Traffic Accident Detection via Self-Supervised Consistency Learning in Driving Scenarios

Jianwu Fang, *Member, IEEE*, Jiahuan Qiao, Jie Bai, Hongkai Yu, *Member, IEEE*, and Jianru Xue, *Member, IEEE*

***Abstract*— With the rapid progress of autonomous driving and advanced driver assistance systems, there are growing efforts to promote their safety in natural driving scenarios, especially for the detection of the traffic accidents. However, because of the dynamic camera motion and complex scene in driving situations, traffic accident detection is still challenging. In this work, we aim to give the ability of Traffic Accident Detection for driving systems by proposing a Self-Supervised Consistency learning framework, termed as SSC-TAD, that involves the appearance, motion, and context consistency learning. The key formulation is to find the inconsistency of video frames, object locations and the spatial relation structure of scene temporally between different frames captured by the dashcam videos. Within this field, different from the previous works which concentrate on predicting the future object locations or frames, we further focus on predicting the visual scene context in driving scenarios and detecting the traffic accident by considering the temporal frame consistency, temporal object location consistency, and the spatial-temporal relation consistency of road participants. In this work, this formulation is fulfilled by a collaborative multi-task consistency learning network and the visual scene context feature is represented by a graph convolution network. The superiority to the state-of-the-art is verified by exhaustive evaluations on two large scale datasets, i.e., the AnAn Accident Detection (A3D) dataset and DADA-2000 dataset collected recently.**

***Index Terms*— Traffic accident detection, frame and location prediction, scene context, adversarial learning.**

I. Introduction

IN RECENT years, there are more and more research efforts to develop the advanced driver assistance, self-driving systems, and the internet of vehicles [1], with the beautiful vision of more comfortable and safer driving experience. That is because the shocking fact that over 1.35 millions of people died in each year [2] investigated by World Heath Organization (WHO). To fulfill this goal, except developing driving techniques in normal driving situations, one key and fundamental aspect of driving systems is to timely react the abnormal situations, especially for the Traffic Accident Detection (TAD) to be explored in this work. Actually, although the traffic accident takes a long-tailed distribution of occurrence for one vehicle, it has severe impact on driving safety when it occurs.

Under this demand, many vision perception systems in automatic driving vehicles utilized the deep learning technique to detect, recognize, or predict the actions of surrounding participants captured by the dashcam videos recently [3]–[7]. Because of the rare frequency of traffic accidents, it is impossible to collect enough and diverse training data online in practical driving. Consequently, there is no sufficient data label to cover all kinds of traffic accidents in the traffic world and achieve a fully-supervised learning for traffic accident detection. Therefore, previous works commonly take an assumption that the traffic accident occurs rarely and model the detection of accident as a fitting problem by the normal situations with the captured videos [3], [8]–[10], and formulate an unsupervised learning prototype. It is worthy noting that the topic in this work is different from the accident prediction wherein the approaches [11], [12] often trim the videos into short clips (about 100 frames) with several accident frames (about ten frames) in the end of the clips, and model a supervised accident prediction.

For TAD, the consistency of spatial-temporal features over multiple video frames is commonly focused, and mainly exploited by video frame prediction [4] or trajectory prediction [3], [13]. For the video frame prediction, it needs to predict the whole video frames and determine the occurrence of accident by the pixel-level difference between the predicted frames and the true ones. Frame prediction framework has the global scene appearance, which is useful for determining the traffic accident without surrounding objects (e.g., the category of "*ego-car hitting road boundary*", as shown in the first row of Fig. 1). However, frame prediction framework is easily to be disturbed by the rapid camera motion and large illumination change in normal driving situations. In terms of the trajectory prediction approaches, they need to do object detection, object tracking and trajectory prediction. Then the accident is determined by the predictive uncertainty (e.g., variance) in the predicted trajectories, in which one pioneering unsupervised

Manuscript received 29 July 2021; revised 16 January 2022 and 1 March 2022; accepted 3 March 2022. Date of publication 14 March 2022; date of current version 8 July 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62036008; in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2022JM-309; and in part by the Fundamental Research Funds for the Central Universities, CHD, under Grant 300102320202. The Associate Editor for this article was S. Wan. *(Corresponding author: Jianwu Fang.)*

Jianwu Fang is with the College of Transportation Engineering, Chang'an University, Xi'an 710064, China, and also with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: fangjianwu@chd.edu.cn).

Jiahuan Qiao and Jie Bai are with the School of Electronic and Control Engineering, Chang'an University, Xi'an 710064, China.

Hongkai Yu is with the Department of Electrical Engineering and Computer Science, Cleveland State University, Cleveland, OH 44115 USA (e-mail: h.yu19@csuohio.edu).

Jianru Xue is with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: jrxue@mail.xjtu.edu.cn).

