IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 November 2024, accepted 17 December 2024, date of publication 25 December 2024, date of current version 3 January 2025.

Digital Object Identifier 10.1109/ACCESS.2024.3522384

RESEARCH ARTICLE

TAD: A Large-Scale Benchmark for Traffic Accidents Detection From Video Surveillance

YAJUN XU[1,2,*], HUAN HU[1,2,*], CHUWEN HUANG[1,2], YIBING NAN[1,2], YUYAO LIU[3], KAI WANG[1,2], ZHAOXIANG LIU[1,2], AND SHIGUO LIAN[1,2], (Member, IEEE)

[1]AI Innovation Center, China Unicom, Beijing 100013, China
[2]Unicom Digital Technology, China Unicom, Beijing 100013, China
[3]School of Information Science and Technology, Tsinghua University, Beijing 100084, China

Corresponding authors: Yibing Nan (nanyb5@chinaunicom.cn) and Shiguo Lian (liansg@chinaunicom.cn)

*Yajun Xu and Huan Hu are co-first authors.

ABSTRACT Automatic traffic accident detection has attracted the attention of the machine vision community for the rapid development of autonomous intelligent transportation systems (ITS). However, previous studies in this domain have been constrained by small-scale datasets with limited scope, impeding their effectiveness and applicability. Specifically, highway traffic accidents, often resulting in severe consequences due to higher speeds, require a more comprehensive approach to detection. The use of video surveillance provides a unique perspective, capturing the entire accident sequence. Unfortunately, existing traffic accident datasets are either not sourced from surveillance cameras, not publicly available, or not tailored for highway scenarios. An open-sourced traffic accident dataset with various scenes from surveillance cameras is in great need and of practical importance. To fulfill the above urgent need, we endeavor to collect abundant video data of real traffic accidents and propose a large-scale traffic accidents dataset, named TAD. Various experiments on image classification, video classification, and object detection tasks, using public mainstream vision algorithms or frameworks are conducted in this work to demonstrate the performance of different methods. The proposed dataset together with the experimental results are presented as a new benchmark to improve computer vision research, especially in ITS. The dataset is publicly available at https://github.com/UnicomAI/UnicomBenchmark/tree/main/TADBench.

INDEX TERMS Traffic accidents, large-scale, surveillance cameras, open-sourced.

I. INTRODUCTION

Over the past decades, automatic traffic accident detection, has become a prominent subject in machine vision and pattern recognition due to its tremendous application potential in developing autonomous intelligent transportation systems (ITS). Traffic accidents are among the life threatening issues facing rural as well as urban communities. Accident detection is crucial for building ITS. According to the statistics from the National Statistics Bureau, the death rate from accidents per 10,000 vehicles stands at 1.57, continuing to pose threats to people's lives. However, accurate identification of traffic accidents is conducive to avoiding secondary accidents, which will effectively improve traffic safety. In light of these applications, researchers in academia and industry are gearing efforts to detect accidents early and potentially reduce their damaging consequences. A significant amount of research has focused on detecting traffic accidents through visual information. Despite their usefulness, those targeting traffic accidents are facing several challenges in data as follows:

First, lack of enough background visual information on accidents. Videos from dashboard cameras have been widely deployed in multiple studies [1], [2], [3], [4], [5] to provide information covering intervals before and after the accidents in the first-person perspective. However, despite their usefulness in addressing deficiencies in data tailored for accident detection, they remain limited in describing the environmental conditions of the involved vehicles, which is important for industrial applications in accident detection. With surveillance cameras installed located at a certain distance from the ground, video surveillance could provide more

The associate editor coordinating the review of this manuscript and approving it for publication was Shaohua Wan.

comprehensive features of traffic conditions such as positions of surrounding vehicles, traffic flow and even the weather, etc. Especially when more than two vehicles are involved in an accident, non-surveillance datasets would be obviously deficient in analyzing causes and recognizing victims. With the increasing popularity of Close Circuit Television (CCTV) in real life for monitoring traffic conditions, datasets from a third-person perspective are in greater need for the capability of better catering to current and future studies in developing smart traffic governance.

Second, small scale specified in accidents limits generalization. With the rapid advancement in deep learning [6], algorithms have evolved to achieve efficient feature learning while requiring high quality data. Traditional studies have categorized it as one of traffic anomalies. The University of Central Florida Crime (UCF-Crime) dataset [4] collected 151 traffic incident surveillance videos from Iowa during 2016-2017. Although it is massive in terms of all anomaly types, contains a rather small proportion of data specified in traffic incidents, which could not satisfy the accuracy of accident detection. Besides, datasets focusing on traffic accidents are less compared to normal activities, as facing resource-constraint problems.

Third, single type of accident classification. With the continuous development of traffic systems as well as increasing car parc, categories of accidents vary and keep growing. Existing datasets of accident detection [7], [8], [9], [10], though large-scale show singularity of types of accidents. For example, AI city competition (2019 track 3) [7] gives a specific perspective on vehicle anomalies while only focusing on stalled vehicle recognition. In fact, the various accidents could be broadly classified in terms of vehicle number or subjects involved, such as rollovers and collisions are often identified as one positive category while demonstrating distinct features of accidents. Accurate and trustworthy collection of such data plays an important role in decision-making for smart ITS applications, especially for those industry applications relying on surveillance cameras.

Fourth, limited scenes of occurred accidents. Currently, open-sourced data of traffic accidents mainly present as part of anomalies. However, researchers [8], [10] in this field often collect data from certain scenes such as urban roads or highways. Other types of roads including countryside and expressway are scarcely retained in relative works, while those are worth noting. Single scene restrains model generalization, thus impeding further development in accident detection.

To address the aforementioned problems in quantity and quality, we present a new dataset named TAD with specific annotations on traffic accidents, which serves as a benchmark to evaluate the robustness and applicability of traffic accident detection algorithms. Our main objective is to offer a benchmark dataset for traffic accident detection within the domain of computer vision, rather than exploring the specific particulars or fundamental causes of each accident. Despite the absence of specific accident descriptions, this design decision is intended to better facilitate the research and development of traffic accident detection algorithms. The dataset is publicly available at https://github.com/UnicomAI/UnicomBenchmark/tree/main/TADBench. The contributions of this work can be summarized as below:

A novel dataset with largest scale and richest types of accidents in the industry. The dataset we introduce in this paper consists of four types of accidents covering several scenes in real life, especially with the highway scene, which, to our knowledge, is the largest-scale open-sourced dataset focusing on various types of accident detection.

A benchmark based on experiments of mainstream algorithms. In this study, our goal is to provide a vision dataset that caters to traffic accident detection of the industry to build ITS. To achieve this purpose, we organize experiments based on three mainstream visual algorithms to ensure their practicality, including image classification, video classification and object detection. The experimental evaluation indicates TAD's capability to serve as a benchmark and possible applications in the future.

The rest of our paper is organized as follows: we first review related studies focusing on detection of traffic accidents in Section II. In Section III, we show that TAD is a large-scale, specified visual database with quality-controlled images and annotations, and details on dataset construction such as data source, structure, type and standard of annotation are all provided. In Section IV, we present several experiments on TAD as application examples. Our goal is to show that TAD can serve as a useful resource for visual recognition. Finally, the future application as well as discussion is provided in Section V and VI.

II. RELATED WORK

A. DATASETS OF TRAFFIC ACCIDENTS

Table 1 presents the existing traffic accident datasets along with their key characteristics. In the following sections, we will provide a detailed introduction to each of these datasets.

1) COMMA-SEPARATED VALUES (CSV) FORMAT

There are several datasets organized in CSV format providing information on traffic events, such as OpenStreetMap [11], US Accident [4], etc. These datasets usually consist of a variety of intrinsic and contextual attributes of traffic events such as time, latitude and longitude of an incident and sensor, casualties, vehicle type, road type, etc. Given these attributes, the objective of these datasets is more inclined to prediction than detection. In addition, apart from the giving indicators, it is often offered with information on traffic flow, weather, period-of-day, and points-of-interest.

2) CONTEXTUAL FORMAT

We find that in the field of traffic detection, the linguistic text description [12] belongs to the secondary processing of the information of the accident scene, thus there is inevitably a corresponding information loss. Furthermore, descriptions

of datasets vary from researcher to researcher, which may bring about the problem that different texts may differ greatly in their description of an identical accident scene, and thus accident scenes obtained through textual reconstruction may also be different. For a certain dataset, such differences are likely to cause many problems, e.g., inconsistent data caliber and relatively poor migratory capability of dataset applications. Considering these limitations, despite the current rapid development of model iterations in natural language processing (NLP) [13], there are few works and public detection results on accident detection using NLP models.

3) VISION FORMAT

Video data can provide more dimensions and direct information than previously mentioned CSV and text-based datasets. Obviously, video data directly visualizes accident scenes, and existing vision-based models are well able to perform image feature extraction works. Given a certain amount of annotated training data, the common features of traffic accidents can be modeled by designing appropriate neural networks, enabling accidents to be detected with relatively little effort.

