# Surveillance-based Spatial Temporal Traffic Accident Detection: A novel Dataset and tailored Algorithm*

Sixuan Xu[1], *Student Member, IEEE*, Tianran Zhang[2], Lei Zhao[3], Wei Zhou[4],*Graduate Student Member, IEEE*, Stefan Nastic[5] and Chen Wang[6*], *Member, IEEE*

***Abstract*— Traffic Accident Detection (TAD) in surveillance videos is a critical task in Intelligent Transportation Systems (ITS). However, current TAD does not analyze the fine-grained information of the specific accident, only identifies the existence or occurrence time of traffic accidents in a video. This study presents a novel Dataset named STTAD that covers fine-grained information such as multiple categories and their Spatial Temporal Occurrence Regions in surveillance videos. Moreover, a tailored deep learning algorithm named STFN is proposed for the implementation of Event-Level TAD. Experimental results demonstrate that STFN could effectively extract the video features and detect the Spatial Temporal Occurrence Regions of multiple accident categories, but further efforts are indeed needed in Event-Level TAD. The STTAD dataset and the tailored algorithm will be open-sourced for research use available through *https://github.com/ZTR02/STTAD.git*.**

## I. INTRODUCTION

With the increase of traffic accidents, more precise and reliable traffic accident detection (TAD) is required for Intelligent Transportation Systems (ITS) to enable timely emergency response and efficient incident management strategies, thereby minimizing secondary casualties, attenuating economic implications and alleviating traffic congestion.

Traffic accidents constitute extreme traffic anomalies characterized by events such as vehicular collisions or rollovers, typically resulting in property damage, injuries, or fatalities [1]. The proliferation of monitoring equipment and the growth of accident-related video datasets have enabled the implementation of automated TAD utilizing advanced Computer Vision methodologies [2]. Existing accident-related video datasets predominantly focus on identifying the moment of accident occurrence, documented via surveillance cameras or vehicle-mounted cameras. Given the comprehensive view and temporal coverage provided by surveillance cameras, this study focuses on surveillance cameras for TAD.

Traditional TAD focuses mainly on detecting the existence, conceptualizing accidents as a subset of anomaly detection, and employing binary (0/1) classification to denote the occurrence of accidents [3], [4]. Although this binary labeling offers simplicity, it provides insufficient granularity and omits critical information of different occurrence patterns for various traffic accidents. In addition, recent studies emphasize temporal analysis, which identify the initiation and termination frames of accidents [5], [6]. However, the occurrence region of traffic accidents (i.e., events), which refers to the sequence of frames where a traffic accident happens and the spatial region within each frame where it occurs, remains underdeveloped and is effectively limited to rudimentary object detection (e.g., pedestrian, bike) [7], [8], [9].

To enhance the precision and reliability of TAD, this study constructs a novel accident-related video datasets STTAD, which performs multiple types of accidents and precise spatial temporal occurrence region via bounding box annotations on each video frame from surveillance cameras. The collected videos are classified into 12 distinct categories, thereby expanding the conventional binary (0/1) classification paradigm to multiple classification tasks. STTAD includes 1,189 accident videos that comprise 64,554 frames in total. The annotation protocol incorporates sequential frame indexing to preserve temporal coherence, followed by precise accident region delineation through bounding box coordinates (centroid x-y coordinates, width, and height dimensions) for each video frame. This spatial demarcation facilitates model attention to salient accident features. Furthermore, we develop a tailored algorithm for event-level TAD, and conduct continuous spatial temporal analysis of accident occurrence regions. The STTAD dataset and corresponding algorithm will be publicly accessible for research purposes via *https://github.com/ZTR02/STTAD.git*. The principal contributions of this study are as follows.

- A novel dataset STTAD is proposed for fine-grained analysis in spatial temporal TAD.
- A tailored algorithm is proposed for the implementation of spatial temporal TAD.
- Various experiments are presented, verifying the effectiveness of proposed Algorithm and demonstrating the challenge of dataset STTAD.

*This study is supported by the National Key R&D Program of China [grant number 2023YFE0106800], Outstanding Youth Foundation of Jiangsu Province [grant number BK20231531], Frontier Technologies R&D Program of Jiangsu [grant number BF2024019], Postgraduate Research & Practice Innovation Program of Jiangsu Province [grant number SJCX24_0100, KYCX25_0510], Austrian Research Promotion Agency (FFG) under the project RapidREC [grant number 903884]

[1]Sixuan Xu is with Intelligent Transportation System Research Center, Southeast University, 2 Southeast University Road, Jiangsu District, Nanjing, 210096, P.R., China 230238937@seu.edu.cn

[2]Tianran Zhang is with Intelligent Transportation System Research Center, Southeast University, 2 Southeast University Road, Jiangsu District, Nanjing, 210096, P.R., China 220243471@seu.edu.cn

[3]Lei Zhao is with Intelligent Transportation System Research Center, Southeast University, 2 Southeast University Road, Jiangsu District, Nanjing, 210096, P.R., China lei_zhao@seu.edu.cn

[4]Wei Zhou is with School of Automation, Nanjing University of Science and Technology, Jiangsu District, Nanjing, 210094, P.R., China vvgod@seu.edu.cn

[5]Stefan Nastic is with Distributed Systems Group, TU Wien, Argentinierstraße 8, 1040 Vienna, Austria snastic@dsg.tuwien.ac.at

[6]Chen Wang is with Intelligent Transportation System Research Center, Southeast University, 2 Southeast University Road, Jiangsu District, Nanjing, 210096, P.R., China wkobec@hotmail.com

## II. RELATED WORKS

### A. Surveillance View Datasets for TAD

As presented in Table I, a systematic review of current surveillance view datasets is provided for TAD tasks. Most of the traffic accident videos are acquired from the website based on real-world surveillance cameras, whereas a minority consist of synthetic datasets generated via gaming environments or traffic simulation systems.