Digital Object Identifier 10.1109/TITS.2022.3157254

TAD work was proposed by Yao *et al.* [3] who predicted the future locations of objects and detected the accident by the variance of the predicted locations. Trajectory prediction can resist the rapid motion influence but miss the global scene appearance information. In addition, we find that some kinds of traffic accidents occur very suddenly within less than about 1 second from normal to accident situation (as demonstrated in the second row of Fig. 1), where the scene inconsistency appears in a very short time window, and the persons to be involved in the accident appear suddenly and have no available temporal window to find them and predict their trajectories.

In this work, we absorb the merits of these two kinds of frameworks, and further advocate a TAD framework by involving the visual scene context consistency captured by daschcam video frames. The main formulation is based on a fact that the normal driving situations obey a relatively regular spatial relation structure between road participants, and the objects in accidents commonly involve a sudden or an irregular change of spatial relation structure. We term this relation feature as "*visual scene context feature*". Therefore, this work constructs a Self-Supervised appearance, motion, and context Consistency learning for Traffic Accident Detection in driving scenarios (*Abbrev.* SSC-TAD). In order to fulfill this insight, we build a collaborative multi-task consistency learning network, and the consistency of visual scene context feature is modeled by a graph-embedded generative adversarial network (Sec. III-D). In our work, we do not need any label annotation and video trimming work, which can be directly utilized to any raw dashcam videos. We evaluate the performance on the AnAn Accident Detection (A3D) dataset [3] and the new dataset called as DADA-2000 [14] collected recently. The superiority is obtained by the comparison with the state-of-the-art. In summarize, the **contributions** are as follows:

- This work proposes a traffic accident detection method by simultaneously learning the appearance, motion and context consistency, and fulfilled by a collaborative multi-task consistency learning framework. It does not need any annotation work and achieves a new unsupervised TAD.
- The proposed TAD method can adapt to most of the traffic accidents, e.g., the "*out of control of ego-car*" (frame-level appearance is the dominant clue for this kind of accident) and other object-related accidents, with the consideration of global frame-level appearance, object-level motion, and the association of them.
- We demonstrate superior performance of the proposed method to several state-of-the-art approaches on two large-scale benchmarks, i.e., the AnAn Accident Detection (A3D) dataset [3] and the new dataset DADA-2000 [14] collected by ourselves, with exhaustively experiment analysis on over 100k testing frames.

The reminder of this work is organized as follows. Section. II presents the related work. Section. II-C describes the proposed approach. The experiments and discussion are demonstrated in Section. IV, and Section. V gives the conclusion.

Fig. 1. Some examples of traffic accidents. Top row shows two traffic accidents without surrounding objects ("*ego-car hitting road boundary*"), and the second row demonstrates an accident with a sudden person crossing.

II. Precedent Works

Relating to this work, the fields of anomaly detection in surveillance videos, traffic accident detection and prediction in dashcam videos are focused, where the technical pipelines of them have some same thinking but with different consideration because of differing scenarios.

A. Anomaly Detection in Surveillance Videos

The most related topic for traffic accident detection in dashcam videos is the video anomaly detection, i.e., finding the abnormal event in surveillance videos. As for this field, the formulations of video anomaly detection are commonly defined by profiling the normal behavior and measuring the spatial-temporal feature consistency. In the early research paradigm, the video anomaly is determined by designing various classification models to evaluate the rarity, sparsity, and irregularity of motion, trajectory patterns or other hand-crafted features [15], [16], such as the Histogram of Optical Flow (HOF) [17], Histogram of Gradient (HOG) [18], spatial-temporal cubes [19], and so on. Within one decades, the classifiers for anomaly detection is based on learning formulation, e.g., supervised (e.g., Hidden Markov Model (HMM) [20], Mixture of Dynamic Texture (MDT) [21], Gaussian Process Regression (GPR) [22], multiple instance learner [23]), unsupervised (e.g., Gaussian Mixture Model (GMM) [24], sparsity reconstruction [25], Hierarchical Dirichlet Process (HDP) [26]), and other approaches (one-class SVM [27], isolation forest [28], etc.).

With the progress of large scale benchmark for video anomaly detection, such as ShanghaiTech [29], UCF-Crime [23], the auto-encoders [30], [31], expressive Convolution Neural Networks (CNNs) [32], [33], and predictive Recurrent Neural Networks (RNNs) [7], [34] are used to minimize the reconstruction, expression, and prediction error of the input samples, respectively. Based on the large-scale data learning, the deep learning based methods generate a better normal video profiling and can determine the video anomaly more robust than previous hand-crafted methods [35]. In addition, the latest works exploited the dependency of the behaviors between frames, such as LSTM predictor [34] and sequential generator [4], [36], [37]. For instance, Liu *et al.* [4] leveraged a future frame prediction based framework for anomaly detection. However, most of these video anomaly detection methods need a stable

scene motion for normal situations, which may be disturbed in practical driving scenarios with rapid camera motion. Besides, because of the rarity of anomaly occurrence in local regions and global frames, the attention based loss or mechanism [37], [38] is taken account for the importance of the learned features.