Video itself has its own temporal property, and each frame can provide spatial information of the captured scene, it also has a spatial-temporal property, which is very helpful to achieve effective traffic prediction. By leveraging vision-based data [14] developed a detection system with YOLOv5 [15] and decision trees to recognize traffic anomaly events. Therefore, based on such data features, in order to make full use of the feature information of videos, existing algorithms [16], [17] usually use multi-dimensional features to extract information from videos. Many algorithms [18], [19], [20] utilize underlying model frameworks with these features to achieve better prediction results than the above CSV and contextual data.

Currently, video-based traffic accident datasets can be mainly divided into two categories: monitored and non-monitored perspective datasets. Visual datasets not collected from surveillance cameras are mostly derived from data provided by automatic driving recorders. References [1] and [9] introduced datasets from dashboard cameras respectively, while the former provided crowd-sourced videos, the latter labeled 1,500 video traffic accidents collected from YouTube. Reference [10] presented a new anomaly detection dataset, named the highway traffic anomaly (HTA) dataset in order to detect anomalous traffic patterns from dashboard cameras of vehicles on highways. Therefore, these datasets are limited in wider applications due to being difficult to migrate to other scenarios of traffic accident detection, except autonomous driving itself.

Datasets from the perspective of surveillance can be classified into general surveillance scenarios and highway surveillance scenarios. The former often includes residential streets, campuses, etc. Some studies are usually dominated by traffic anomalies datasets, such as the UCF-Crime video dataset [21] and AI city challenge dataset [22]. Anomalies defined in these datasets have a wide range of classes, mostly including crimes committed on streets such as shooting, stealing, vandalism, and robbery. Some of them also include pedestrians. However, traffic accidents are often not included or simply classified as one of the anomaly types, leaving specific information on vehicle anomalies less provided.

In summary, the data collected from video surveillance could provide a more macro, comprehensive and objective perspective of traffic accident detection, improving urban governance and the construction of intelligent transportation system applications. With the extensive use of surveillance cameras in public places, computer vision-based scene understanding has gained a lot of popularity amongst the CV research community. Visual data contains rich information compared to other information sources such as GPS, mobile location, and radar signals. Thus, it plays a vital role in detecting or predicting congestion, accidents and other anomalies apart from collecting statistical information about the status of road traffic. Moreover, open-sourced datasets for traffic accident monitoring in high-speed scenarios are also of great significance. Considering the advantages of surveillance video to capture a variety of realistic anomalies, the dataset we proposed is comprised of videos and images from surveillance cameras installed on various scenes, providing different information from the non-surveillance paradigm like the first-person view.

B. DETECTION ALGORITHMS OF TRAFFIC ACCIDENTS

The main purpose of our proposed dataset is to provide a benchmark and help improve research on traffic accident detection. Accordingly, this section mainly discusses and compares previous works on traffic accident detection algorithms. There are three major algorithm frameworks, including classification methods, causal inference methods, and object detection methods.

1) CLASSIFICATION ALGORITHM

In this field, researchers usually choose to develop a supervised learning method that works as a binary classifier to distinguish between images containing damaged vehicles as positive class and not ones as negative class [16].

As the research progressed, some works focused on positive samples with accidents and carried out the next level of classification refinement based on the severity of the accident occurrence [20]. Statistical methods both parametrical and non-parametrical can be found in recent studies [12]. Based on rules of statistics and data mining methods for classification decisions, widely used models are support vector machine (SVM) [23], Random Forest (RF) [24], Multilayer Perceptron (MLP) [25], Naive Bayes [26], etc. With the development of the deep learning (DL) method, DL techniques have been proven effective in visual classification tasks [27], [28], [29], [30]. Models based on Convolutional Neural Network (CNN) [31] such as Convolution 3-dimensional (C3D) [32] perform efficiently in traffic

TABLE 1. Summary of traffic accident datasets and their key characteristics.

Dataset Name	Key Characteristics
OpenStreetMap [11], US Accident [4]	Designed for predictive traffic analysis, tends towards prediction rather than detection, including intrinsic and contextual attributes (e.g., time, location, casualties, vehicle type, road type), often with additional data on traffic flow, weather, time of day, and points of interest.
Textual Description Data	Information loss due to secondary processing of accident scene details; inconsistent data standards; weak dataset transferability; limited research on accident detection using NLP models.
DADA-2000 Dataset [1]	Crowdsourced video; Limited by varying video quality and annotation consistency; difficult to migrate to other scenarios of traffic accident detection except autonomous driving itself.
YouTube Traffic Accident Dataset [9]	Crowdsourced video; Contains 1,500 labeled traffic accident videos; Quality inconsistencies, metadata unreliability, and copyright concerns limit its usefulness; difficult to migrate to other scenarios of traffic accident detection except autonomous driving itself.
Highway Traffic Anomaly (HTA) Dataset [10]	Uses dashboard cameras for detecting anomalous traffic patterns; Limited to highway scenarios, potentially lacks generalization to other road types, and may have subjective definitions of "anomaly".
UCF-Crime Video Dataset [21]	Contains a wide range of anomaly classes but traffic accidents are often not separately classified or simply treated as one anomaly type, lacking specific details on vehicular anomalies.
AI City Challenge Dataset [22]	Comprehensive dataset spanning multiple tasks including traffic flow analysis, pedestrian tracking, vehicle recognition; May not focus exclusively on a single type of anomaly detection (e.g., traffic accidents) due to its multi-task nature.
TAD (ours)	Novel dataset with the largest scale and richest types of accidents in the industry; covers four types of accidents across multiple real-life scenes, notably including highway scenes; provides a benchmark for evaluating the robustness and applicability of traffic accident detection algorithms; supports experiments with mainstream visual algorithms including image classification, video classification, and object detection, indicating its capability as a benchmark and potential future applications.

anomaly detection on surveillance videos by distinguishing positive and negative cases of traffic accidents. Methods of time series [33], [34] based classification models such as Recursive Neural Network (RNN) [35], Long Short-Term Memory (LSTM) [36] also reach high accuracy and recall rate.

2) CAUSAL INFERENCE ALGORITHM

Some researchers find that there is a close relationship between the occurrence of traffic accidents and driver behaviors as well as the driving environment before accidents happen. Studies in this field can make early anticipation of traffic accidents. By proposing definitions of risks influencing drivers' behaviors, [37] introduced a framework via casual inference and demonstrated favorable performance on the Honda Research Institute Driving Dataset. Reference [38] developed a dynamic spatial-temporal attention network using videos from dashboard cameras.

3) OBJECT DETECTION ALGORITHM

Object detection is one of the most important computer vision tasks that deals with detecting instances of visual objects of certain classes such as humans, animals, or cars in digital images.

The algorithms used in object detection tasks have achieved rapid development, from the histogram of oriented gradient (HOG) Detector [39] to deformable parts model (DPM) [40] and then to CNN-based models. Among them, RCNN [41], Faster-RCNN [42], RetinaNet [43], and YOLO series are considered milestones. The former two models use CNN-based two-stage Detectors, while YOLO [44] is one of the CNN-based One-stage Detectors. Due to advantages in addressing the problem of data imbalance, RCNN and YOLO are seen as milestones, winning wide popularity especially in industries. Reference [38] made comparative analysis among R-FCN [45], Mask R-CNN [46], SSD [47], and YOLOv4 [48], finding that YOLOv4 outperforms in accurately detecting difficult road targets under complex road scenarios and weather conditions in an identical testing environment. Reference [18] and [49] both used R-CNN models for object detection of accidents from surveillance perspective and achieved high detection rate and low false alarm rate. Reference [50] used YOLOv3 [51] as the base model to extract object features in video datasets, and carried out object detection of accidents including vehicle rollovers. YOLOv8 [52] and YOLOv9 [53], as the latest additions to the YOLO series. YOLOv8 leverages model scaling, anchor-free detection, and improved training strategies to achieve better

results across various benchmarks. YOLOv9 further refines the architecture and training procedures. It incorporates novel features like a more efficient backbone network, improved head design, and advanced data augmentation techniques. Additionally, recent research has focused on Transformer-based object detection algorithms, such as the DETR [54] series, which achieve full end-to-end object detection without the need for pre-defined anchors or Non-Maximum Suppression [55] post-processing, thereby enhancing detection efficiency and performance.

As for the application of datasets, video datasets are very suitable for object detection tasks because of their rich information content and direct expression compared to CSV and text datasets. The current algorithm models for traffic accident detection are relatively well developed, but there are bottlenecks in development due to the limitation of data sources. Therefore, the dataset provided in this paper is of great importance for future practical applications.