Several datasets (e.g., MP-RAD [16], CTAD [15], Iowa DOT [12], IITH [11], and UCF Crime [10]) exclusively provide temporal annotations, demarcating the initiation and termination frames of traffic accidents. These datasets facilitate binary (0/1) classification of accident occurrences, which can differentiate accident videos from non-accident videos. Such annotation is conducive to real-time traffic accident alert systems (i.e., require expeditious detection and notification) with subsequent analytical processes executed manually by human operators. However, the absence of fine-grained information regarding accident categories or occurrence regions renders these datasets insufficient for sophisticated analytical tasks like directly locating the event in each frame.

Recently, several datasets offer fine-grained annotations, encompassing accident locations, accident categories and weathers. The TADS classifies traffic accidents into 12 distinct categories based on participant involvement patterns. It furnishes multidimensional annotations, incorporating spatial temporal information, accident categories, weathers, etc. However, it provides spatial annotations using gaze areas formed by eye-tracking, rather than precise bounding boxes or segmentation regions [17]. The FAD dataset demonstrates enhanced granularity, delineating accidents into 26 categories and annotating each instance with event-level spatial temporal information, accident categories, severity stratification, and weathers [13]. SO-TAD classifies accidents into four principal categories (i.e., vehicle–pedestrian, inter-vehicle, vehicle–two-wheeler, and single-agent accidents) [18]. TAD-2 categorizes accidents into four labels (i.e., collision, crash, rollover and victim) with object-level annotations [14]. These datasets with fine-grained annotation paradigms are promising to facilitate sophisticated TAD tasks such as spatial temporal occurrence region detection, thereby mitigating manual interventions (e.g., encompassing automatic liability determination and incident report generation). Nevertheless, there is still a lack of available datasets that encode fine-grained contextual information, thereby constricts the capacity to capture the temporal evolution and infer the causal relationships in event sequences.

As such, we propose a novel fine-grained dataset (STTAD) refining the categories of traffic accident, and introducing spatial temporal occurrence region of each event.

### B. Vision-Based TAD Methods

*1) Frame-Level TAD:* The objective of frame-level TAD involves identifying accident temporal windows within video sequences. You and Han introduced the novel paradigm of causality recognition in traffic accidents [19]. They adopted the Temporal Segment Network (TSN) [20] as a baseline for action classification and evaluated three models for action localization. Srinivasan et al. employed DETR for salient object detection in video sequences, including vehicles and bicycles, and subsequently utilized Random Forest classification methodologies to differentiate between accident and non-accident frames [21], [22]. Vijay et al. proposed a dual-branch CNN to extract spatiotemporal features from videos and used softmax for binary accident classification [16].

*2) Object-Level TAD:* Many object-level TAD methods identify and track traffic participants' motion trajectories to determine whether an accident has occurred. Basheer et al. used YOLOv5 and DeepSORT as vehicle detection and tracking models, assigning a unique ID to each vehicle to track and monitor its movement [23], [24]. Chand et al. utilized the Mask R-CNN framework for vehicle detection and applied a centroid tracking algorithm to follow the detected vehicles [25], [26]. Santhosh et al. generated pseudo-labels for normal and accident trajectories during the training phase, and then used Convolutional Neural Networks (CNNs) and Variational Autoencoders (VAEs) to classify trajectory features [27]. Despite *Object-Level TAD* exhibiting measurable efficacy to some extent, the lack of *Event-Level TAD* circumscribes the capacity of current models for sophisticated inference and interpretation.

As such, this study constructs a tailored Algorithm (STFN) for further event-level TAD and spatial temporal analysis using the proposed STTAD.

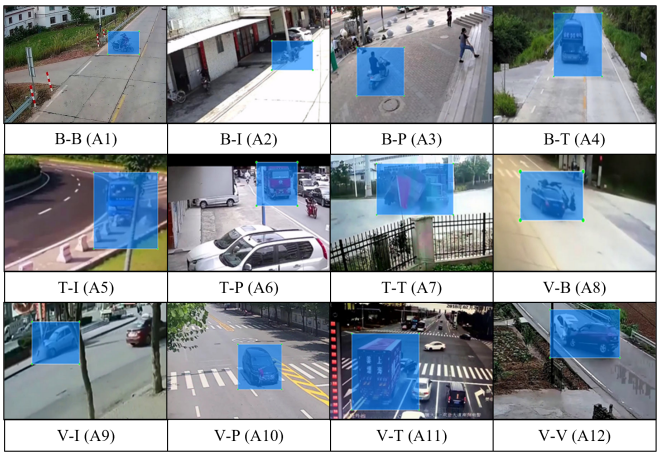


Fig. 1. Twelve accident categories and frame examples in STTAD

## III. DATASET

### A. Data collection and Accident categories

We collected surveillance-view road traffic accident videos of various categories from multiple video platforms. To ensure data quality, we retained only the accident segments and short time windows before and after the accidents, making sure each video contains a complete accident event. Our dataset comprises 12 categories of traffic accidents, with a total of 1,189 accident videos. We define the abbreviations