B. Traffic Accident Detection in Dashcam Videos

Compared with the anomaly detection in surveillance videos, TAD in dashcam videos begins to be focused recently. Some previous works also absorbed the insight of video anomaly detection in surveillance systems, and detected the road crashes by motion feature reconstruction [39], [40].

Different from the video surveillance systems, the dashcam videos captured in moving cars are full of rapid motion and fast in and out of road participants. Consequently, the motion feature based methods sometimes are vulnerable to the dynamic camera motion. Therefore, for the traffic accident detection in dashcam videos, most of the works focused on predicting the object trajectories. For example, Yao *et al.* [3] proposed an unsupervised traffic accident detection work on dashcam videos, which determined the accident by predicting the future locations of objects. Then, they extended this work for object-centric and spatial-temporal accident detection [10]. However, they do not consider the spatial-temporal relations between road participants.

C. Traffic Accident Prediction in Dashcam Videos

Relating to TAD, some works concentrate on the traffic accident prediction problem with dashcam videos. For instance, Kataoka *et al.* [5], [6] and Chan *et al.* [7] detected the traffic accident through evaluating the predicted trajectory with labeled ground-truth by adaptive loss and early collision loss, respectively. However, these formulations need to label the ground-truth of object trajectories in advance in dashcam videos. Some excellent object trackers [41]–[43] are useful for these methods. However, long-term object tracking is difficult in complex driving scenarios. Hence, Kilicarsla and Zheng [44] exploited the motion profile change in horizontal and vertical orientation, specially for the zero-flow (optical flow close to zero), and calculated the Time to Collision (TTC) value by the scale change of the target. The work of [44] is interesting and will be considered in future. Besides, Bao *et al.* [12] contributed a traffic accident anticipation approach by utilizing a Graph Convolutional Recurrent Network (GCRN) [11] to model the spatial-temporal relations between different objects, while it needs to trim the occurrence of the accident to the end of the sequence and annotate the tracklets of objects to learn the spatial-temporal relation. In our case, the topic in this work is different from the traffic accident prediction, and we do not need any label annotation and video trimming work.

III. Our Approach

In driving scenarios, traffic accident commonly involves a collision of road participants. Naturally, there are commonly sudden visual scene change, inconsistency movement of road users, and chaotic change of spatial-relation between different participants from normality to accident situation. Therefore, we model the traffic accident detection from the consideration of frame appearance consistency, object motion consistency and scene context consistency simultaneously. In fact, traffic anomaly usually demonstrates a similar scene inconsistency. Hence, the traffic accident detection in this work aims to localize the temporal window of "*anomaly-to-accident*" (A2A), where the starting time of A2A is activated once the object to be involved in accident appears in the scene. With the definition of A2A, this work may be adaptable for early traffic accident detection (to be verified in experiments).

Fig. 2 demonstrates the pipeline of the self-supervised consistency learning framework. Manifestly, the main parts consist of a frame prediction module, object location prediction module and the collaborative multi-task consistency learning module. Within the collaborative multi-task consistency learning module, we have a Driving Scene Context Representation (DSCR) module. The main method pipeline is described as follows.

Specifically, we input a dashcam video clip with successive T frames $\mathcal{I} = \{I_1, \ldots, I_t, \ldots, I_T\}$ with the same image size (256×256). From these video frames, we detect the objects by some popular object detectors and obtain the object bounding boxes $\mathcal{C} = \{C_1^{1:N_1}, \ldots, C_t^{1:N_t}, \ldots, C_T^{1:N_T}\}$, where $N_t (t = 1, \ldots, T)$ represents the number of detected objects in the t^{th} frame. Because the pixel-level motion change among video frames is an important clue to find the object with sudden movement, we also obtain the optical flow images $\mathcal{F} = \{F_1, \ldots, F_t, \ldots, F_{T-1}\}$ of video frames (Notably, two video frames generates one optical flow image). With these input preparations, we predict the $(T+1)^{th}$ frame $\hat{I}_{T+1}$ and the object locations $C_{T+1}^{1:N_{T+1}}$. With $\hat{I}_{T+1}$ and $C_{T+1}^{1:N_{T+1}}$, we represent the driving scene context by DSCR module (to be described in Sec. III-C), which is learned by Graph Convolution Network (GCN).

In the following, frame prediction, object location prediction, and DSCR modules are described in detail, respectively. Then, the collaborative multi-task consistency learning and traffic accident determination are presented.

