATASET.

Methods	Input	Paradigm	AUC (%)
ConvAE [10]	Gray	Single-Stage	64.3
ConvAE [10]	Flow	Two-Stage	66.3
ConvLSTMAE [11]	Gray	Single-Stage	53.8
ConvLSTMAE [11]	Flow	Two-Stage	62.5
AnoPred [6]	RGB	Single-Stage	67.5
AnoPred [6]	Mask RGB	Two-Stage	64.8
FOL-STD [32]	Box	Two-Stage	66.7
FOL-STD [32]	Box + Flow	Two-Stage	69.1
FOL-STD [32]	Box + Flow + Ego	Two-Stage	69.7
FOL-Ensemble [9]	RGB + Box + Flow + Ego	Two-Stage	73.0
STFE [8]	RGB + Box	Two-Stage	<u>79.3</u>
TTHF-Base	RGB	Single-Stage	75.8
TTHF	RGB	Single-Stage	**84.7**

spatio-temporal annotations of anomalous objects for traffic anomaly detection in driving scenarios. The dataset contains 4677 dashcam video clips with a resolution of 1280×720 pixels, captured under various weather and lighting conditions. Each video is annotated with the start and end time of the anomaly and assigned to one of nine categories, which we summarize in Table II. The DADA-2000 dataset consists of 2000 dashcam videos with a resolution of 1584×660 pixels, each annotated with driver attention and one of 54 anomaly categories. In our experiments, we use the standard train-test split as used in [9], [23] and other previous works.

2) Metrics: Following prior works [8], [9], [54], we use Area under ROC curve (AUC) metric to evaluate the performance of different TAD approaches. The AUC metric is calculated by computing the area under a standard frame-level receiver operating characteristic (ROC) curve, which plots the true positive rate (TPR) against the false positive rate (FPR). The larger AUC prefers better performance.

C. Competitors

To verify the superiority of the proposed framework, we compare with the following state-of-the-art TAD approaches: ConvAE [10], ConvLSTMAE [11], AnoPred [6], FOL-STD [32], FOL-Ensemble [9], DMMNet [55], SSC-TAD [12] and STFE [8]. Among them, the ConvAE [10] and ConvLSTMAE [11] methods contain two variants. The variant utilizing the grayscale image as input belongs to the single-stage paradigm, while the variant using optical flow as input belongs to the two-stage paradigm. The AnoPred method [6] also contains two variants. The variant employing the full video frame as input falls within the single-stage paradigm, whereas the variant utilizing pixels of foreground objects belongs to the two-stage paradigm. Besides, the DMMNet method [55] follows the single-stage paradigm, while the methods FOL-STD [32], FOL-Ensemble [9], SSC-TAD [12], and STFE [8] fall under the two-stage paradigm. Note that the experimental results for all these methods and their variants are obtained from the published papers [8], [9], [12]. In addition, we consider a CLIP-like TAD framework, denoted as TTHF-Base, as our baseline approach. This baseline lacks temporal High-Frequency

Modeling and the attention anomaly focusing mechanism and utilizes only general text prompts for alignment.

D. Quantitative Results

1) Overall results: We conduct a comparative analysis of TTHF with a wide range of competitors and their variants in terms of AUC metric. Table III presents the AUC performance of various competitors, along with labels indicating their respective variants (*i.e.*, different inputs) and paradigms employed. Overall, our framework demonstrates the superior performance on the DoTA dataset in terms of AUC. Specifically, our method outperforms the previously two-stage paradigm-based leading TAD method, STFE [8], by +5.4% AUC. Although in previous methods, the two-stage paradigm method employs a perception algorithm in the first stage to mitigate the impact of dynamic background resulting from the ego-vehicle movement, and generally outperforms single-stage TAD methods [6], [10], [11], such approaches are susceptible to the performance of the perception algorithm in the first stage, potentially leading to error propagation. In contrast, our proposed single-stage TAD method explicitly characterizes dynamic changes by modeling high frequency in the temporal domain, achieving a significant performance improvement over all previous methods and establishing a new state-of-the-art in traffic anomaly detection. Note that our baseline method outperforms all previous single-stage paradigm-based methods by at least +8.3% AUC. This is mainly attributed to our introduction of text prompts and the alignment of driving videos with text representations in a high-dimensional space, which facilitates the detection of traffic anomalies.