III. TRAFFIC ACCIDENTS DATASET (TAD)

This section describes how we constructed the TAD and presents its properties through statistical analysis and comparison versus existing traffic accident datasets. While constructing a traffic dataset is an arduous task, we endeavor to build a large-scale traffic accident dataset named TAD from surveillance perspective in various scenes. TAD contains serious traffic incidents caused by rain, vandalism or other factors, with a total of 344 videos covering 277 positive ones with traffic accidents and 127 negative ones without traffic accidents.

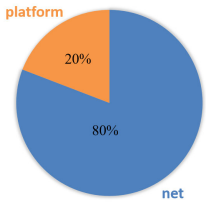
In Section III-A, we first describe the whole process of how TAD is developed, including the methods deployed in collecting data, extracting frames and generating labels set for tasks on classification and detection. In Section III-B, we detail the statistical characteristics of TAD from various perspectives through examples and explanations. We also compare TAD with several mainstream traffic accident datasets, as described in Section III-C.

A. CONSTRUCTION OF DATASET

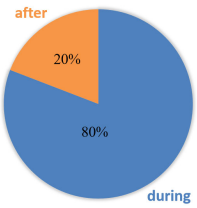
TAD is a large-scale traffic accident dataset from surveillance perspective, including serious traffic incidents caused by rain, vandalism or other accidental factors. As for the recording time, the accident data we provided includes both at the time of the accident and after the accident. From the perspective of collection sources, our accident data includes those downloaded from traffic video analysis platforms and those downloaded from mainstream video broadcasters such as Weibo [56]. Mainstream video broadcasters have become an important channel for obtaining instant information, especially during emergencies such as traffic accidents. Official media outlets often respond swiftly, posting on-site images, videos, and other intuitive information on these mainstream video platforms to help the public stay informed about developments. We use keywords such as ''car accident'', ''traffic accident'' and ''accident scene'' to conduct efficient searches on platforms like Weibo and quickly lock in relevant content. It is worth noting that the data we collect from platforms like Weibo is not merely user-uploaded, casual recordings. Instead, we carefully select the data to ensure that it originates from a surveillance perspective. This data may come directly from on-site surveillance footage or be high-quality reproductions of such footage, ensuring its authenticity and accuracy. From the perspective of accident types, there are collisions between vehicles, collisions between cars and pedestrians or cyclists who suddenly cross the road, collisions between cars and road obstacles or roadside fences, and rollovers caused by emergency braking and other factors. We select video data with more obvious accident characteristics from the visual point of view, and the statistical distribution results are shown in Table 2.

TABLE 2. A summary of our TAD videos in the number of different video sizes.

Size	Number	Average Time (s)	Average Resolution (px)
$< 1M$	92	10	(862,530)
$>= 1M \&\& < 5M$	165	20	(1382, 883)
$>= 5M \&\& < 10M$	16	40	(1490,923)
$>= 10M$	4	60	(1432,990)



(a) Data source distribution: displaying the ratio between internet videos and platform-sourced videos.



(b) Time window distribution: comparing videos captured during the accident occurrence with those recorded afterwards.

FIGURE 1. Overview distribution of TAD's data source.

1) COLLECTING DATA

a: DATA SOURCE

To ensure the quality of our dataset, we have edited video data, setting video time windows as before, during and after a traffic accident. We collect different categories of traffic accidents from various internet websites and continue to extend the dataset. The preset position of surveillance cameras is diverse at different times, it is hard to capture continuous video frames covering the complete process of

an accident. Therefore, videos acquired from our traffic video analysis platform mainly cover time windows after accidents happen. While videos collected from Internet video mostly provide time windows before and after an accident. In order to obtain as many kinds of accident data as possible, we utilize several search terms for retrieval, including ''vehicle wreck'', ''road accidents'', ''highway accidents'', ''non-motor vehicle accidents'', etc. In particular, videos including blurred images, pranks, minor cuts, and minor rear-end accidents are abandoned in the process of data collection. Meanwhile, final data retained from two ways generally contains traffic accidents that are serious and recorded completely. Accordingly, the ratio between data from the internet and our platform is 4:1 (Figure 1(a)), sharing the same ratio between data when an accident is happening and after an accident happens and (Figure 1(b)).

b: RESOLUTION

The resolution of videos we acquired ranged from (862,530) to (1432,990). Data collected from the video analysis platform developed by our teams usually consists of 2-3 video clips per accident scene with both near and far views. The mentioned platform can automatically analyze various traffic incidents in real-time from surveillance cameras. The abnormal events warnings such as pedestrians crossing the highway, vehicles parked on the side of the road and traffic accidents can be delivered in the format of captured images and video clips. We have access to download playback videos with high resolution from a network video recorder (NVR), the same resolution as original video streams. The resolution of data collected from Internet video platforms such as Weibo is rather lower.

2) CLEANING DATA

The videos we provided in our dataset have different memory sizes varying from less than 1M to over 10M. In Table 2, we display a summary of dataset distributions in the number of videos divided into different sizes. Videos less than 5M accounts for 92 % in total while the proportion of videos smaller than 1M is 33 %. The count of videos larger than 5M, or even 10M is less. They are mostly a collection of several clips of traffic events. The frame interval for videos less than 5M is set sparsely and the ones equal to or larger than 5M is set more dense. We apply two parameters setting of frame interval to balance image quantity between different sizes of videos, while capturing as much data as possible at the moment of traffic accident happens.

3) ANNOTATING DATA

TAD is constructed with multiple dimensions in three kinds of annotation in total, including video-level classification labels, image-level classification labels and image-level rectangular annotation boxes. In the first level, our goal is to divide video data into two groups, including the presence and absence of traffic accidents. Each traffic accident shot is a series of consecutive video clips containing the trajectories of moving accident targets. In the second level, we further filter more rigorously, so that images obtained from the same video clip would be divided into two candidate sets depending on whether existing obvious abnormal characters of accident targets in each frame. In general, Video frames before an accident happens are set as the ''normal'' category, and video frames that generate an accident or have generated an accident scene are considered as the ''accident'' category. In view of the low resolution that existed in some videos, from which the extracted frames could not demonstrate distinctive features of traffic accidents we set these frames as ''normal'' ones. For some frames in high-resolution videos, they still could not be distinguished from the ''accident'' ones. For example, some accident vehicle is obscured by passing vehicles. In the third level, we add annotating types in the format of location coordinates. In order to improve the quality of the dataset, we hire five professional workers to perform the labeling task. All annotators should make more pertinent and stable data annotation data for which individuals are responsible first. Then, we combine the whole results to unify the annotation level and verify the quality. The rectangular annotation boxes have four positive labels to indicate accident, including ''collision'', ''wreck'', ''roll over'' and ''victims'', as illustrated in Figure 2.

collision Collisions among motorized vehicles, such as collisions between two cars.

wreck

- Collisions between motorized and non-motorized vehicles, such as a collision between a car and a motorcycle.
- Collisions between pedestrian and motorized vehicles.
- Injured motorized vehicles such as trucks and cars.

roll over Rollovers.

victims Injured people as well as broken non-motorized vehicles such as bicycles.

B. DATASET STATISTICS

This section presents the results from statistical analysis, showing the distribution of data from different perspectives. We aim to give a clear illustration of TAD.

1) ACCIDENTS TYPES

There are a total of four types of accidents in our dataset including collisions between multiple vehicles, collisions between vehicles and bicycles/motorcycles, collisions between vehicles and inanimate entities and rollovers. Figure 3 shows some sampled typical video frames of the whole image sequences of four accident types. The red dotted line marks the accident area for each image. Some videos are video collections integrated with several accident clips, we count the number of each accident type by individual accident clips when plotting the distribution as depicted in Figure 5.

2) ACCIDENTS SCENES

There are totally four scenarios of accidents in TAD including cross streets, city roads, village roads, and highways. Figure 4

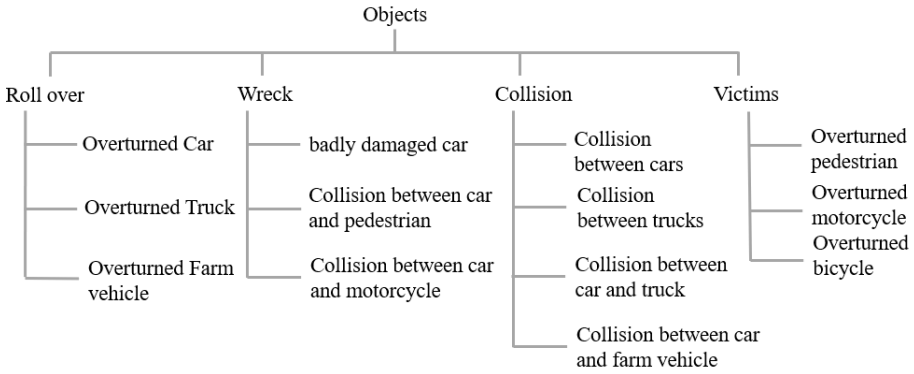


FIGURE 2. TAD classification tree.