A. Frame Prediction

The structure of frame prediction module is shown in Fig. 3, which contains a two-branch of frame encoder, optical flow motion image encoder, and one path of future frame decoder. For the feature encoding of appearance and motion within frames, we temporally concatenate T RGB frames and $T-1$ optical flow images into frame tensors in the time dimension, respectively, where the optical flow images are obtained by pre-trained FlowNet 2.0 [45] and pseudo-colorized by [45]. They are fed separately into the two-branch encoders with shared weights in each branch, interleaved with multiple *convolution*, *max-pooling*, and *relu* layers. The feature embedding of T RGB frames and $T-1$ optical flow images are defined as:

$$\begin{aligned} \mathbf{Z}_T &= \phi(\mathcal{I}, \Phi_I), \\ \mathbf{m}_{1:(T-1)} &= \phi(F, \Phi_I), \end{aligned} \tag{1}$$

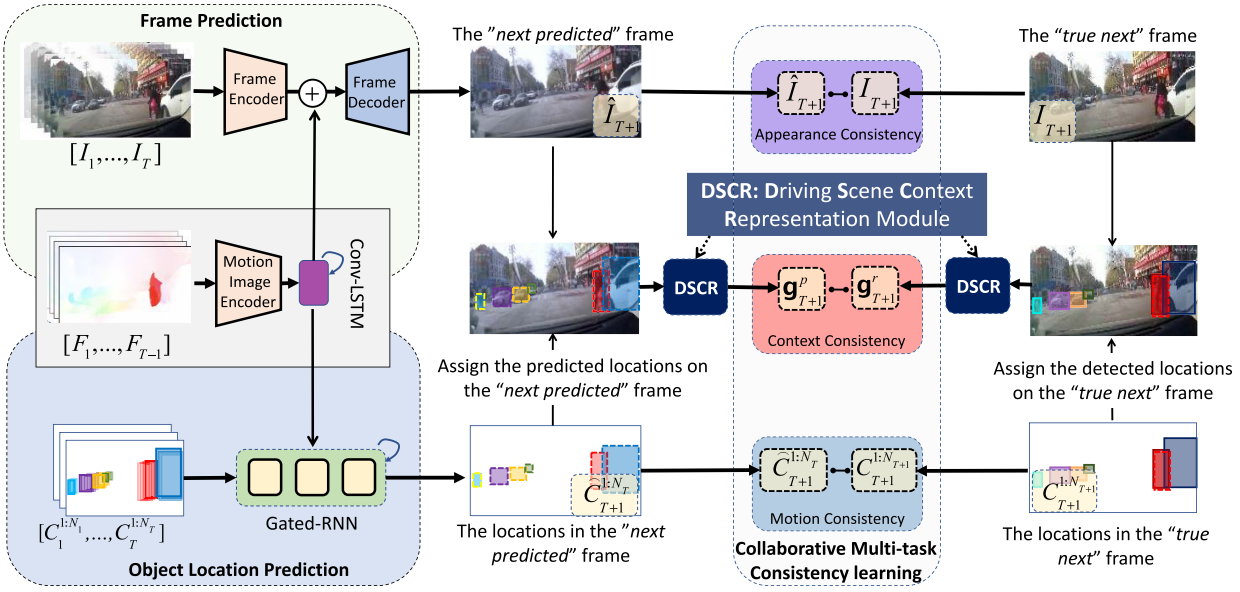


Fig. 2. The framework of appearance, motion, and context consistency learning, where $\mathbf{g}^p_{T+1}$ and $\mathbf{g}^r_{T+1}$ represent the predicted visual scene context feature and the real one of the $(T+1)^{th}$ frame. The shared motion encoder output the embedding of optical flow images and feed the motion embedding to the frame prediction and object location prediction modules.