2) Per-class results: To investigate the ability of our proposed method to detect traffic anomalies in different categories, we compared the detection performance of different methods for ego-involved and non-ego traffic anomalies. Based on the nine traffic anomalies divided by the DoTA dataset, detailed in Table II, we summarize the AUC performance of the different methods as well as the average AUC in Table IV. Our method achieves significant improvements in all categories of traffic anomalies except ST*, and in particular, achieves an average AUC of at least +9.9% on egos involving traffic anomalies. This further validates our idea that characterizing dynamic changes in driving scenarios is important for traffic anomaly detection. Simultaneously, it also demonstrates the effectiveness of our proposed approach to model the temporal high frequency of driving videos to characterize the dynamic changes of driving scenes.

3) Generalization performance: To explore the generalization performance of our method for unseen types of traffic anomalies, we perform a generalization experiment on the DADA-2000 dataset. Specifically, we compare the AUC performance of our TTHF and TTHF-Base without any fine tuning on the DADA-2000 dataset with previous trained models, summarized in Table V. As we can see, our proposed TTHF-base and TTHF methods outperform previously trained TAD methods, bringing at least +0.8% and +4.2% improvement in AUC respectively, indicating the strong generalization performance of the proposed approach. This is mainly attributed to

TABLE IV

THE AUC ↑ (%) OF DIFFERENT METHODS FOR EACH INDIVIDUAL ANOMALY CLASS ON THE DoTA DATASET IS PRESENTED. THE ∗ INDICATES NON-EGO ANOMALIES, WHILE EGO-INVOLVED ANOMALIES ARE SHOWN WITHOUT ∗. N/A INDICATES THAT THE AUC PERFORMANCE FOR THE CORRESPONDING CATEGORY IS NOT AVAILABLE. WE **BOLD** THE BEST PERFORMANCE.

Methods	ST	AH	LA	OC	TC	VP	VO	OO	UK	AVG
AnoPred [6]	69.9	73.6	75.2	69.7	73.5	66.3	N/A	N/A	N/A	71.4
AnoPred [6]+Mask	66.3	72.2	64.2	65.4	65.6	66.6	N/A	N/A	N/A	66.7
FOL-STD [32]	67.3	77.4	71.1	68.6	69.2	65.1	N/A	N/A	N/A	69.7
FOL-Ensemble [9]	73.3	81.2	74.0	73.4	75.1	70.1	N/A	N/A	N/A	74.5
STFE [8]	<u>75.2</u>	<u>84.5</u>	72.1	<u>77.3</u>	72.8	71.9	N/A	N/A	N/A	75.6
TTHF-Base	72.8	79.6	<u>83.7</u>	76.4	<u>82.6</u>	<u>72.3</u>	<u>81.8</u>	<u>80.4</u>	**72.7**	<u>78.0</u>
TTHF	**86.7**	**90.5**	**89.7**	**87.0**	**89.5**	**77.1**	**87.6**	**90.1**	<u>70.9</u>	**85.5**
Methods	ST*	AH*	LA*	OC*	TC*	VP*	VO*	OO*	UK*	AVG
AnoPred [6]	70.9	62.6	60.1	65.6	65.4	64.9	64.2	57.8	N/A	63.9
AnoPred [6]+Mask	72.9	63.7	60.6	66.9	65.7	64.0	58.8	59.9	N/A	64.1
FOL-STD [32]	75.1	66.2	66.8	74.1	72.0	69.7	63.8	69.2	N/A	69.6
FOL-Ensemble [9]	<u>77.5</u>	69.8	68.1	<u>76.7</u>	73.9	<u>71.2</u>	65.2	69.6	N/A	71.5
STFE [8]	**80.6**	65.6	<u>69.9</u>	76.5	<u>74.2</u>	N/A	<u>75.6</u>	<u>70.5</u>	N/A	<u>73.2</u>
TTHF-Base	75.0	<u>71.5</u>	67.2	72.5	70.6	64.3	69.9	68.3	<u>68.1</u>	69.7
TTHF	74.9	**76.0**	**76.4**	**79.8**	**81.5**	**79.2**	**79.0**	**77.5**	**68.9**	**77.0**

TABLE V

THE AUC ↑ (%) OF DIFFERENT METHODS ON THE DADA-2000 DATASET.

Methods	Trained	Ego-Involved	Non-Ego	Both
AnoPred [6]	✓	55.7	56.9	56.1
FOL-STD [32]	✓	71.3	57.1	66.6
DMMNet [55]	✓	73.0	56.3	67.5
SSC-TAD [12]	✓	67.6	58.7	66.5
TTHF-Base	✗	<u>78.7</u>	<u>59.4</u>	68.3
TTHF	✗	**80.9**	**64.0**	**71.7**

TABLE VI

ABLATION RESULTS OF DIFFERENT COMPONENTS ON DoTA DATASET. NOTE THAT FOR FAIR COMPARISON, IN THE EXPERIMENTS WITHOUT THFM, WE FINE-TUNE THE PARAMETERS OF THE VISUAL ENCODER. LARGER AUC PREFERS BETTER PERFORMANCE.