TABLE 3. Comparison results of TAD with mainstream anomaly detection datasets. C: CCTV footage, H:highway scene, S:spatial annotation,T:temporal annotation.

Dataset	Total duration	Average #frames	#positives/#all	C	H	S	T
TAD	1.2hours	896	277/404	Y	Y	Y	Y
UCF-Crime	-	-	151/1900	Y	N	N	Y
CADP	5.2hours	366	1416/1416	Y	N	Y	Y
DAD	2.4hours	100	630/1730	N	N	Y	Y
HTA	-	-	-	N	Y	N	Y

illustrates typical samples for each scene with randomly selected images. It is worth noting that the same scenario contains different monitoring angles, enriching the visual information of each accident type.

Figure 5(b) shows that The number of data collected from ''highway'' scenes accounts for the largest among the four scenes. In other words, TAD is specifically built from the monitoring view of the highway, which addresses the challenging problem facing visual recognition studies where open-sourced datasets of traffic accidents under highway surveillance are rather rare. Generally, accidents are more likely to happen on cross streets due to the higher traffic flow and probability of abnormal activities such as red-light running.

3) ACCIDENTS SCALE

TAD aims to provide a wide coverage of the accidents happening in the real world. There are 277 videos, with 294 video clips totally provided in TAD. Figure 6 shows the histogram distribution of the number of images as well as objects per label.

C. TAD AND RELATED DATASETS

In this section, we present the properties of the traffic accidents dataset (TAD) in comparison to other popular traffic incidents datasets. The main differences are summarized in Table 3. Up to now, traffic accidents are often considered as one category of traffic anomaly detection and there are few open-sourced video datasets related to traffic accidents. Therefore, dataset comparison in this section targets traffic video datasets commonly applied in traffic anomaly detection.

1) THIRD-PERSON PERSPECTIVE

HTA and DAD provide video data of accidents captured by dashboard cameras mounted on driving vehicles, providing the first-person perspective. More information on surrounding conditions such as traffic flow, vehicles driving nearby is provided from CCTV footage with a wider view. UCF-Crime, CADP and our dataset can provide surveillance videos from third-person perspective like CCTV footage, which is critical to improving intelligent traffic system.

2) MUTIPLE ACCIDENT SCENES

Our collected videos mainly focus on accidents that happened on highways as shown in Figure 5(b). UCF-Crime, CADP and DAD, though large in scale, lack scenes recorded on highways. HTA and Ours record highway-level visual observations of traffic accidents. Other scenes ''cross street'', ''city road'', and ''village road'' generally seen are also included in Ours for comprehensive visual understanding as well as application.

3) COMPLETE ANNOTATION TYPES

Annotations at different levels of accidents dataset is rarely seen in recent public traffic accident datasets. UCF-Crime dataset concludes all types of accidents in one single category. So does HTA. CADP's annotations are made upon both temporal and spatial levels. Our dataset, TAD is also temporal-spatial level. Ours further provides detailed annotation in describing accident types, including rollovers,

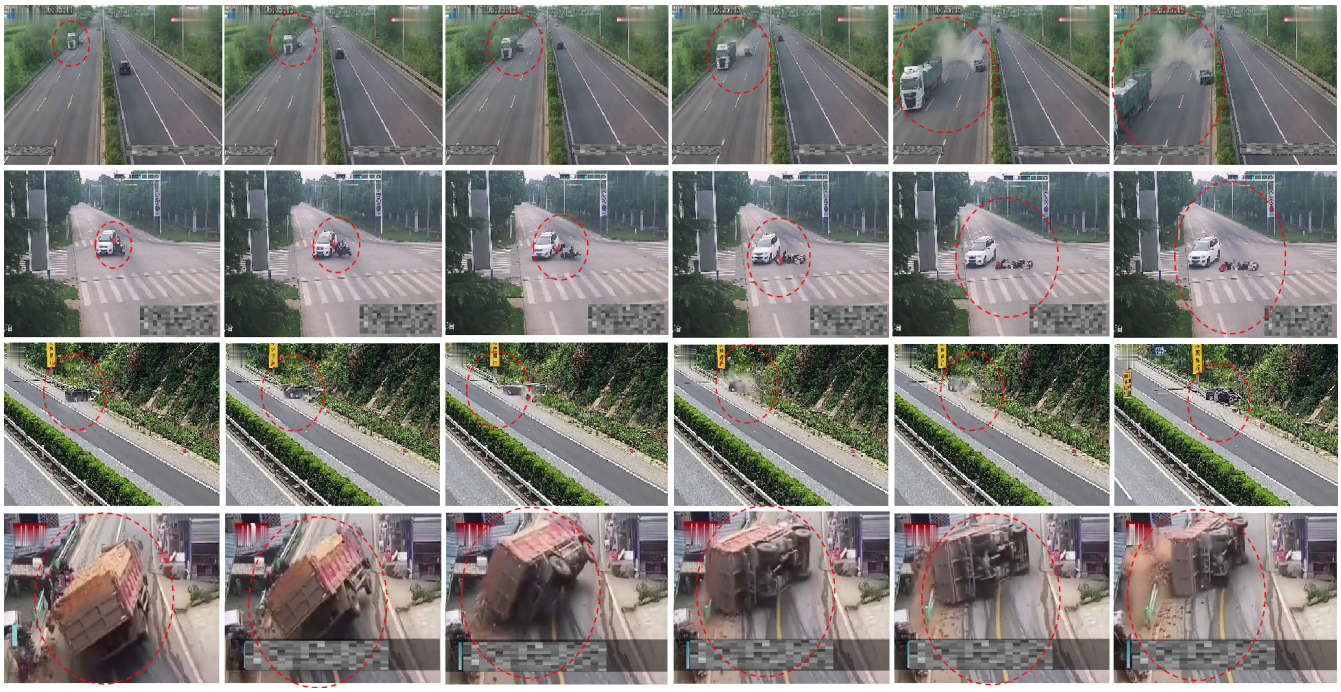


FIGURE 3. A snapshot of different kinds of traffic accidents in TAD.

collisions between vehicles, or other pedestrians, cyclists suddenly crossing, collision road obstacles, roadside fences, etc. This diverse visual information offers complementary characteristics in the training of more vision models.

4) SCALE AND LENGTH

There are 404 videos with 277 positive samples with accidents included in our dataset. The total duration of TAD is 1.2 hours in total. The longest video has 5.2 hours and DAD has 2.4 hours. But our dataset has the longest frame number on average. TAD is 896 frames per video, which means more complete process provided in each accident.

IV. TRAFFIC EVENTS BENCHMARK

In this section we make several experiments to test the performance of our dataset and provide meaningful references in actual application. Conducted experiments are organized in image classification, video classification and object detection tasks. In each task, we select recent classical mainstream algorithms and construct algorithm-related datasets from 404 video resources. In order to compare fairness, we choose fixed 32 video resources to test the accident recognition effect of different algorithms, so as to find the algorithm model that is more in line with the actual application scenario.

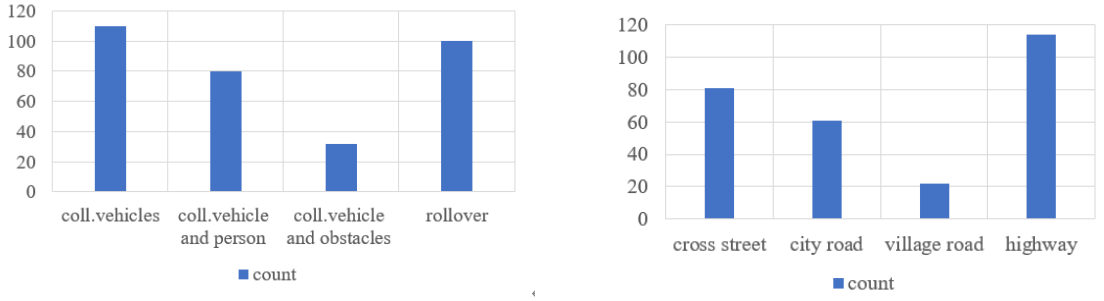
A. IMAGE CLASSIFICATION

TAD is collected from two sources varied in resolution and the number of accident types, we divide image collection extracted from TAD into two subsets referring to ''TrafficAccident-net'' and ''TrafficAccident-platform'', which means they are selected from the internet website and our traffic analysis platform respectively. Similarly, the dataset ''TrafficAccident-all'' indicates the merged version from the above two sources.

To better verify TAD's performance in the test set, we design all experiments based on comparison within three datasets, including ''TrafficAccident-net'', ''TrafficAccident-all'' and RoadAccident [4]. On the one hand, comparison between ''TrafficAccident-net'' and ''TrafficAccident-all'' aims to clearly present the improvement of accuracy due to data collected from the traffic analysis platform. The two training sets each separate 10% of data as validation set and the rest 90% is used for training. On the other hand, there are lots of low-resolution shots of traffic accidents in RoadAccident and we chose this accessible accident dataset to test the effect of image quality on the classification algorithm. All three datasets share the same test data, in total 1548 images were randomly selected 32 scenes downloaded from the traffic analysis platform. The ratio of positive and negative data is kept 1:1 in the test.