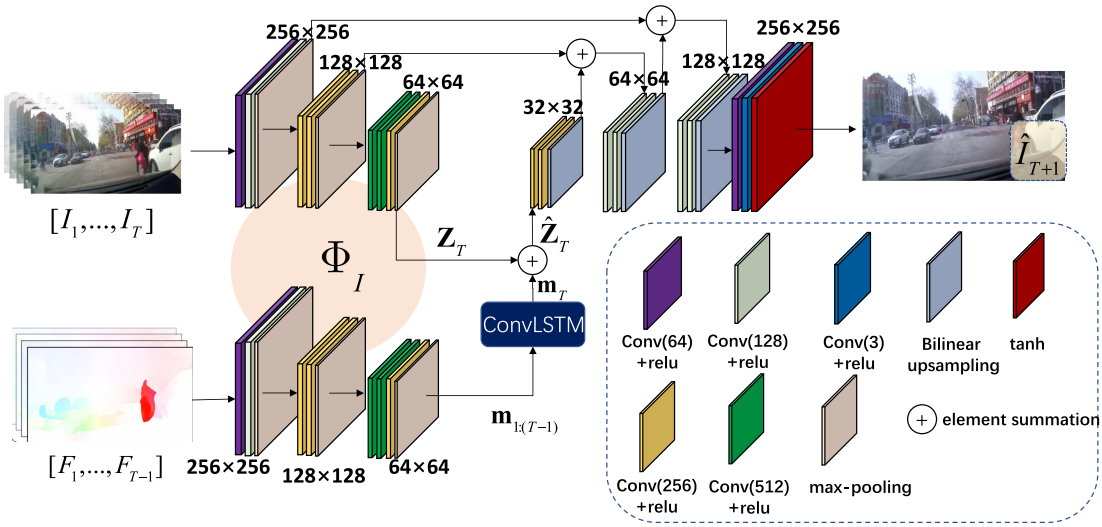


Fig. 3. The architecture of the frame prediction network, where Φ_I is the shared weight of the encoding of RGB frames and optical flow images.

where the shared weight Φ_I is the parameter to be optimized, $\mathbf{Z}_T \in R^{32\times32\times256}$ is the feature embedding tensor of T video frames, and $\phi(\cdot,\cdot)$ specifies the feature mapping function of the encoding module. $\mathbf{Z}_T$ captures the temporal correlation within T RGB frames, and $\mathbf{m}_{1:(T-1)} \in R^{(T-1)\times32\times32\times256}$ contains the dynamic motion feature of the previous $(T-1)$ frames. Notably, frame encoder and motion encoder have the same structure with seven convolution layers, three max-pooling layers, where the kernel size of convolution in our case is 3×3, and the number of kernels is denoted in the bracket of each convolution layer, as shown in Fig. 3.

1) Temporal Motion Prediction For $\hat{I}_{T+1}$: In order to predict $\hat{I}_{T+1}$, one straightforward way is to estimate the motion change between I_T and I_{T+1}. However, I_{T+1} is unknown in encoding process. Therefore, we predict the pseudo motion feature between the $(T-1)^{th}$ frame and the T^{th} frame by passing $\mathbf{m}_{1:(T-1)}$ through a Convolutional-LSTM module (Conv-LSTM) [46], defined as:

$$\mathbf{m}_T = \mathbf{ConvLSTM}(\mathbf{m}_{1:(T-1)}, \Phi_F), \tag{2}$$

where Φ_F refers to the parameters of ConvLSTM, and $\mathbf{m}_T \in R^{32\times32\times256}$ is the predicted pseudo motion feature tensor between the $(T-1)^{th}$ frame and the T^{th} frame. In this work, the number of hidden states of ConvLSTM is set as 256.

2) Appearance Motion Feature Fusion: For the frame prediction, we fuse $\mathbf{m}_T$ and $\mathbf{Z}_T$ by:

$$\hat{\mathbf{Z}}_T = \mathbf{m}_T + \mathbf{Z}_T. \tag{3}$$

Thus, $\hat{\mathbf{Z}}_T$ contains the information of T RGB frames and the predicted motion change between the I_T and $\hat{I}_{T+1}$. With that, $\hat{\mathbf{Z}}_T$ is fed into the future frame decoder to output $\hat{I}_{T+1}$.

3) Future Frame Decoding: In the future frame decoding process, we de-convolute $\hat{\mathbf{Z}}_T$ to $\hat{I}_{T+1}$ with a future frame decoder interleaved with upsampling and convolution layers. In order to maintain the spatial details of future frame in decoding process, we add the skip connection from each max-pooling layer in the frame encoder to the first convolution layer after bilinear upsampling in future frame decoder with the same resolution.

B. Object Location Prediction

Object location prediction network aims to give the positions of the objects in the future frame, which implies the motion consistency feature when measuring the occurrence degree of traffic accident. Different from the work of [3], we need not to know the camera extrinsic parameters to estimate the ego-motion, and the motion feature of the video frames in future frame prediction is directly utilized in the location prediction to fulfill an end-to-end learning. In this work, assume we obtain N_t road participants in the t^{th} video frame by object detectors. Their locations at time t here are denoted as $C_t^{1:N_t} = \{(x_i^c, y_i^c, h_i, w_i)\}_{i=1}^{N_t}$, where x_i^c, y_i^c, h_i, w_i refer to the center point (x_i^c, y_i^c), the height h_i and width w_i of the i^{th} participant in the t^{th} frame. In order to fulfill the future location prediction, the temporal object associations in successive frames are needed, and this work takes a multi-object tracker, e.g., DeepSort [47], to associate the tracklets. Notably, object location prediction not only predict the trajectories of road participants, but also the bounding boxes of them are preferred in following driving scene context representation.