Arch.	Visual	Textual	AAFM	THFM	AUC (%)
TTHF	✓	✗	✗	✗	61.0
TTHF	✓	✓	✗	✗	75.8
TTHF	✓	✓	✓	✗	76.8
TTHF	✓	✓	✓	✓	**84.7**

TABLE VII

ABLATION RESULTS ON HOW AAFM CONTRIBUTES TO TRAFFIC ANOMALY DETECTION ON THE DoTA DATASET. LARGER AUC PREFERS BETTER PERFORMANCE.

Arch.	VFS	LFS	AUC (%)
TTHF	–	–	75.8
TTHF	✓	✗	76.3
TTHF	✗	✓	76.5
TTHF	✓	✓	**76.8**

TABLE VIII

ABLATION RESULTS OF DIFFERENT BACKBONES ON DoTA DATASET. LARGER AUC PREFERS BETTER PERFORMANCE.

Arch.	Visual	Textual	AUC (%)
TTHF	RN-50	Text-Transformer	84.7
TTHF	RN-50x64	Text-Transformer	84.8
TTHF	ViT-B-32	Text-Transformer	84.0
TTHF	ViT-L-14	Text-Transformer	**85.0**

types of traffic anomalies and the feasibility of soft text representation in our framework.

1) Visualization of various types of traffic anomalies: As presented in Fig. 5, we show five representative traffic anomalies from top to bottom as examples: a) The other vehicle collides with another vehicle that turns into or crosses a road. b) The ego-vehicle collides with another oncoming vehicle. c) The ego-vehicle collides with another vehicle moving laterally in the same direction. d) The ego-vehicle collides with another vehicle waiting. e) The ego-vehicle is out-of-control and leaving the roadway to the left. From the above visualization results of different types of traffic anomalies, we can summarize as follows. Overall, our TTHF exhibits superior detection performance on various types of traffic anomalies. Secondly, while the most intuitive classify-based approach (It has the same network architecture as the visual encoder of TTHF, but directly classifies the visual representation, denoted as Classifier in Fig. 5) also follows a single-stage paradigm, our proposed text-driven TAD approach offers a more comprehensive representation in high-dimensional space than orthogonal one-hot vectors. Consequently, both our proposed TTHF and its variants outperform the Classifier. Third, incorporating AAFM allows our method to better perceive different types of traffic anomalies, as evident in Fig. 5 when comparing the Base and AAFM variants across various traffic anomalies. Finally, capturing dynamic changes in driving scenarios significantly

our introduction of a text-driven video-text alignment strategy for traffic anomaly detection from a new perspective, as well as the proposed attentive anomaly focusing mechanism and temporal high-frequency modeling for traffic anomaly detection.