Our image classification experiments are based on ResNet50 [57] as its backbone in Convolutional Architecture for Fast Feature Embedding (Caffe) deep learning frameworks running on one GPU of NVIDIA GeForce RTX2080Ti. We have chosen Adam as the optimizer, with a base learning rate set to le-3, a batch size of 64, and trained the model for 200 epochs. We divide whole images into two categories: ''Accident'' and ''Normal'', which refer to scenarios with and without accidents respectively according to mentioned annotated ways in Section III-A3. In Figure 7, the top three lines represent sampling sequences of ongoing accident videos, while the fourth line is those with a single shot. The

FIGURE 4. A snapshot of different scenes of traffic accidents in TAD.



(a) Different accident kinds of TAD.

(b) Different accident scenes of TAD.

FIGURE 5. Shots distribution of traffic accidents in TAD.

select area in the red box on the left refers to positive samples with distinctive features of accidents, while the right shows negative ones without accidents.

The comparison results of the image classification task within three datasets evaluated by Accuracy, Precision, Recall, and F1-score in Table 4. Accuracy represents the proportion of correctly predicted instances out of the total instances. It is calculated as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP, TN, FP, and FN stand for True Positives, True Negatives, False Positives, and False Negatives, respectively. Precision quantifies the fraction of correctly predicted positive instances among all instances predicted as positive. Its formula is:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

A high Precision indicates a model's tendency to label only those instances with high confidence as positive. Recall assesses the model's ability to find all actual positive

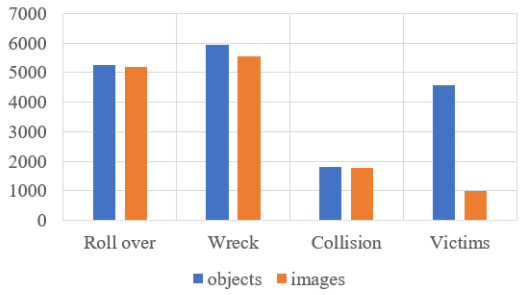


FIGURE 6. Histogram of image and object counts per accident label in TAD dataset.

TABLE 4. Comparison results of image classification task in different datasets.

Dataset	Recall	Precision	F1-score	Accuracy
TrafficAccident-net	0.74	0.59	0.66	0.63
TrafficAccident-all	0.62	0.61	0.61	0.62
RoadAccident	0.09	0.74	0.16	0.54

TABLE 5. Comparison results of TAD in different video classification algorithms.

Model	Recall	Precision	F1-score	Accuracy
SlowFast	0.94	0.65	0.77	0.72
VideoSwinTransformer	0.88	0.74	0.80	0.78

instances. It is calculated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

A high Recall signifies the model's effectiveness in retrieving most of the positive instances. F1-score serves as a harmonic mean of Precision and Recall, providing a balanced view of the model's performance. It is computed as:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

A high F1-score indicates a good trade-off between Precision and Recall.

RoadAccident tops in Precision because of its uniformity in accident types which helps to achieve model convergence while declines in Recall. This indicates its incapability in detecting the majority of accidents. With richer accident information provided, TrafficAccident-net and TrafficAccident-all, though receive lower Precision, are superior in F1-score and Accuracy. We found that our datasets are conducive to detecting accidents compared to RoadAccident.

B. VIDEO CLASSIFICATION

Video classification is almost the same as the task of image classification except for the data format. Similarly, the video classification task is designed to differentiate videos into two types referring to "Accident" and "Normal" based on mainstream video classification models. In this section, 261 videos are used as positive samples and 111 videos without accidents are taken as negative ones. We construct the training set and validation set with the ratio of 9:1, while the test set remains the same video source utilized in IV-A. To create a video frame dataset, we systematically extract frames from each video segment, forming processing units of 64 frames by sampling every 4th frame. As shown in Figure 8, positive samples with consecutive sequences used in video classification are framed with red lines one (Figure 8(a)) with negative ones framed in orange (Figure 8(b)). To facilitate reproducibility and encourage further research, the constructed video frame dataset, alongside the preprocessed videos, is publicly accessible at https://github.com/yajunbaby/TAD-benchmark.

Selected models are SlowFast [58] and VideoSwinTransformer [59], which are both classical algorithms utilized in video classification. SlowFast is a typical CNN-based model in the field of video classification. Due to Transformer's powerful modeling capability, the mainstream backbone of visual tasks has gradually changed from CNN to Transformer. VideoSwinTransformer is a transformer-based model of video recognition, and its efficiency exceeds that of previous decomposed spatio-temporal modeling models. Both networks use the same configuration of parameters with clip len setting 64 and frame interval setting 4 trained in the Pytorch [60] deep learning framework. The original size of images is randomly selected from 128 to 160. Later we randomly crop the images into (112,112) size square covering the target object. Finally, the input to the models are images sized in (64,112,112).

Models are all trained from scratch for 200 epochs with an initial learning rate set at le-3. Evaluation results of TAD trained on SlowFast and VideoSwinTransformer are compared in Table 5. Here we employ the same evaluation metrics with image classification. As described in Table, The performance of VideoSwinTransformer is better than that of SlowFast in all evaluation metrics.

C. OBJECT DETECTION

Experiments in this section are designed based on images to verify the capability of spotting and detecting accidents of mainstream object detection models. We conducted performance validation on several representative object detection algorithms, encompassing the anchor-based classical algorithms Faster R-CNN (a two-stage detector) and YOLOv5 (a one-stage detector), the latest anchor-free YOLO series algorithms, YOLOv8 and YOLOv9, and the Transformer-based representative detection algorithm, RT-DETR [61]. Notably, for YOLO series models, we uniformly employed their medium-scale parameter versions in our experiments. We still use two versions of TAD, both "TrafficAccident-net" and "TrafficAccident-all" to train the model. The accident objects for the object detection task are annotated into four labels, referring "roll over", "wreck",

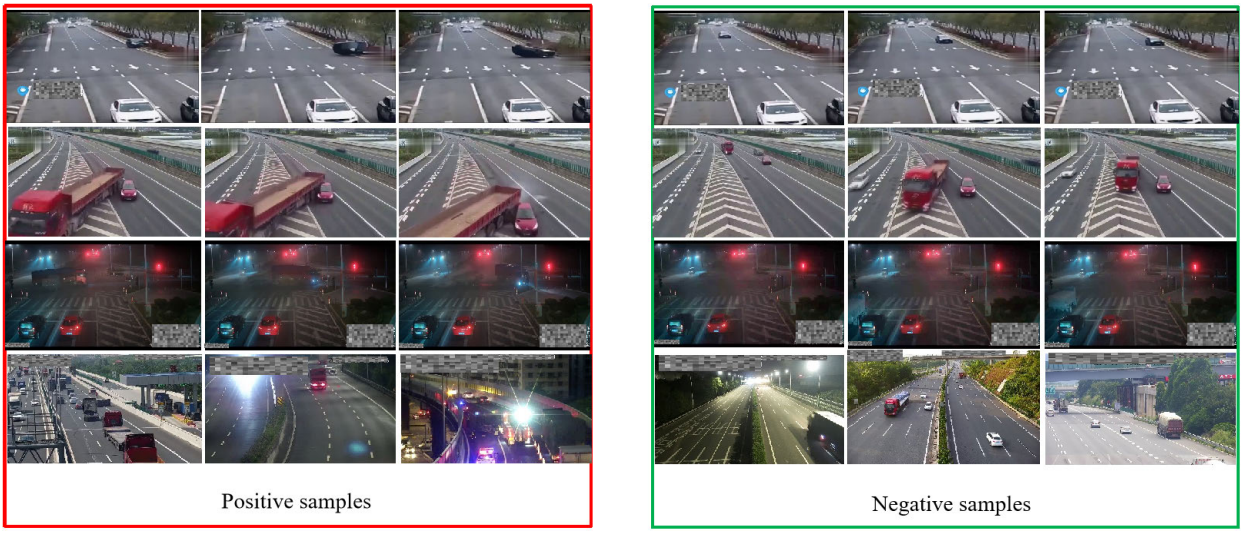


FIGURE 7. A snapshot of positive and negative samples for image classification.

TABLE 6. The detection performance results on test sets with different datasets and backbones. The average precision with a conf threshold of 0.4 is used.