1) Missing Object Consideration: It is worthy noting that although object detection methods have made significant progress, it is inevitable to miss or appear wrongly detected results. In order to fulfill a robust consistency learning of following driving context, we adopt DeepSort to associate the object bounding boxes **backward**, i.e., *from time T to time* 1, where the object feature within the bounding box is extracted by self-configured deep feature model in DeepSort. By this strategy, the number of temporal associations of road participants can be maintained to the same to the ones in the $(T+1)^{th}$ frame. For the object missing problem, this work fixes the length (T) of the temporal associations, which is fulfilled by **backward** copying the first point coordinates of the trajectory to the first time step within T video frames. Backward copying is used to consider the newly appeared object(s). After the object association, the object locations $\{C_t^{1:N_t}\}_{t=1}^{T}$ in T frames is changed into $\{C_t^{1:N_T}\}_{t=1}^{T}$, which generates N_T associated tracklets with the same length of T.

2) Location Prediction: Assume we obtain N_T temporal associations with the indexes (IDs) over T video frames. Each association is a 4-dimensional coordinate chain from time 1 to time T. In each association, the future location prediction can be conducted by some Gated Recurrent Neural Networks (Gated-RNN). In this work, Gate Recurrent Unit (GRU) [48] is adopted. As shown in Fig. 4, both the object locations along

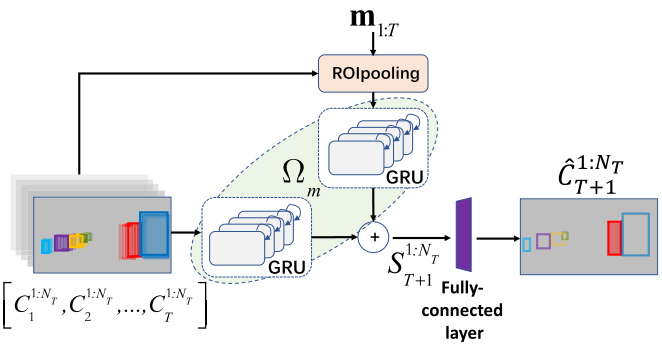


Fig. 4. The architecture of the object location prediction module.

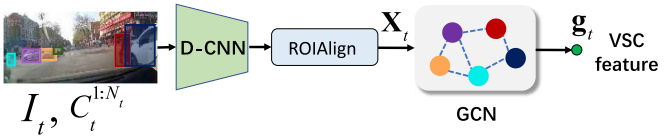


Fig. 5. The driving scene context representation module.

the temporal associations and the motion feature within the object bounding boxes are the input of GRU.

In this work, we directly link the encoded motion features $[\mathbf{m}_1, \ldots, \mathbf{m}_T]$ of T frames from frame prediction network to fulfill an end-to-end learning. The motion feature vector of each object at time t is aligned by a ROI pooling operation with the pooling size of 1×1 on the encoded motion feature tensor $\mathbf{m}_t$ $(t = 1, \ldots, T)$, defined as:

$$\mathbf{MO}_t = [\mathbf{mo}_t^1, \ldots, \mathbf{mo}_t^{N_T}] = \text{ROIpool}(\mathbf{m}_t, C_t^{1:N_T}), \quad (4)$$

To be specific, the location prediction with T frames of observation is operated by:

$$S_{T+1}^{1:N_T} = \mathbf{GRU}(C_{1:T}^{1:N_T}, \Omega_m) + \mathbf{GRU}(\mathbf{MO}_{1:T}, \Omega_m), \quad (5)$$

where $S_{T+1}^{1:N_T}$ refers to the predicted object location state of N_T temporal associations for the $(T+1)^{th}$ frame, and the shared $\mathbf{\Omega}_m$ refers to the parameters of GRU cells for the temporal associations of locations and the motion feature embeddings. The number of hidden states of GRU are set as 256. Here, $S_{T+1}^{1:N_T}$ is decoded by a fully-connected layer to output the future locations of N_T objects $\hat{C}_{T+1}^k = \eta(S_{T+1}^k)$, where $k = 1, \ldots, N_T$, and $\eta(\cdot)$ denotes a fully connected structure.

C. Driving Scene Context Representation (DSCR)

In this work, DSCR models the scene relations within the video frame. In order to resist influence of the wrongly detected road participants, this work extracts the informative relation within the participants, as well as the whole video frame, for the traffic accident detection. For this goal, we introduce the graph representation over the feature embeddings of the specific road participants and the whole frame.