TABLE I

SURVEILLANCE VIEW DATASETS FOR DETECTION TASKS AND THEIR CHARACTERISTICS, INCLUDING ANNOTATIONS, NUMBER OF ACCIDENT CATEGORIES, NUMBER OF VIDEOS, NUMBER OF ACCIDENTS, OPEN SOURCE AVAILABILITY, SYNTHETIC (S) / REAL-WORLD (R), AND YEAR

| Datasets | Annotations* | Space | Classes | Videos | Accidents | Open Source | S/R | Year |
|---|---|---|---|---|---|---|---|---|
| UCF Crime [10] | T | / | 2 (0/1) | 1900 | 150 | ✓ | R | 2018 |
| IITH [11] | T | / | 2 (0/1) | 7 | 7 | ✗ | R | 2020 |
| Iowa DOT [12] | T | / | 2 (0/1) | 200 | 50 | ✗ | R | 2020 |
| FAD [13] | S, T, C | events | 26 | 3996 | 2393 | ✗ | R | 2022 |
| TAD-2 [14] | S, T, C | objects | 4 | 333 | 261 | ✓ | R | 2022 |
| CTAD [15] | T | / | 2 (0/1) | 1100 | 1100 | ✓ | S | 2023 |
| MP-RAD [16] | T | / | 2 (0/1) | 2000 | 400 | ✓ | S | 2023 |
| TADS [17] | S, T, C | gaze areas | 12 | 966 | 966 | ✓ | R | 2024 |
| SO-TAD [18] | T, C | / | 4 | 2186 | 282 | ✓ | R | 2024 |
| STTAD (Ours) | S, T, C | events | 12 | 1189 | 1189 | ✓ | R | 2025 |

*S: Spatial — provides spatial information; T: Temporal — provides temporal information; C: Categories — distinguishes different accident types.

as follows: "B" represents electric bikes, "I" represents road infrastructure, "P" represents bicycles and pedestrians, "T" represents large vehicles like trucks, "V" represents small vehicles. For example, "V-I" indicates a collision between a small vehicle and a roadside barrier, or a single-vehicle rollover. Frame examples corresponding to each accident category are shown in Fig. 1.

### B. Annotation attributes

Firstly, we formally define the concept of the Spatial Temporal Occurrence Region. Spatial Temporal Occurrence Region is our detection target, representing the sequence of frames during which a traffic accident occurs and the corresponding spatial region within each frame where the accident takes place. In contrast to traditional object detection that focuses on identifying traffic participants such as vehicles or pedestrians, Spatial Temporal Occurrence Region is designed to directly localize and characterize the accident event itself in both time and space. We represent the Spatial Temporal Occurrence Region using bounding boxes.

We perform the spatial temporal annotation of each traffic accident video using the LabelImg tool. In STTAD dataset, we have extensively annotated the attribute information of each traffic accident contained in the videos. The annotated attributes include Video number, Accident start frame, Accident end frame, Spatial Temporal Occurrence Region, and Accident categories. Table II presents the detailed information of these annotation attributes. Fig. 2 illustrates an example of a V-V (A12) accident, with annotations for the accident's start and end points. Thereinto, T1 corresponds to the first frame of the video, T2 represents the accident start frame, and T3 indicates the accident end frame. The interval from T2 to T3 constitutes the accident window.

### C. Training and testing sets splitting

When splitting the training set and testing set, a stratified random sampling strategy was adopted to ensure a balanced distribution of accident categories across both subsets at a ratio of 7:3. Then Fig. 3 illustrates the twelve accident categories along with their corresponding quantities and the quantitative proportion of each accident category within the training and testing sets.

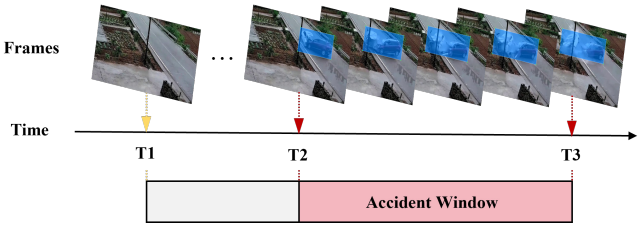


Fig. 2. An example of a V-V accident

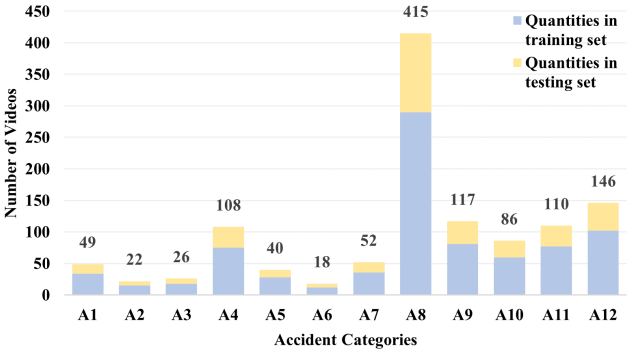


Fig. 3. Proportion of each accident category in training and testing set.

## IV. ALGORITHM DESIGN

In Section IV-A, we introduce the 3D backbone utilized for spatiotemporal feature extraction, as well as the 2D backbone adopted by the tailored Spatial Temporal Fusion Network (STFN). In Section IV-B, we describe our feature fusion method, which integrates the spatiotemporal features extracted in Section IV-A. The overall framework of our proposed STFN is shown in Fig. 4.

### A. 2D and 3D Backbone

Spatial feature extraction is the first step in STFN framework. To capture multiple spatial features, we adopt YOLOv8 as the 2D backbone. Furthermore, we only apply up-sampling operations on the spatial features obtained by the 2D backbone, without applying additional convolution operations to accelerate the inference efficiency.