Dataset	Backbone	mAP@50	mAP@50:95	FPS
TrafficAccident-all	FasterRCNN	0.255	0.157	18.5
TrafficAccident-all	YOLOv5m	0.286	0.191	121.9
TrafficAccident-all	YOLOv8m	**0.385**	**0.252**	147.5
TrafficAccident-all	YOLOv9m	0.361	0.238	153.8
TrafficAccident-all	RT-DETR	0.332	0.214	102.5
TrafficAccident-net	FasterRCNN	0.232	0.144	18.5
TrafficAccident-net	YOLOv5m	0.269	0.182	121.9
TrafficAccident-net	YOLOv8m	**0.363**	**0.249**	147.5
TrafficAccident-net	YOLOv9m	0.355	0.213	153.8
TrafficAccident-net	RT-DETR	0.303	0.202	102.5

TABLE 7. Image-level classification performance of TAD among different vision tasks.

Task	Model	Recall	Precision	F1-score	Accuracy
Image classification	ResNet50	0.74	0.59	0.66	0.63
Image detection	YOLOv8m	0.52	0.98	0.68	0.76

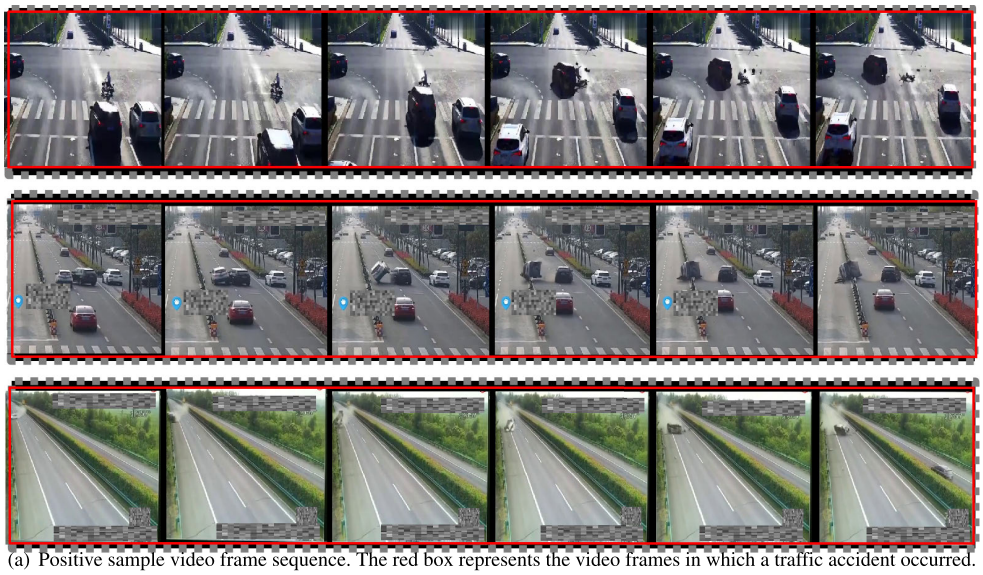
"collision" and "victims" by the Labelme [62] tool. Figure 9 shows some annotation samples applied in our dataset. Each row consists of 6 shots with the same object type from different scenes. It's worth noting that labeled objects are bounded slightly bigger than their original size. Specifically, each bounding box is expanded 1/3 times from the minimum external rectangle box of the target area. This intends to better capture background features apart from the object in the area where accidents happen, improving the detection performance of recognizing accidents.

Experiments are conducted on two GPUs of NVIDIA GeForce RTX2080Ti in the Ubuntu system. For the training settings, We adopt mosaic image preprocessing technology, which joins four images performed random cropping, scaling, and rotation operations together to generate the final synthesized image, and multi-scale strategy ranging from 0.6 to 1.5. The input resolution is 640×640, the mini-batch size is 4 on 2 GPUs and 150 epochs have been trained. To evaluate the real-time performance, we measure the Frames Per Second (FPS) of each model during inference on the test set. FPS is a critical metric for practical deployment,higher FPS indicates faster processing speeds. We follow the general evaluation metrics for object detection with MAP (Mean Average Precision). MAP is a metric that quantifies the average performance of a model across multiple classes in detection tasks. It is derived by computing the arithmetic mean of the AP (Average Precision) values for each class, where AP is calculated as the area under the Precision-Recall curve, reflecting the model's capability in detection within a single class. The MAP is calculated as:

$$\text{MAP} = \frac{1}{K}\sum_{i=1}^{K}\text{AP}_i$$

where K is the number of classes and AP_i is the AP value for the i-th class. A higher MAP value signifies better overall performance of the model across all classes. MAP50 is a specific version of MAP that only considers detections with an IoU threshold of 0.5. When calculating the AP for each category, only detections with an IoU greater than or equal to 0.5 with the ground truth boxes are taken into account. MAP50 reflects the model's detection ability under a relaxed matching criterion. MAP50:95 is a stricter performance metric that calculates MAP at all IoU thresholds from 0.5 to 0.95 (typically with a step size of 0.05) and averages these MAP values. This approach comprehensively evaluates the model's detection performance across different IoU thresholds, providing a more robust indication of the model's ability under precise matching criteria.

Table 6 shows the detection performance of different backbones on our test set. It can be seen that TrafficAccident-all outperforms TrafficAccident-net in all evaluation metrics with the same backbone, which shows the importance of accident data collected from our video analysis platform



(a) Positive sample video frame sequence. The red box represents the video frames in which a traffic accident occurred.

(b) Negative sample video frame sequence. The orange box represents the video frames of normal traffic operation.

FIGURE 8. **Positive and negative samples for video classification.**

TABLE 8. **Video-level classification performance of TAD among different vision tasks.**

Task	Model	Recall	Precision	F1-score	Accuracy
Image classification	ResNet50	0.81	0.62	0.70	0.66
Image detection	YOLOv8m	0.57	1.00	0.73	0.78
Video classification	SwinVideoTransformer	0.88	0.74	0.80	0.78

for model learning. Furthermore, among all backbones, while YOLOv9m exhibits slightly better FPS performance, YOLOv8m maintains the best overall detection performance on our dataset.

In order to compare the performance of accident recognition at the image level, we use the same evaluation metrics including Recall, Precision, F1-score and Accuracy in two vision tasks. For the object detection task, if an object with a conf threshold greater than the specified threshold is detected on the image, we predict that this image belongs to the positive category. In our experiments, the best detect conf threshold is 0.4. As shown in Table 7, the model of the object detection task is superior to the classification task in Accuracy because it can learn more useful information such as the location of accident areas, and the type of accident targets.

Figure 10 visualizes test results in our object detection experiments. Areas colored in red refer to the detected accident targets in the first three rows, while the yellow-colored ones, located at the bottom, indicate wrong predictions. Specifically, it's worth noting that the top two rows are darker

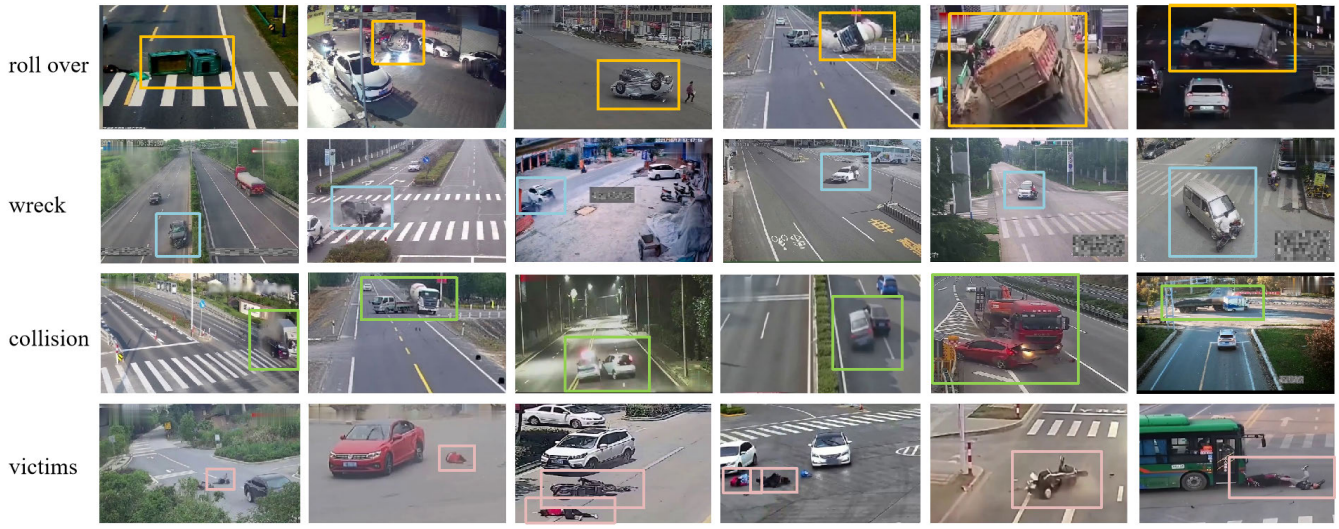


FIGURE 9. Annotation samples for object detection.