As shown in Fig. 5, assume we have obtained N_t objects (detected by Mask-RCNN [49] in this work) $C_t^{1:N_t}$ for the t^{th} video frame I_t. We first extract the feature embeddings $\mathbf{f} =$

$\{f_1, f_2, \ldots, f_k, \ldots, f_{N_t}, f_v\}$ of these objects and the whole frame, where $f_k \in \mathbb{R}^{1\times d}$ refers to the d-dimensional feature vector of the k^{th} object at time t, and f_v denotes the feature embedding of whole frame. With $\mathbf{f}$, we re-organized it into a matrix $\mathbf{X}_t \in \mathbb{R}^{(N_t+1)\times d}$ and then the similarity of different nodes is defined as:

$$\mathbf{S}_t = \phi(\mathbf{X}_t)^T \phi'(\mathbf{X}_t), \tag{6}$$

where $\phi(\mathbf{X}_t) = \mathbf{w}_T \mathbf{X}_t$ and $\phi'(\mathbf{X}_t) = \mathbf{w}'_T \mathbf{X}_t$ are the linear transformation of $\mathbf{X}_t$ fulfilled by a fully connected layer in our case. $\mathbf{S}_t \in \mathbb{R}^{(N_t+1)\times(N_t+1)}$ is the similarity matrix of nodes. $\mathbf{w}_T, \mathbf{w}'_T \in \mathbb{R}^{d\times d}$ are the weight matrixes of the transformations ϕ and ϕ', respectively.

1) Node Representation: As for the feature embeddings of objects and whole frame, Deformable Convolution Neural Network (D-CNN) is adopted to extract the whole frame to resist the disturbance of the background within the object bounding boxes. After D-CNN, we take the ROIAlign operation [49] to generate the feature vector of each object. D-CNN here is interleaved with four stacked layers of blocks, where the first three blocks all contain a *deformable convolution* layer, a *max-pooling*, a *relu* operation, and the last block only consists of a *deformable convolution* layer and a *relu* operation. The kernel size of *deformation convolution* is 3×3, and the feature channels from shallow convolution layer to deep convolution layer are 64, 128, 256, and 512, respectively.

2) Graph Representation: Based on the node representation of N_t objects and the whole frame, we extract the visual scene context by a graph convolution network (GCN) with three stacked layers. Notably, for the graph representation, we focus on the relation between the road participants and whole frame, the self-relation of the node is not considered to reduce the computation. Therefore, GCN in our case is denoted as:

$$\begin{aligned}
\mathbf{X}_t^{(1)} &= Relu(\mathbf{S}_t \mathbf{X}_t \mathbf{W}^0), \\
\mathbf{X}_t^{(2)} &= Relu(\mathbf{S}_t \mathbf{X}_t^{(1)} \mathbf{W}^1), \\
\mathbf{g}_t &= softmax(\mathbf{S}_t \mathbf{X}_t^{(1)} \mathbf{W}^{(2)}),
\end{aligned} \tag{7}$$

where $\mathbf{g}_t$ denotes the visual scene context feature (named as VSC feature) among N_t road participants and the whole frame in the t^{th} video frame, $Relu()$ denotes the *relu* operation, and $\mathbf{W}^{(0)}, \mathbf{W}^{(1)}, \mathbf{W}^{(2)}$ specify the parameters of the 1^{st}, the 2^{nd}, and the 3^{rd} layer of graph convolution. In this work, $\mathbf{g}_t \in \mathbb{R}^{1\times p}$, where p denotes the number of hidden state in GCN, denoted as 256 in this work.

D. Collaborative Multi-Task Consistency Learning

With the predicted $\hat{I}_{T+1}$ and the object locations $\hat{C}_{T+1}^{1:N_T}$, we feed them into our DSCR to obtain the VSC feature $\mathbf{g}_{T+1}^p$ of the predicted $(T+1)^{th}$ frame. Similarly, the VSC feature $\mathbf{g}_{T+1}^r$ of the true $(T+1)^{th}$ frame is also obtained. In order to learn the optimal VSC, we design a graph generative adversarial network (Graph-GAN) model to adversarially training the visual scene context generator $\mathcal{G}$ and the discriminator $\mathcal{D}$. Therefore, the loss function of collaborative multi-task consistency learning is defined as:

$$\min_{\mathcal{G},\mathcal{D}} : \underbrace{\lambda_f \mathcal{L}_{appearance} + \lambda_l \mathcal{L}_{motion} + \lambda_{\mathcal{G}} \mathcal{L}(\mathcal{G}|\mathcal{D})}_{A} + \underbrace{\lambda_{\mathcal{D}} \mathcal{L}(\mathcal{D}|\mathcal{G})}_{B}, \tag{8}$$