Upon the completion of both the backbone and enhanced feature extraction networks, a 1×1 convolution is applied

TABLE II

ANNOTATION ATTRIBUTES OF TRAFFIC ACCIDENT VIDEOS

| Attributes | Details |
|---|---|
| Video number | A unique number to each video contained in the STTAD. |
| Accident start frame | The location of the frame from which the two entities involved in the accident began to make contact. |
| Accident end frame | The location of the frame from which the two entities involved in the accident begin to cease relative motion after the collision, or the last frame of the video. |
| Spatial Temporal Occurrence Region | The bounding box of the traffic accident occurrence region, covering both involved parties, used to directly localize and characterize the accident event itself in both time and space. |
| Accident categories | Divided into 12 categories according to the type of the entities involved in the accident. |

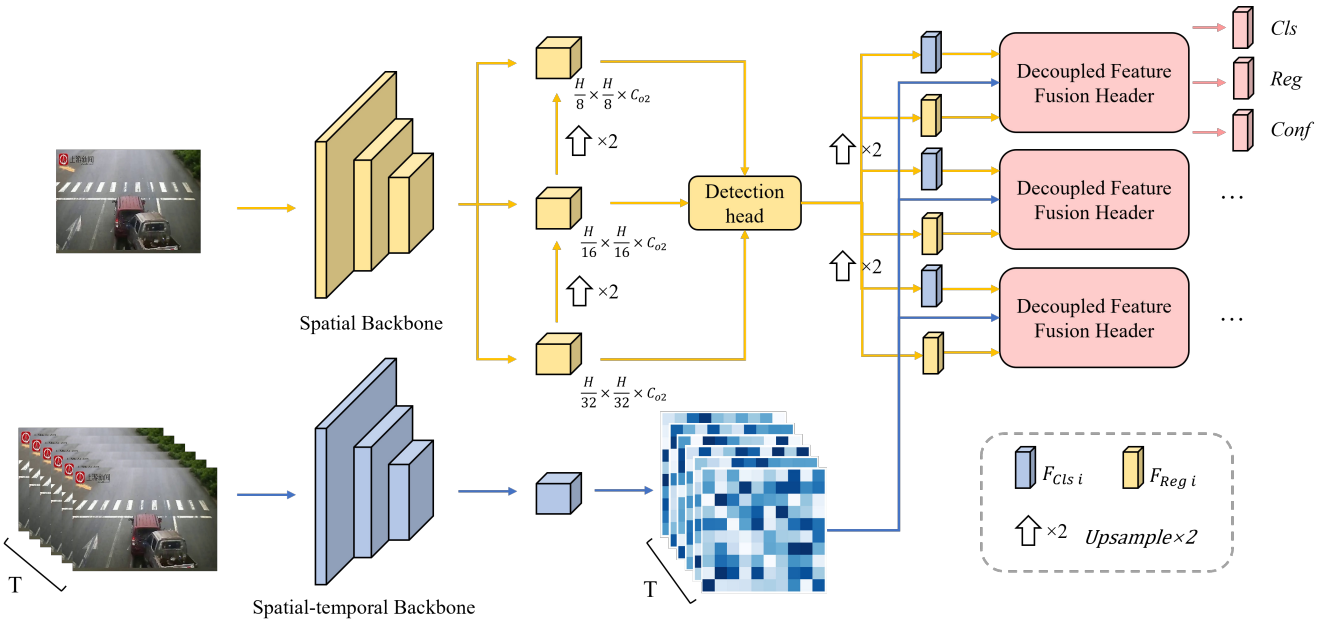


Fig. 4. Structure of the STFN framework. The input video frames are processed separately by the 2D backbone and the 3D backbone to extract multiple features. The 2D backbone applies a feature pyramid to capture multi-level features. T represents the number of frames in the input video, which is 16 in this study. Up-sampling is utilized to combine the spatiotemporal features and the spatial features.

to compress the channel dimensions of the spatial features derived by YOLOv8, reducing the channel dimension of each feature level $F_{\text{stem } i}$ to 256. Next, we perform channel decoupling and utilize two parallel 3×3 standard convolutions, each applied twice, to extract decoupled features. Subsequently, a standard 1×1 convolution is employed to fuse the separated feature channels back into 64 dimensions, which helps accelerate the convergence. The decoupling process is illustrated in Eq. (1).

$$\begin{aligned} F_{\text{stem } i} &= f_{\text{conv}1\times1}\left(F_{\text{level } i}\right) \\ F_{\text{head } i} &= f_{\text{conv}1\times1}\left(f_{\text{conv}3\times3}\left(f_{\text{conv}3\times3}\left(F_{\text{stem } i}\right)\right)\right) \end{aligned} \tag{1}$$

A pre-trained YOLOv8 model based on the COCO dataset [28] is utilized for training the STFN framework. Specifically, only the pre-trained weights of the 2D backbone are loaded into the spatial feature detection branch of the STFN. For the 3D backbone, we adopt efficient 3D CNN architectures to capture temporal features without significantly increasing model complexity. This is accomplished by extending classical lightweight networks into 3D domain, replacing 2D convolutions and pooling operations with their 3D counterparts. This design enables the 3D branch to effectively capture spatiotemporal features. Finally, we apply upsampling to the output feature layer to facilitate the fusion of spatiotemporal features with the derived spatial features through concatenation.

### B. Channel Fusion and Attention Convolution Mix module

Feature fusion is a critical step in STFN framework, aiming to effectively combine spatiotemporal features into unified representations to enhance the detection capability. In this study, we design a Channel Fusion and Attention Convolution Mix (CFACM) module, which integrates convolutional operations with a self-attention mechanism to achieve comprehensive feature fusion. The structure of CFACM module is illustrated in Fig. 5. By incorporating both the standard convolution and simulated Transformer branches, the model is able to capture local receptive field information effectively while simultaneously perceiving long-range dependencies, achieving a balanced and enriched representation of spatial temporal features.