E. Qualitative Results

In this subsection, we visualize some examples to further illustrate the detection capability of our TTHF across various
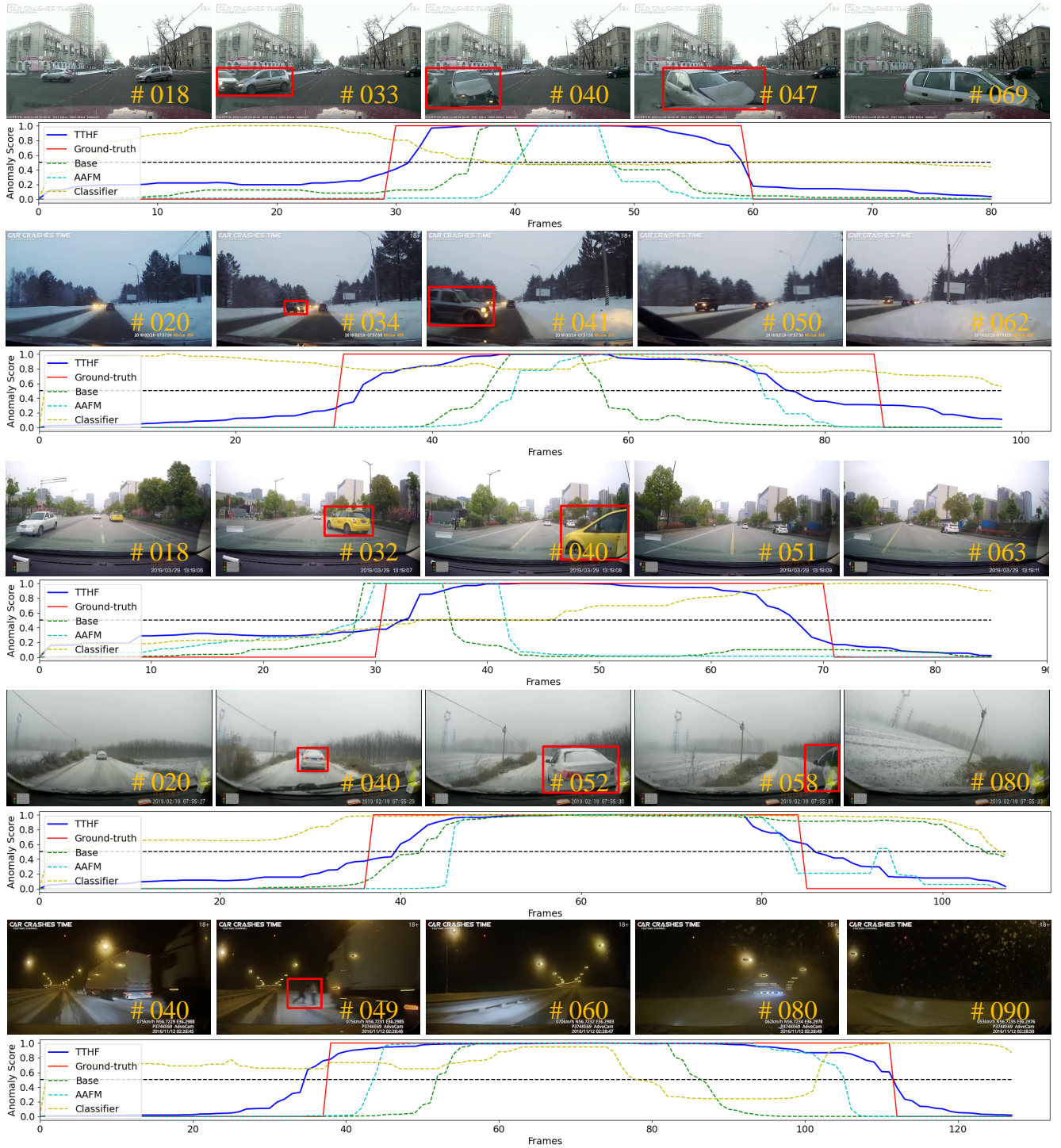
Fig. 5. The visualization of anomaly score curves for traffic anomaly detection of different variants on the DoTA dataset. The first row of each case shows the extracted video frames of the driving video, where the red boxes mark the object involved in or causing the anomaly. The second rows show the anomaly score curves of different methods on the corresponding whole videos. For brevity, we label the TTHF-Base variant as Base and TTHF-Base with AAFM as AAFM, while Classifier denotes the classify-based TAD method. Better viewed in color.

enhances traffic anomaly detection. This highlights the effectiveness of our approach in characterizing dynamic changes in driving scenarios by modeling high frequency in the temporal domain.

2) Visualization of the weights used for soft text representation: We further investigate the feasibility of soft text representations. Specifically, as shown in Fig. 6, we use

three cases from the test set as examples. For video frames captured at different moments in driving videos, we visualize the weights employed to compute the soft text representation and compare it with the real fine-grained text representation. From the visualization results, we observe that the text representation associated with the maximum weight (indicated by

Fig. 6. Visualization of the weights used for computing soft text representations. We present three illustrative cases, each involving video frames captured at different times. These frames are accompanied by the corresponding weight values used in the computation of soft text representations. Notably, we employ a blue-to-red color scale, where increasing redness signifies higher weights. Additionally, we label the ground-truth fine-grained text representations (denoted as T_i) associated with specific frames. Among them, T_1 corresponds to the text *"The ego vehicle collision with another vehicle"* (as described in Table I), T_4 corresponds to the text *"The non-ego vehicle collision with another vehicle"*, T_7 corresponds to the text *"The ego vehicle out-of-control and leaving the roadway"*, and T_11 corresponds to the text *"The vehicle is running normally on the road"*.

the darkest red) consistently aligns with the real fine-grained text representation. The above results indicate that the way we calculate the soft text representation is effective and can well reflect the real anomaly category.

F. Ablation Investigation

In this subsection, we conduct ablation studies by analyzing how different components of TTHF contribute to traffic anomaly detection on DoTA dataset.

1) Variants of our architecture: We first evaluate the effectiveness of different components in our TTHF framework including the visual encoder, the textual encoder, the attentive anomaly focusing mechanism (AAFM), and the temporal high-frequency modeling (THFM). The ablation results are summarized in Table VI. Note that when only the visual encoder is applied, we add a linear classification head after the visual representation. This adaptation formulates the traffic anomaly detection task as a straightforward binary classification task. The results presented in Table VI demonstrate that introducing linguistic modalities and aligning visual-text in high-dimensional space greatly facilitates anomaly detection in driving videos compared to the classifier, achieving an AUC improvement of +14.8%. Based on this, the designed AAFM helps guide the model to adaptively focus on the visual context of interest and thus enhance the perception ability of various types of traffic anomalies. Lastly, the incorporation of the modeling of temporal high frequency to capture dynamic background during driving significantly improves traffic anomaly detection, resulting in an AUC improvement of +7.9%.