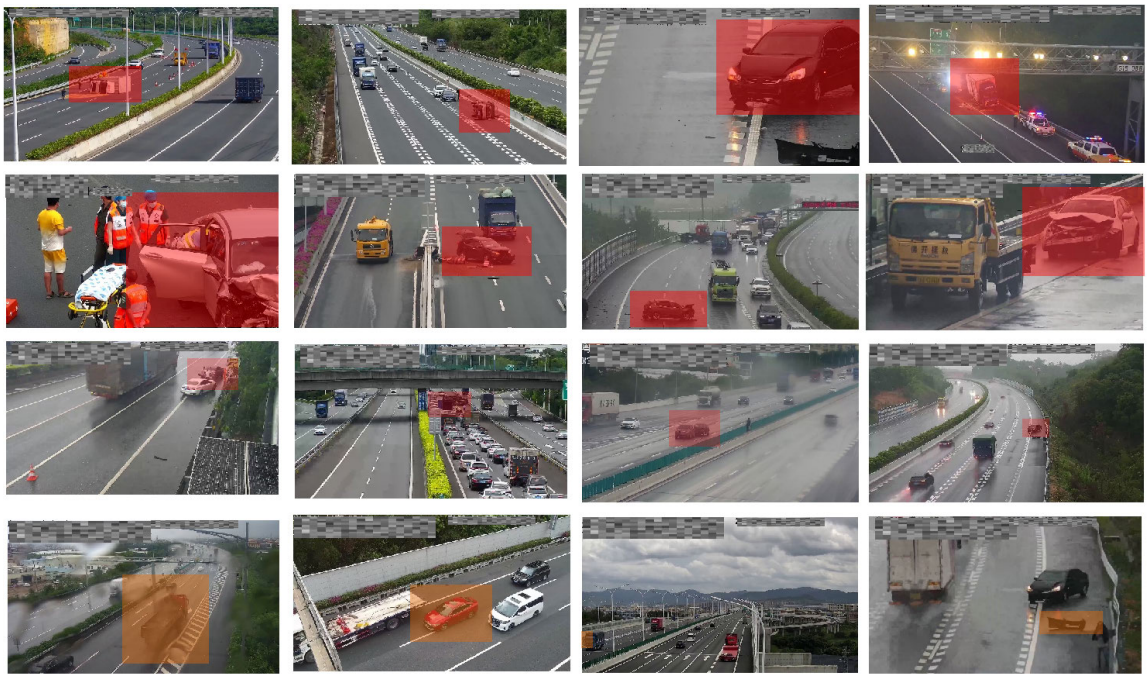


FIGURE 10. Visualization results for the test data.

in red for higher confidence in detecting ''wreck'' and ''roll over'' compared to the third row. Lower confidence is mainly due to the smaller size of distant targets, making it harder to capture visual features, which could be improved by model iterations through the addition of more negative samples.

D. EXPERIMENTS SUMMARY

The above three experiments focus on traffic accident prediction methods based on various vision tasks.

The classification and object detection tasks we explore are image level, learning and utilizing two-dimensional visual features of images. Results show that classification models mainly learn and capture the global features of images, focusing more on the overall regions of images. In contrast, guided by the supervised learning logic, detection models are more refined as they concentrate on local regions while mining information on accident types and locations. As shown in Table 7, the Precision and Accuracy of detection task are higher than those of classification task, indicating that fine-grained regions of accidents are more beneficial to comprehensive detection of accident types. In terms of Recall and F1-score, the detection task is currently lower than classification, possibly due to the performance difference among different categories. In particular, ''collision'' and ''victims'', pull down the overall recall rate.

In concrete practice, the best way to discern an accident is to analyze it by the whole process of its occurrence. Therefore, we use a sequence of video frames to determine the likelihood of an accident at video level. Table 8 shows the prediction results of three vision tasks on the test set at video level. After adjusting various threshold parameters, for the image classification task, we take the threshold value 0.6 as the prediction boundary of the positive sample video, that is to say, when the proportion of video frames predicted as positive samples accounts for 60% of the whole video sequence, we think that an accident has occurred in the video. For the image object detection task, the threshold value of 0.5 can get the best result. The video-level test results confirm that the accuracy of the image object detection task is still higher than that of the image classification task. In general, The video classification task achieved the highest Accuracy and F1 score. From the perspective of learning dimension, video-level models achieve higher accuracy than image-level. This is because the former also utilizes sequential information of the time dimension in addition to the two-dimensional features of images.

V. DISCUSSION

In this paper, we have presented the Traffic Accident Detection (TAD) dataset, a novel and comprehensive resource that encapsulates a diverse range of accident scenes sourced from CCTV footage. The TAD dataset has been meticulously curated to drive the advancement of visual recognition technologies in identifying various types of accidents occurring on roadways. Our dataset stands out for its inclusiveness of a wide spectrum of accident types, spanning from vehicular rollovers and collisions to incidents involving bicycles, pedestrians, and static obstacles. Statistical analyses and algorithmic evaluations conducted on the dataset confirm its superiority in terms of informativeness, highlighting its potential to catalyze significant progress in the field. Two pivotal applications emerge from the TAD dataset:

A benchmark Dataset: TAD aims to serve as a benchmark for extensive research on visual applications. We believe that accident datasets with high quality, specificity and large scale will enable the advancement in object detection and visual classification tasks.

Accident Vision Research: TAD offers annotation for multiple types of accidents from CCTV footage. This caters to the desperate need for open-sourced datasets on traffic research.

VI. FUTURE WORK

Here we list no-exhaustive aspects to be improved for further application and advancement. We intend to include more objects in our dataset. More specifically, we propose to refine the annotation of vehicles in categories such as cars and trucks. This would add more visual information provided by a certain accident video for future research. More scenes of accidents with further variations in weather and roads are expected to be included in an extended version of TAD. We plan to enrich the categories of accidents. For example, we are planning to divide the label ''collision'' into several subcategories, such as head-on collision, rear-end collision, and side-impact collision.

REFERENCES

[1] J. Fang, D. Yan, J. Qiao, J. Xue, H. Wang, and S. Li, ''DADA-2000: Can driving accident be predicted by driver attention? Analyzed by a benchmark,'' 2019, *arXiv:1904.12634*.

[2] 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, Jan. 2023.

[3] S. Bhattacharyya, A. Seal, and A. Mukherjee, ''Real-time traffic incidence dataset,'' in *Proc. SoutheastCon*, Apr. 2019, pp. 1–5.

[4] S. Moosavi, M. H. Samavatian, S. Parthasarathy, and R. Ramnath, ''A countrywide traffic accident dataset,'' 2019, *arXiv:1906.05409*.

[5] T. You and B. Han, ''Traffic accident benchmark for causality recognition,'' in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2020, pp. 540–556.

[6] Y. LeCun, Y. Bengio, and G. Hinton, ''Deep learning,'' *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] M. Naphade, Z. Tang, M.-C. Chang, D. C. Anastasiu, A. Sharma, R. Chellappa, S. Wang, P. Chakraborty, T. Huang, J. Hwang, and S. Lyu, ''The 2019 AI city challenge,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jan. 2019, pp. 452–460.

[8] A. P. Shah, J.-B. Lamare, T. Nguyen-Anh, and A. Hauptmann, ''CADP: A novel dataset for CCTV traffic camera based accident analysis,'' in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–9.

[9] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, ''Anticipating accidents in dashcam videos,'' in *Proc. 13th Asian Conf. Comput. Vis. (ACCV)*, Taipei, Taiwan. Cham, Switzerland: Springer, Jan. 2017, pp. 136–153.

[10] H. Singh, E. M. Hand, and K. Alexis, ''Anomalous motion detection on highway using deep learning,'' in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 1901–1905.

[11] M. Haklay and P. Weber, ''OpenStreetMap: User-generated street maps,'' *IEEE Pervasive Comput.*, vol. 7, no. 4, pp. 12–18, Oct. 2008.

[12] V. Rovšek, M. Batista, and B. Bogunović, ''Identifying the key risk factors of traffic accident injury severity on Slovenian roads using a non-parametric classification tree,'' *Transport*, vol. 32, no. 3, pp. 272–281, Jun. 2014.

[13] K. Chowdhary and K. Chowdhary, ''Natural language processing,'' in *Fundamentals of Artificial Intelligence*, 2020, pp. 603–649.

[14] A. Aboah, M. Shoman, V. Mandal, S. Davami, Y. Adu-Gyamfi, and A. Sharma, ''A vision-based system for traffic anomaly detection using deep learning and decision trees,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4202–4207.

[15] G. Jocher. (May 2020). *YOLOv5 by Ultralytics*. [Online]. Available: https://github.com/ultralytics/yolov5

[16] V. Ravindran, L. Viswanathan, and S. Rangaswamy, ''A novel approach to automatic road-accident detection using machine vision techniques,'' *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 11, 2016.