As shown in Fig. 5, we combine convolution and self-attention mechanisms to effectively fuse spatial and temporal features while minimizing computational overhead and parameter complexity. Initially, two standard $3\times3$ convolutions

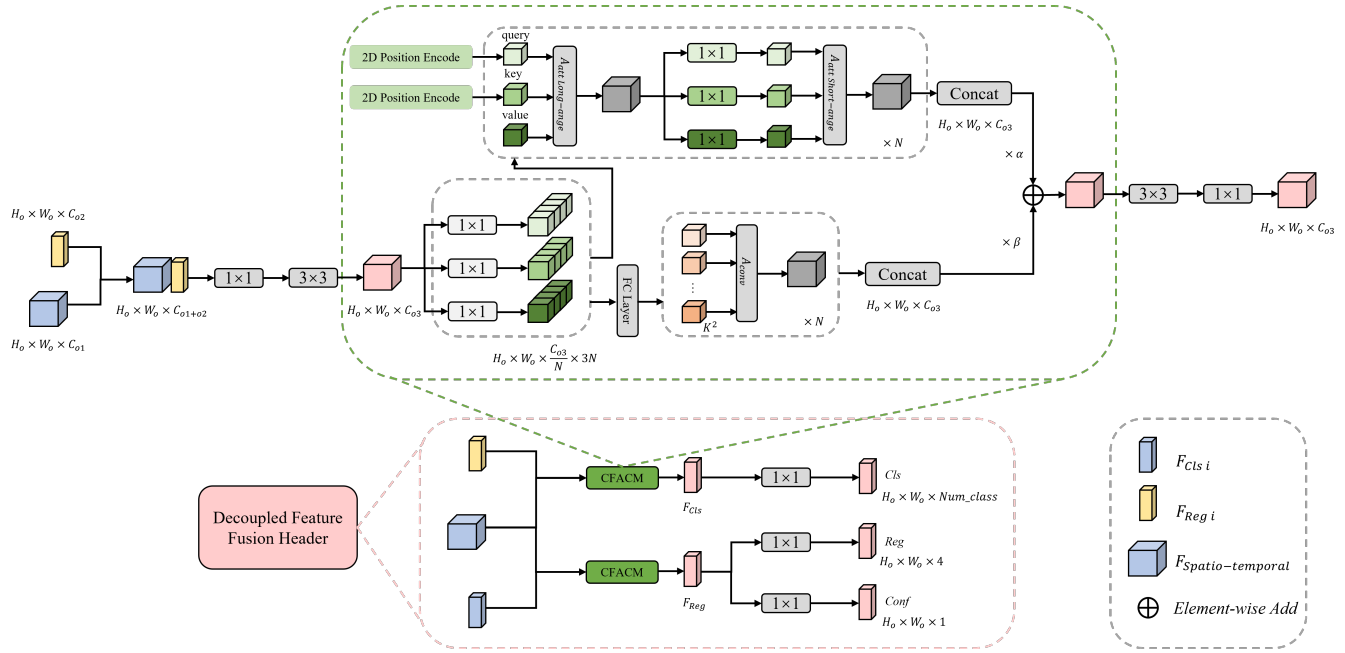


Fig. 5. Structure of the CFACM Module. The Decoupled Feature Fusion Header inputs $F_{\text{Cls } i}$ and $F_{\text{Reg } i}$ are fused with the derived features utilizing the CFACM module. 2D spatial encoding is performed before the computation of the self-attention branch. Then, the self-attention and convolution branches are concatenated by applying adaptive weights $\alpha$ and $\beta$.

(CBR) are applied to extract features from the concatenated inputs. Then, three $1 \times 1$ convolutions map the spatial temporal features, which are subsequently reshaped into $N$ segments, resulting in a set of intermediate features with $3 \times N$ feature maps. These intermediate features are then split into two branches, both sharing the outputs of the three $1 \times 1$ convolutions. One branch carries out self-attention by grouping intermediate features into $N$ sets, with each set comprising three features serving as query, key, and value, adhering to the standard multi-head self-attention structure.

Shift and aggregation mechanisms are utilized to capture information from the local receptive field in the self-attention computation branch, similar to traditional convolution. Convolutional operations are performed in a parallel branch. A fully connected layer first generates $K^2$ feature maps, followed by a $3 \times 3$ convolution to merge these $K^2$ features. The resulting output is then concatenated with the self-attention branch. Additionally, to reduce computational cost, we divide the window and apply a cross-sparse self-attention mechanism. As shown in Fig. 6, based on window division, the projected $Q$, $K$, and $V$ are partitioned into $P_L$ groups, where each group contains $Q_L$ positions, resulting in $P_L$ local groups $\{X_p\}_{p=1}^{P_L}$. The computations of the attention weights $A_p$ and enhanced features $Z_p$ are provided in Eq. (2).

$$\begin{aligned} A_p &= \text{Softmax}\left(\frac{Q_p^T K_p}{\sqrt{d}}\right) \in \mathbb{R}^{Q_L \times Q_L} \\ Z_p &= A_p V_p \in \mathbb{R}^{Q_L \times C'} \end{aligned} \tag{2}$$

Here, $Q_p$, $K_p$, and $V_p$ denote the query, key, and value features corresponding to the $p$-th divided window; $d$ refers to the dimensionality of $Q_p$, and $C'$ indicates the number of feature channels following the projection. While the adoption of window partitioning for local self-attention computation effectively reduces computational complexity, it also restricts the flow of global contextual information. To address this limitation, the second stage introduces a feature mixing operation across windows using the Permute method, followed by another round of self-attention computation within each window. This design enables the model to capture long-range dependencies across different windows. The attention mechanism computation in this second stage is formally defined in Eq. (3).

$$\begin{aligned} A_q &= \text{Softmax}\left(\frac{Q_q^T K_q}{\sqrt{d}}\right) V_q \\ &\in \left(\mathbb{R}^{Q_S \times Q_S} \times \mathbb{R}^{Q_S \times C'}\right) \end{aligned} \tag{3}$$

Here, $Q_q$, $K_q$, and $V_q$ correspond to the query, key, and value features of the segmented window; $d$ signifies the dimensionality of $Q_q$; $C'$ indicates the number of feature channels following the projection; and $Q_s$ indicates the spatial positions per group resulting from the second-stage window partitioning.