2) Analysis of the AAFM: To investigate how the proposed attentive anomaly focusing mechanism (AAFM) contributes to traffic anomaly detection, we perform ablation on each component in the AAFM. The ablation results are presented in Table VII. We can conclude that both the Visually Focused Strategy (VFS) and the Linguistically Focused Strategy (LFS) explicitly guide the model to pay attention to the visual context most relevant to the representations of visual and linguistic modalities, respectively. This enhances the ability to perceive traffic anomalies with different characteristics, thereby improving traffic anomaly detection in driving videos. Our AAFM achieves the best detection performance when both VFS and LFS are applied.

3) Network Architecture: Different network architectures of visual encoder may exhibit different representation capabilities. We now evaluate the performance of traffic anomaly detection when ResNet50 [52], ResNet50x64 [13], ViT-B-32 [56] and ViT-L-14 [56] are used. Specifically, the results of these visual encoders can be found in Table VIII, respectively. As can be noticed, for the task of traffic anomaly detection in driving videos, we observe that the ResNet-based network achieves comparable performance to the Transformer-based network. The larger model sizes perform slightly better, with ViT-L-14 achieving an AUC performance of 85.0%. Therefore, considering both computing resources and performance gains, we ultimately chose ResNet50 as an example as our visual encoder in all other experiments.
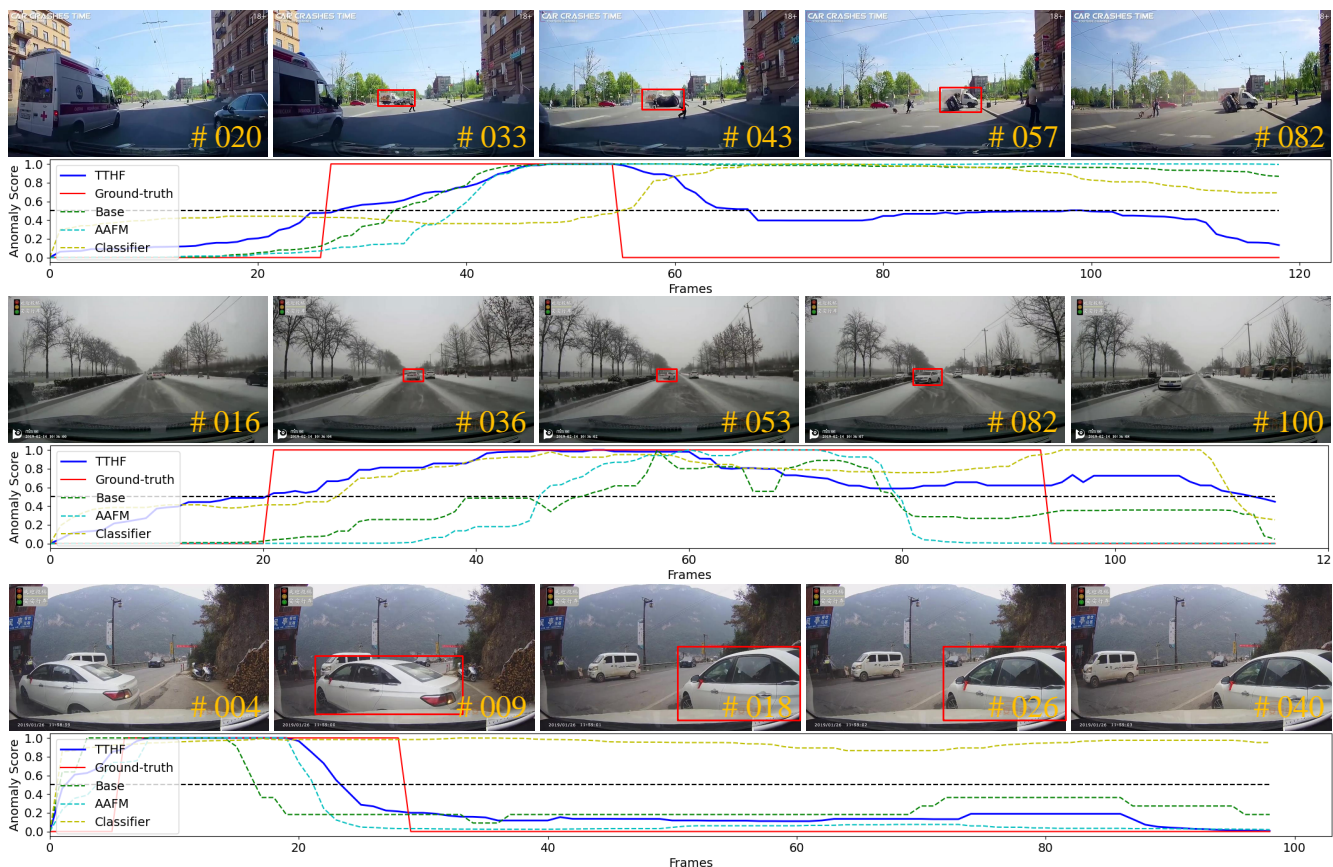
Fig. 7. Visualization of some bad cases of the proposed TTHF. The first row of each case shows the extracted video frames of the driving video, where the red boxes mark the objects involved in the anomaly. The second rows show the anomaly score curves of different methods on the corresponding whole videos. Better viewed in color.