[17] J. Zhao, Z. Yi, S. Pan, Y. Zhao, Z. Zhao, F. Su, and B. Zhuang, ''Unsupervised traffic anomaly detection using trajectories,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Long Beach, CA, USA, Jun. 2019, pp. 133–140. [Online]. Available: http://openaccess.thecvf.com/content_CVPRW_2019/html/AI_City/Zhao_Unsupervised_Traffic_Anomaly_Detection_Using_Trajectories_CVPRW_2019_paper.html

[18] E. P. Ijjina, D. Chand, S. Gupta, and K. Goutham, ''Computer vision-based accident detection in traffic surveillance,'' in *Proc. 10th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2019, pp. 1–6.

[19] S. Habib, A. Hussain, W. Albattah, M. Islam, S. Khan, R. U. Khan, and K. Khan, ''Abnormal activity recognition from surveillance videos using convolutional neural network,'' *Sensors*, vol. 21, no. 24, p. 8291, Dec. 2021.

[20] X. Jianfeng, G. Hongyu, T. Jian, L. Liu, and L. Haizhu, ''A classification and recognition model for the severity of road traffic accident,'' *Adv. Mech. Eng.*, vol. 11, no. 5, May 2019, Art. no. 1687814019851893.

[21] W. Sultani, C. Chen, and M. Shah, ''Real-world anomaly detection in surveillance videos,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6479–6488.

[22] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M.-C. Chang, Y. Yao, L. Zheng, M. S. Rahman, M. S. Arya, A. Sharma, Q. Feng, V. Ablavsky, S. Sclaroff, P. Chakraborty, S. Prajapati, A. Li, S. Li, K. Kunadharaju, S. Jiang, and R. Chellappa, "The 7th AI city challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023.
[23] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
[24] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.
[25] M. W. Gardner and S. R. Dorling, "Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences," *Atmos. Environ.*, vol. 32, nos. 14–15, pp. 2627–2636, Aug. 1998.
[26] G. I. Webb, *Naïve Bayes*. Boston, MA, USA: Springer, 2010, pp. 713–714, doi: 10.1007/978-0-387-30164-8_576.
[27] D. Tian, C. Zhang, X. Duan, and X. Wang, "An automatic car accident detection method based on cooperative vehicle infrastructure systems," *IEEE Access*, vol. 7, pp. 127453–127463, 2019.
[28] K. Liu and H. Ma, "Exploring background-bias for anomaly detection in surveillance videos," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019.
[29] D. Koshti, S. Kamoji, N. Kalnad, S. Sreekumar, and S. Bhujbal, "Video anomaly detection using inflated 3D convolution network," in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, Feb. 2020, pp. 729–733.
[30] H. Kim, S. Park, and J. Paik, "Pre-activated 3D CNN and feature pyramid network for traffic accident detection," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2020, pp. 1–3.
[31] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.
[32] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.
[33] S. Robles-Serrano, G. Sanchez-Torres, and J. Branch-Bedoya, "Automatic detection of traffic accidents from video using deep learning techniques," *Computers*, vol. 10, no. 11, p. 148, Nov. 2021.
[34] Z. Xu, J. Hu, and W. Deng, "Recurrent convolutional neural network for video classification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.
[35] G. Hinton, *Recursive Distributed Representations*. Cambridge, MA, USA: MIT Press, 1991.
[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[37] C. Li, S. H. Chan, and Y.-T. Chen, "Who make drivers stop? Towards driver-centric risk assessment: Risk object identification via causal inference," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 10711–10718.
[38] B. Mahaur, N. Singh, and K. K. Mishra, "Road object detection: A comparative study of deep learning-based algorithms," *Multimedia Tools Appl.*, vol. 81, no. 10, pp. 14247–14282, Apr. 2022.
[39] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.
[40] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[41] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[42] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, Dec. 2015, pp. 91–99.
[43] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[44] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[45] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
[46] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[47] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham, Switzerland: Springer, 2016, pp. 21–37.
[48] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[49] T.-N. Le, S. Ono, A. Sugimoto, and H. Kawasaki, "Attention R-CNN for accident detection," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 313–320.
[50] P. Wang, C. Ni, and K. Li, "Vision-based highway traffic accident detection," in *Proc. Int. Conf. Artif. Intell., Inf. Process. Cloud Comput.*, Dec. 2019, pp. 1–5.
[51] G. Jocher. (May 2018). *YOLOv3 by Ultralytics*. [Online]. Available: https://github.com/ultralytics/yolov3
[52] G. Jocher, A. Chaurasia, and J. Qiu. (Jan. 2023). *Ultralytics YOLO*. [Online]. Available: https://github.com/ultralytics/ultralytics
[53] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," 2024, *arXiv:2402.13616*.
[54] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2020, pp. 213–229.
[55] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, 2006, pp. 850–855.
[56] G. Cao. *Weibo Sina*. [Online]. Available: https://overseas.weibo.com/
[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[58] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6201–6210.
[59] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video Swin transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3192–3201.
[60] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2019.
[61] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16965–16974.
[62] K. Wada. *Labelme: Image Polygonal Annotation With Python*. [Online]. Available: https://github.com/wkentaro/labelme

YAJUN XU received the M.S. degree from the Institute of Information Engineering, School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China.

She is currently working as a Software Development Engineer with AI Innovation Center, China Unicom. Her research interests include computer vision and artificial intelligence.

HUAN HU received the B.S. degree from the School of Science, Hubei University of Technology, Hubei, China, in 2013, and the M.S. degree from the School of Electronic Science and Engineering, University of Electronic Science and Technology of China, Sichuan, China, in 2017.

In 2017, he joined CloudMinds Technologies Inc., as an Algorithm Engineer. Since 2020, he has been working as a Deep Learning Algorithm Engineer with AI Innovation and Center, China Unicom Digital Technology Company Ltd., Beijing, China. His research interests include machine vision, computer vision, and deep learning.

CHUWEN HUANG received the M.S. degree in statistics from the Renmin University of China.

She has been working as a Software Engineer with AI Innovation Center, China Unicom engaged in algorithm development and application. Her research interests include but are not limited to deep learning, big data, computer vision, and natural language processing.

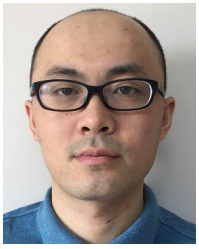
YIBING NAN received the Ph.D. degree from Beijing Institute of Technology, China.

He was a Senior Engineer at CloudMinds Technologies Inc., from 2016 to 2019. He has been working as a Senior Algorithm Expert with AI Innovation Center, China Unicom, since 2019. He is the author of some 20 refereed international papers and held more than 50 patents. His research interests include artificial intelligence, deep learning, computer vision, and intelligent video analysis.

YUYAO LIU received the B.Sc. degree in automation from Tsinghua University, Beijing, China, in 2024. He is currently pursuing the Ph.D. degree in mathematics with Hong Kong University of Science and Technology, Hong Kong. His research interests include anomaly detection in computer vision, pattern recognition, machine learning, and bioinformatics.

KAI WANG received the Ph.D. degree from Nanyang Technological University, Singapore, in 2013.

He has been working as the AI Director with AI Innovation Center, China Unicom, since 2019. Before that, he worked at CloudMinds Technologies Inc., and the Central Research Institute of Huawei Technologies. He has published more than 20 refereed papers on international journals and conferences and granted more than 40 patents. His research interests include generative AI, computer vision, computer graphics, and human–computer interaction.



ZHAOXIANG LIU received the B.S. and Ph.D. degrees from the College of Information and Electrical Engineering, China Agricultural University, Beijing, China, in 2006 and 2011, respectively.

He joined VIA Technologies, Inc., Beijing, in 2011. From 2012 to 2016, he was a Senior Researcher at the Central Research Institute, Huawei Technologies, Beijing. He was a Senior Manager at CloudMinds Technologies Inc., Beijing, from 2016 to 2019. Since 2019, he has been working as the Director of AI research with AI Innovation Center, China Unicom. He has published more than 20 refereed papers on international journals and conferences and hold more than 40 patents. His current research interests include artificial intelligence, computer vision, deep learning, robotics, and human–computer interaction.

SHIGUO LIAN (Member, IEEE) received the Ph.D. degree from Nanjing University of Science and Technology, China.

He was a Research Assistant at the City University of Hong Kong, in 2004. From 2005 to 2010, he was a Research Scientist at France Telecom Research and Development, Beijing. He was a Senior Research Scientist and a Technical Director of Huawei Central Research Institute, from 2010 to 2016. He was a Senior Director of CloudMinds Technologies Inc., from 2016 to 2019. Since 2019, he has been working as the Chief AI Scientist with China Unicom. He has authored more than 100 refereed international journal articles covering topics of artificial intelligence, robotics, human–computer interface, and multimedia communication. He has authored or co-edited more than ten books and hold more than 200 patents.

Dr. Lian is an editor board of several refereed international journals.