## V. EXPERIMENTS AND RESULTS

### A. Evaluation Metrics

Mean Average Precision (mAP) is a commonly used metric for evaluating object detection models. It measures performance by averaging the precision across categories. The overlap between predicted and ground truth boxes is quantified by Intersection over Union (IoU). We focus on

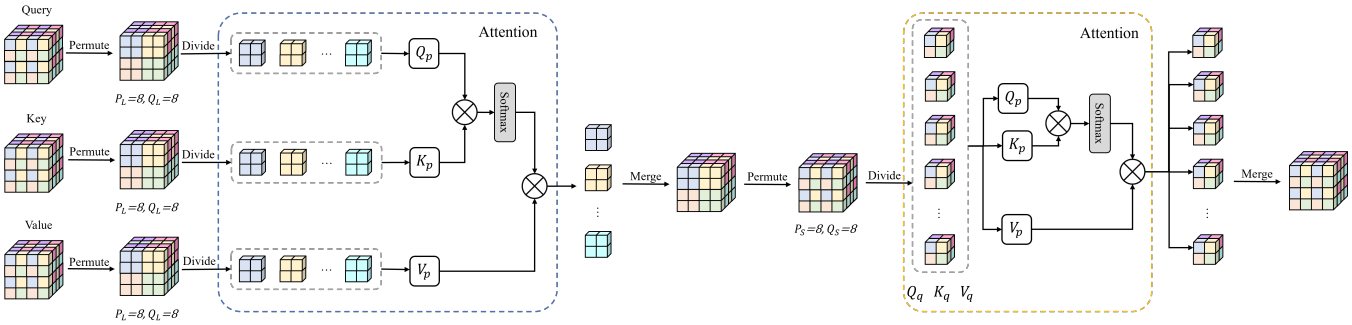


Fig. 6. Self-attention computing branch of CFACM Module. The calculation process of the two attention mechanisms is consistent, but their inputs $Q$, $K$, $V$ come from different sources. In the first self-attention calculation, $Q$, $K$, $V$ come from three different features. In the second self-attention calculation, $Q$, $K$, $V$ come from the same feature that has been integrated.

accident classification (mAP#12). For Spatial Temporal Occurrence Region detection, we report mAP at IoU thresholds of 0.33, 0.5, 0.66, and 0.75 (mAP@33, mAP@50, mAP@66, mAP@75). A region is considered positive if its IoU exceeds the threshold; otherwise, it is negative.

### B. Implementation Details

During the training process, we train the network using the PyTorch framework on an NVIDIA RTX 4090 GPU (24GB). On the proposed STTAD dataset, we use the AdamW optimizer with an initial learning rate of 1×10-4 for 20 epochs, 10-5 for 40 epochs, 10-6 for 40 epochs as a total of 100 epochs with the batch size set to 32. For model input, we evenly select 16 frames of each video clip as the initial $N_i$ images and resize each frame to 224 × 224 in width and height. To alleviate model overfitting, we set Dropout to 0.5.

### C. Spatial Temporal Occurrence Region Detection

We evaluate the performance of the proposed method with various 3D backbone networks. The results indicate that employing I3D [29] as the backbone enables our STFN algorithm to achieve the highest performance across all evaluation metrics, including mAP@75, mAP@66, mAP@50, and mAP@33. Subsequently, we compare STFN with baseline models on the test set of the STTAD dataset for the occurrence region detection task. The mAP#12 results under different IoU thresholds, used to determine whether a region detection is considered correct, are presented in Table III.

However, as shown in the table, the performance of occurrence region detection remains unsatisfactory, with the best mAP@75 reaching only 23.6% using our model. Although our model achieves 48.0% on mAP@33, this result is still considerably lower compared to performance on the same metric in conventional object detection tasks. These findings highlight the inherent difficulty of event-level occurrence region detection, which requires comprehensive understanding of the entire accident process. Furthermore, we compare the AP of each accident category under different thresholds in Fig.7, where the evaluation is restricted to the 12 accident categories, i.e., A1∼A12. By contrasting Fig.3 and Fig. 7, it is evident that categories with larger amounts of data tend to achieve higher detection accuracy.

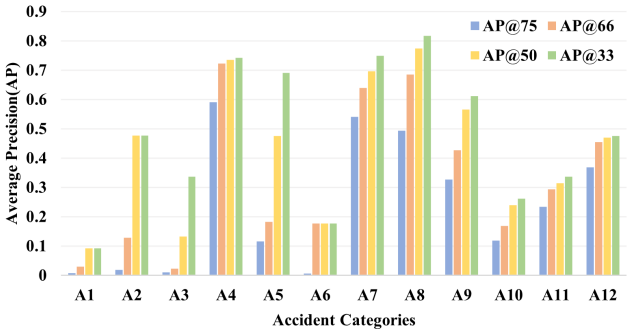


Fig. 7. AP of each accident category under different thresholds.

The reason for this decrease in performance is due to the variety of accidents. Even with the refined definition of the accident category, the fine-grained accident classification is quite challenging. The distribution of high and low performance on specific accident categories is roughly like the distribution of data quantity in the proposed dataset. This indicates that further increasing the samples of each fine-grained accident category could effectively learn the occurrence mode of accidents discriminatively.

### D. Model Visualization

To further demonstrate the effectiveness of the proposed model, we visualize representative detection results of accident occurrence regions on the STTAD testing set. For clearer comparison, we select the best-performing version of each method. As shown in Fig. 8, while our model can effectively localize the occurrence region, the predicted bounding boxes may not tightly align with the actual accident area. This indicates that accurate detection of occurrence regions in video remains a challenging problem requiring further investigation.