G. Disscusion

In this subsection, we discuss the limitations of the proposed framework. We experimentally found that the detection accuracy of our proposed method needs improvement for two specific cases: 1) long-distance observation of traffic anomalies; and 2) subtle traffic anomalies involving other vehicles when the ego-vehicle is stationary. Fig. 7 shows several cases where the accuracy of our method needs to be further improved. In the first scenario, the other vehicle at a distance collide with a turning or crossing vehicle. The second scenario depicts a distant vehicle losing control and veering to the left side of the road. The third scenario involves a slowly retreating vehicle experiencing friction with other stationary vehicles. By analyzing the anomaly score curve in Fig. 7, we can conclude that our method faces challenges primarily due to the traffic anomalies occurring in these scenarios involve non-ego vehicles and cause minor anomaly areas. These anomalies include small local anomalies that are caused when non-ego vehicles are abnormal at a distance, and slow and slight traffic anomalies that are observed for other vehicles when the ego-vehicle is at rest. These slight traffic anomalies may not be well focused on the corresponding abnormal regions by modeling the dynamic changes of the driving scene as well as using text guidance. This also explains that the ability of our method in detecting non-ego involved traffic anomalies is not as

good as in detecting ego-involved traffic anomalies, especially ST* in Table IV. Despite the significant improvement of our approach over previous TAD methods, addressing these more challenging traffic anomalies undoubtedly requires a greater effort from the community.

V. CONCLUSION

This paper have proposed an accurate single-stage TAD framework. For the first time, this framework introduces visual-text alignment to address the traffic anomaly detection task for driving videos. Notably, we verified that modeling the high frequency of driving videos in the temporal domain helps to characterize the dynamic changes of the driving scene and enhance the visual representation, thereby greatly facilitating the detection of traffic anomalies. In addition, the experimental results demonstrated that the proposed attentive anomaly focusing mechanism is indeed effective in guiding the model to adaptively focus on the visual content of interest, thereby enhancing the ability to perceive different types of traffic anomalies. Although extensive experiments have demonstrated that the proposed TTHF substantially outperforms state-of-the-art competitors, more effort is required to accurately detect the more challenging slight traffic anomalies.

REFERENCES

[1] Z. Yuan, X. Song, L. Bai, Z. Wang, and W. Ouyang, "Temporal-channel transformer for 3d lidar-based video object detection for autonomous driving," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2068–2078, 2022.

[2] L. Claussmann, M. Revilloud, D. Gruyer, and S. Glaser, "A review of motion planning for highway autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 1826–1848, 2020.

[3] M. Jeong, B. C. Ko, and J.-Y. Nam, "Early detection of sudden pedestrian crossing for safe driving during summer nights," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 6, pp. 1368–1380, 2017.

[4] L. Yue, M. A. Abdel-Aty, Y. Wu, and A. Farid, "The practical effectiveness of advanced driver assistance systems at different roadway facilities: System limitation, adoption, and usage," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3859–3870, 2020.

[5] Y. Yuan, D. Wang, and Q. Wang, "Anomaly detection in traffic scenes via spatial-aware motion reconstruction," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1198–1209, 2017.

[6] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection – a new baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2018, pp. 6536–6545.

[7] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, "A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 13 588–13 597.

[8] Z. Zhou, X. Dong, Z. Li, K. Yu, C. Ding, and Y. Yang, "Spatio-temporal feature encoding for traffic accident detection in vanet environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19 772–19 781, 2022.

[9] Y. Yao, X. Wang, M. Xu, Z. Pu, Y. Wang, E. Atkins, and D. J. Crandall, "Dota: Unsupervised detection of traffic anomaly in driving videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 444–459, 2023.

[10] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2016, pp. 733–742.

[11] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Proc. Adv. Neural Networks*, 2017, pp. 189–196.

[12] J. Fang, J. Qiao, J. Bai, H. Yu, and J. Xue, "Traffic accident detection via self-supervised consistency learning in driving scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9601–9614, 2022.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. conf. mach. learn.*, vol. 139, 2021, pp. 8748–8763.

[14] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. conf. mach. learn.*, vol. 139, 2021, pp. 4904–4916.

[15] Y. Yang, W. Huang, Y. Wei, H. Peng, X. Jiang, H. Jiang, F. Wei, Y. Wang, H. Hu, L. Qiu, and Y. Yang, "Attentive mask clip," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 2771–2781.

[16] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *Proc. Int. Conf. Learn. Represent.*, 2022.

[17] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, "Groupvit: Semantic segmentation emerges from text supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2022, pp. 18 134–18 144.

[18] S. Chen, Q. Xu, Y. Ma, Y. Qiao, and Y. Wang, "Attentive snippet prompting for video retrieval," *IEEE Trans. Multimed.*, pp. 1–12, 2023.

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

[20] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, 2022.

[21] H. Rasheed, M. U. Khattak, M. Maaz, S. Khan, and F. S. Khan, "Fine-tuned clip models are efficient video learners," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2023, pp. 6545–6554.

[22] Y. Li, J. Ye, L. Zeng, R. Liang, X. Zheng, W. Sun, and N. Wang, "Learning hierarchical fingerprints via multi-level fusion for video integrity and source analysis," *IEEE Trans. Consum. Electron.*, pp. 1–11, 2024.

[23] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "Dada: Driver attention prediction in driving accident scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4959–4971, 2022.

[24] Y. Zhong, X. Chen, Y. Hu, P. Tang, and F. Ren, "Bidirectional spatio-temporal feature learning with multiscale evaluation for video anomaly detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8285–8296, 2022.

[25] M. I. Georgescu, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, "A background-agnostic framework with adversarial training for abnormal event detection in video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4505–4523, 2022.

[26] S. Zhang, M. Gong, Y. Xie, A. K. Qin, H. Li, Y. Gao, and Y.-S. Ong, "Influence-aware attention networks for anomaly detection in surveillance videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5427–5437, 2022.

[27] X. Zeng, Y. Jiang, W. Ding, H. Li, Y. Hao, and Z. Qiu, "A hierarchical spatio-temporal graph convolutional neural network for anomaly detection in videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 200–212, 2023.

[28] C. Huang, J. Wen, Y. Xu, Q. Jiang, J. Yang, Y. Wang, and D. Zhang, "Self-supervised attentive generative adversarial networks for video anomaly detection," *IEEE Trans. Neural Netw. Learn. Systems*, vol. 34, no. 11, pp. 9389–9403, 2023.

[29] Y. Gong, C. Wang, X. Dai, S. Yu, L. Xiang, and J. Wu, "Multi-scale continuity-aware refinement network for weakly supervised video anomaly detection," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2022, pp. 1–6.

[30] Y. Wang, X. Luo, and Z. Zhou, "Contrasting estimation of pattern prototypes for anomaly detection in urban crowd flow," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–15, 2024.

[31] Y. Yuan, J. Fang, and Q. Wang, "Incrementally perceiving hazards in driving," *Neurocomputing*, vol. 282, pp. 202–217, 2018.

[32] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised traffic accident detection in first-person videos," in *Proc. IEEE Int. Conf. Intell. Rob. Syst.*, 2019, pp. 273–280.

[33] G. Sun, Z. Liu, L. Wen, J. Shi, and C. Xu, "Anomaly crossing: New horizons for video anomaly detection as cross-domain few-shot learning," *arXiv preprint arXiv:2112.06320*, 2022.

[34] R. Liang, Y. Li, Y. Yi, J. Zhou, and X. Li, "A memory-augmented multi-task collaborative framework for unsupervised traffic accident detection in driving videos," *arXiv preprint arXiv:2307.14575*, 2023.

[35] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017.

[36] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2017.

[37] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Processing*, 2017, pp. 3645–3649.

[38] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Trans. Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[39] R. Liang, Y. Li, J. Zhou, and X. Li, "Stglow: A flow-based generative framework with dual-graphormer for pedestrian trajectory prediction," *IEEE Trans. Neural Netw. Learn. Systems*, pp. 1–14, 2023.

[40] L. Yao, J. Han, X. Liang, D. Xu, W. Zhang, Z. Li, and H. Xu, "Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2023, pp. 23 497–23 506.