We further provide the temporal range of detected accident window to demonstrate the accuracy of temporal boundary judgment (see Fig. 9). Due to the strong temporal modeling ability brought by STFN, it can accurately judge whether an accident occurs in each video clip so that accident windows with accurate temporal ranges can be obtained. As illustrated in Fig. 9, the proposed STFN demonstrates a strong capacity

TABLE III
COMPARISON WITH THE METHODS ON VARIOUS AVERAGE PRECISION ON THE TESTING SET OF THE STTAD DATASET

| Methods | Backbone | Params | mAP@75 | mAP@66 | mAP@50 | mAP@33 |
|---|---|---|---|---|---|---|
| YOWO [30] | YOLOv2 + RESNEXT101 | 121.04 M | 0.135 | 0.157 | 0.201 | 0.273 |
| YOWOv2-Medium [31] | YOLO_free_large + shufflenetv2_2x | 52.0 M | 0.168 | 0.205 | 0.278 | 0.312 |
| YOWOv2-Large [31] | YOLO_free_large + RESNEXT101 | 109.7 M | 0.198 | 0.287 | 0.353 | 0.387 |
| STFN(Ours) | YOLOv8 + Mobilenet V2_1x [32] | 89.69 M | 0.231 | 0.317 | 0.371 | 0.402 |
| | YOLOv8 + ShuffleNet V2_1.5x [33] | 89.49 M | 0.217 | 0.298 | 0.349 | 0.393 |
| | YOLOv8 + I3D [29] | 99.23 M | **0.236** | **0.327** | **0.425** | **0.480** |

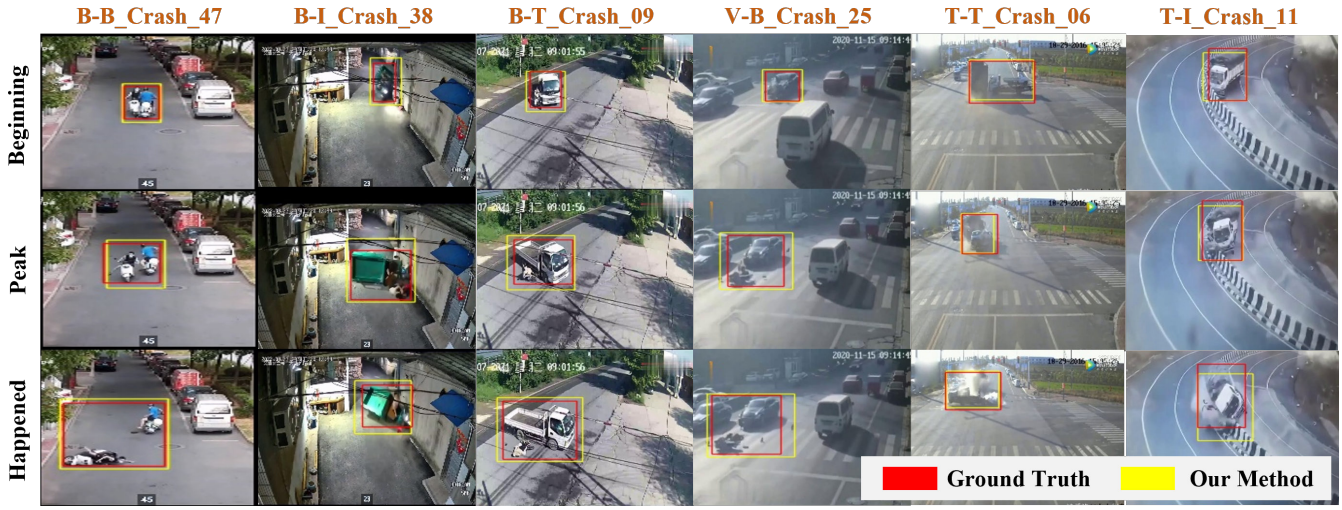


Fig. 8. Visualized comparisons of occurrence region detection on the testing set of the STTAD dataset. The sequence numbers are shown at the top. Each row shows the beginning, peak, and happened state of the accident respectively. The magenta bold bounding boxes indicate the ground truth. The colored boxes indicate the results of the corresponding methods respectively. Our method can estimate a more accurate accident occurrence region.

to accurately localize traffic accidents and delineate their temporal boundaries (i.e., the onset and termination times), even in the presence of complex environment and visual challenges. Specifically, the model maintains robust performance under conditions of severe occlusion, where critical visual cues may be partially or entirely obscured; under low-light scenarios such as nighttime scenes, which typically impair feature visibility and compromise conventional detection frameworks; and within complex traffic environments characterized by dense vehicular flow, heterogeneous agent behaviors, and intricate scene dynamics. These results highlight the ability of STFN to learn discriminative spatiotemporal representations and to generalize effectively across diverse real-world traffic conditions.

## VI. CONCLUSION AND FUTURE WORK

In this study, we introduced the challenging STTAD dataset for traffic surveillance scenarios and defined a detailed task for accident analysis, specifically spatio-temporal occurrence region detection. To address fine-grained TAD, we proposed a tailored algorithm named STFN. We also present the performance of our model alongside baseline models on the STTAD dataset, demonstrating the effectiveness of our approach in fusing video features for each task. We believe STFN can serve as a robust baseline for future research. Furthermore, the experimental results highlight that fine-grained accident detection is more complex than traditional accident analysis tasks, such as binary classification or object detection. In future work, given the spatial temporal nature of traffic accidents, incorporating scene priors should be a key focus for fine-grained accident analysis.

## REFERENCES


[1] K. K. Santhosh, D. P. Dogra, and P. P. Roy, "Anomaly detection in road traffic using visual surveillance: A survey," *Acm Computing Surveys (CSUR)*, vol. 53, no. 6, pp. 1–26, 2020.

[2] J. Fang, J. Qiao, J. Xue, and Z. Li, "Vision-based traffic accident detection and anticipation: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 1983–1999, 2023.