[41] Z. Zhou, Y. Lei, B. Zhang, L. Liu, and Y. Liu, "Zegclip: Towards adapting clip for zero-shot semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2023, pp. 11 175–11 185.

[42] A. Baldrati, L. Agnolucci, M. Bertini, and A. Del Bimbo, "Zero-shot composed image retrieval with textual inversion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 15 338–15 347.

[43] M. Tschannen, B. Mustafa, and N. Houlsby, "Clippo: Image-and-language understanding from pixels only," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2023, pp. 11 006–11 017.

[44] S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang, "Zero-shot temporal action detection via vision-language prompting," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 681–697.

[45] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, and R. Ji, "X-clip: End-to-end multi-grained contrastive learning for video-text retrieval," in *Proc. ACM Int. Conf. Multi.*, 2022, pp. 638–647.

[46] W. Wu, Z. Sun, and W. Ouyang, "Revisiting classifier: Transferring vision-language models for video recognition," in *Proc. AAAI Conf. Art. Intel.*, vol. 37, 2023, pp. 2847–2855.

[47] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, and H. Ling, "Expanding language-image pretrained models for general video recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 1–18.

[48] P. Wu, X. Zhou, G. Pang, L. Zhou, Q. Yan, P. Wang, and Y. Zhang, "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection," *arXiv preprint arXiv:2308.11681*, 2023.

[49] R. Zhang, Z. Zeng, Z. Guo, and Y. Li, "Can language understand depth?" in *Proc. ACM Int. Conf. Multi.*, 2022, p. 6868–6874.

[50] Z. Liang, C. Li, S. Zhou, R. Feng, and C. C. Loy, "Iterative prompt learning for unsupervised backlit image enhancement," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 8094–8103.

[51] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2022, pp. 16 816–16 825.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2016.

[53] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 8, no. 1, pp. 1–9, 2019.

[54] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1705–1714.

[55] S. Li, J. Fang, H. Xu, and J. Xue, "Video frame prediction by deep multi-branch mask network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1283–1295, 2021.

[56] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *In Proc. Int. Conf. Learn. Representat.*, 2021, pp. 1–22.
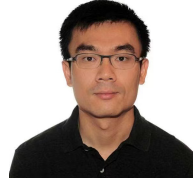
Jiantao Zhou (Senior Member, IEEE) received the B.Eng. degree from the Department of Electronic Engineering, Dalian University of Technology, in 2002, the M.Phil. degree from the Department of Radio Engineering, Southeast University, in 2005, and the Ph.D. degree from the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, in 2009. He held various research positions with University of Illinois at Urbana-Champaign, Hong Kong University of Science and Technology, and McMaster University. He is an Associate Professor with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, and also the Interim Head of the newly established Centre for Artificial Intelligence and Robotics. His research interests include multimedia security and forensics, multimedia signal processing, artificial intelligence and big data. He holds four granted U.S. patents and two granted Chinese patents. He has co-authored two papers that received the Best Paper Award at the IEEE Pacific-Rim Conference on Multimedia in 2007 and the Best Student Paper Award at the IEEE International Conference on Multimedia and Expo in 2016. He is serving as the Associate Editors of the IEEE TRANSACTIONS on IMAGE PROCESSING and the IEEE TRANSACTIONS on MULTIMEDIA.

Rongqin Liang (Student Member, IEEE) received the B.Eng. degree in communication engineering from Wuyi University, Guangdong, China, in 2018 and M.S. degree in Information and Communication Engineering from Shenzhen University, Shenzhen, China, in 2021. He is currently a Ph.D. candidate at the College of Electronics and Information Engineering from Shenzhen University. His current research interests include trajectory prediction, anomaly detection, computer vision and deep learning.

Xia Li (Member, IEEE) received her B.S. and M.S. in electronic engineering and SIP (signal and information processing) from Xidian University in 1989 and 1992 respectively. She was later conferred a Ph.D. in Department of information engineering by the Chinese University of Hong Kong in 1997. Currently, she is a member of the Guangdong Key Laboratory of Intelligent Information Processing. Her research interests include intelligent computing and its applications, image processing and pattern recognition.

Yuanman Li (Senior Member, IEEE) received the B.Eng. degree in software engineering from Chongqing University, Chongqing, China, in 2012, and the Ph.D. degree in computer science from University of Macau, Macau, 2018. From 2018 to 2019, he was a Post-doctoral Fellow with the State Key Laboratory of Internet of Things for Smart City, University of Macau. He is currently an Assistant Professor with the College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China. His current research interests include multimedia security and forensics, data representation, computer vision and machine learning.