[3] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part IV 13*, pp. 136–153, Springer, 2017.

[4] W. Zhou, L. Wen, Y. Zhan, and C. Wang, "An appearance-motion network for vision-based crash detection: Improving the accuracy in congested traffic," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 12, pp. 13742–13755, 2023.

[5] Y. Chen, C. Wang, and Y. Xie, "Modeling the risk of single-vehicle run-off-road crashes on horizontal curves using connected vehicle data," *Analytic Methods in Accident Research*, vol. 43, p. 100333, 2024.

[6] L. Zhao, W. Zhou, S. Xu, Y. Chen, and C. Wang, "Multi-agent trajectory prediction at unsignalized intersections: An improved generative adversarial network accounting for collision avoidance behaviors," *Transportation Research Part C: Emerging Technologies*, vol. 171, p. 104974, 2025.


Fig. 9. Visualizations of accident windows. The sequence numbers are shown at the top. The magenta bold bounding boxes indicate the ground truth. The yellow band and pink band indicates the detected and ground truth accident window respectively. The blue dashed box indicates a case of severe occlusion between vehicles during the accident window


[7] S. Xu, M. Li, W. Zhou, J. Zhang, and C. Wang, "An evolutionary game theory-based machine learning framework for predicting mandatory lane change decision," *Digital Transportation and Safety*, vol. 3, no. 3, pp. 115–125, 2024.

[8] S. Xu, X. Xie, C. Wang, and J. Yan, "On the safety effects of off-peak hour speed characteristics of urban arterials," *Multimodal Transportation*, vol. 4, no. 2, p. 100206, 2025.

[9] Y. Chen, Y. Xie, C. Wang, L. Yang, N. Zheng, and L. Wu, "Time-dependent effect of advanced driver assistance systems on driver behavior based on connected vehicle data," *Analytic Methods in Accident Research*, vol. 45, p. 100370, 2025.

[10] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479–6488, 2018.

[11] D. Singh and C. K. Mohan, "Deep spatio-temporal representation for detection of road accidents using stacked autoencoder," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 879–887, 2018.

[12] M. Naphade, Z. Tang, M.-C. Chang, D. C. Anastasiu, A. Sharma, R. Chellappa, S. Wang, P. Chakraborty, T. Huang, J.-N. Hwang, *et al.*, "The 2019 ai city challenge.," in *CVPR workshops*, vol. 8, p. 2, 2019.

[13] H. Yu, X. Zhang, Y. Wang, Q. Huang, and B. Yin, "Fine-grained accident detection: database and algorithm," *IEEE transactions on image processing*, vol. 33, pp. 1059–1069, 2024.

[14] Y. Xu, H. Hu, C. Huang, Y. Nan, Y. Liu, K. Wang, Z. Liu, and S. Lian, "Tad: A large-scale benchmark for traffic accidents detection from video surveillance," *IEEE Access*, 2024.

[15] H. Luo and F. Wang, "A simulation-based framework for urban traffic accident detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[16] T. K. Vijay, D. P. Dogra, H. Choi, G. Nam, and I.-J. Kim, "Detection of road accidents using synthetically generated multi-perspective accident videos," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 2, pp. 1926–1935, 2022.

[17] Y. Chai, J. Fang, H. Liang, and W. Silamu, "Tads: a novel dataset for road traffic accident detection from a surveillance perspective," *The Journal of Supercomputing*, vol. 80, no. 18, pp. 26226–26249, 2024.

[18] X. Chen, H. Xu, M. Ruan, M. Bian, Q. Chen, and Y. Huang, "Sotad: A surveillance-oriented benchmark for traffic accident detection," *Neurocomputing*, vol. 618, p. 129061, 2025.

[19] T. You and B. Han, "Traffic accident benchmark for causality recognition," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pp. 540–556, Springer, 2020.

[20] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*, pp. 20–36, Springer, 2016.

[21] A. Srinivasan, A. Srikanth, H. Indrajit, and V. Narasimhan, "A novel approach for road accident detection using detr algorithm," in *2020 international conference on intelligent data science technologies and applications (IDSTA)*, pp. 75–80, IEEE, 2020.

[22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020.

[23] M. I. Basheer Ahmed, R. Zaghdoud, M. S. Ahmed, R. Sendi, S. Alsharif, J. Alabdulkarim, B. A. Albin Saad, R. Alsabt, A. Rahman, and G. Krishnasamy, "A real-time computer vision based approach to detection and classification of traffic incidents," *Big data and cognitive computing*, vol. 7, no. 1, p. 22, 2023.

[24] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*, pp. 3645–3649, IEEE, 2017.

[25] D. Chand, S. Gupta, and I. Kavati, "Computer vision based accident detection for autonomous vehicles," in *2020 IEEE 17th India Council International Conference (INDICON)*, pp. 1–6, IEEE, 2020.

[26] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[27] K. K. Santhosh, D. P. Dogra, P. P. Roy, and A. Mitra, "Vehicular trajectory classification and traffic anomaly detection in videos using a hybrid cnn-vae architecture," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 11891–11902, 2021.

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755, Springer, 2014.

[29] J. Carreira and 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*, pp. 6299–6308, 2017.

[30] O. Köpüklü, X. Wei, and G. Rigoll, "You only watch once: A unified cnn architecture for real-time spatiotemporal action localization," *arXiv preprint arXiv:1911.06644*, 2019.

[31] Z. Jiang, J. Yang, N. Jiang, S. Liu, T. Xie, L. Zhao, and R. Li, "Yowov2: A stronger yet efficient multi-level detection framework for real-time spatio-temporal action detection," in *International Conference on Intelligent Robotics and Applications*, pp. 33–48, Springer, 2024.

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[33] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, 2018.