where $\mathcal{L}_{appearance}$ adopted the summation of the same intensity loss and gradient loss to [4], $\mathcal{L}_{motion}$ denotes the loss of future location prediction, specified as $\mathcal{L}_l = \sum_k \mathbf{RMSE}(\hat{C}_{T+1}^k, C_{T+1}^k)$, respectively. $\mathbf{RMSE}(\cdot,\cdot)$ calculates the root mean square error. $\mathcal{L}(\mathcal{G}|\mathcal{D})$ and $\mathcal{L}(\mathcal{D}|\mathcal{G})$ are the generative loss and discriminative loss, which are denoted as:

$$\begin{aligned}
\mathcal{L}(\mathcal{G}|\mathcal{D}) &= \frac{1}{2}(1 - \mathbf{MSE}(\mathbf{g}_{T+1}^{p,\pi}, \mathbf{g}_{T+1}^{r,\pi-1})), \\
\mathcal{L}(\mathcal{D}|\mathcal{G}) &= \frac{1}{2}(1 + \mathbf{MSE}(\mathbf{g}_{T+1}^{p,\pi-1}, \mathbf{g}_{T+1}^{r,\pi})),
\end{aligned} \tag{9}$$

where π specifies the iteration step for optimization, and $\mathbf{MSE}(\cdot,\cdot)$ is the mean square error [50].

For optimizing Eq. 8, the terms A and B are alternately trained. This setting means that we want $\mathbf{MSE}(\mathbf{g}_{T+1}^{p,\pi}, \mathbf{g}_{T+1}^{r,\pi-1})$ to be true when training $\mathcal{G}$ and make $\mathbf{MSE}(\mathbf{g}_{T+1}^{p,\pi}, \mathbf{g}_{T+1}^{r,\pi-1})$ be false when training $\mathcal{D}$.

E. Traffic Accident Determination

After training Eq. 8, we adopt it to determine the occurrence degree of accident for each frame in testing. As aforementioned, we aim to consider the advantages of frame prediction and location prediction, and the context relation prediction between frames to fulfill the appearance, motion, and visual scene context consistency learning. This work designs a simple yet efficient fusion strategy, as illustrated by Fig. 6, to combine the consistency measurement, denoted as:

$$\begin{aligned}
\text{SSC-TAD}_t^a &= D(I_{T+1}, \hat{I}_{T+1}), \\
\text{SSC-TAD}_t^{am} &= \text{SSC-TAD}_t^a \times (\text{SSC-TAD}_t^m + 1), \\
\text{SSC-TAD}_t^{amc} &= \text{SSC-TAD}_t^{am} \times (\text{SSC-TAD}_t^c + 1),
\end{aligned} \tag{10}$$

where SSC-TAD_t^a is the appearance consistency measurement, $D(I_{T+1}, \hat{I}_{T+1})$ denotes the Peak Signal to Noise Ratio (PSNR) difference between I_{T+1} and $\hat{I}_{T+1}$, where SSC-TAD_t^m specifies the motion consistency measurement, and is computed by $\mathbf{IOU}(\hat{C}_{T+1}^{1:K}, C_{T+1}^{1:K})$ which computes the Intersection over Union (IOU) of all predicted locations with the true ones. SSC-TAD_t^c measures the visual scene context consistency. Notably, before the weighted summation in Eq. 10, each metric is normalized into [0,1].

The behind meaning of Eq. 10 is that because the learning of frame prediction module is supervised by the raw video frames, which has no label error. However, the object location prediction and driving scene context prediction modules must do the object detection in advance, which may include the detection error. Therefore, we set the appearance consistency as the baseline and fuse the motion and context consistency gradually. The computation of the visual context consistency may cause fluctuation because of the object detection error. Therefore, we introduce a Savitzky-Golay filter [51] to smooth

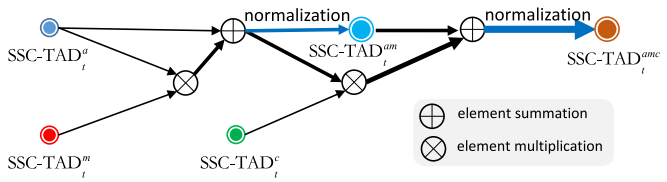


Fig. 6. The fusion strategy of three kinds of consistency metrics. The operation of "normalization", marked by blue color, is max-min normalization with the range of [0,1]. The thickness of the lines represents the amount of fused consistency information.

TABLE I
THE CHARACTERISTICS OF TESTING SET OF A3D AND DADA-2000. AT.: ANNOTATION TYPES, BT.: BEHAVIOR TYPES

Datasets	videos	frames	AT.	BT.
A3D	254	20,839 (10fps)	temporal	no
DADA-2000	212	80,071 (30fps)	temporal, spatial	yes

the obtained visual scene context consistency scores between frames incrementally, and defined as:

$$\text{SSC-TAD}^c_t = \frac{1}{M+1}\sum_{c=t-M}^{t}\frac{1}{2}(\textbf{Cosine}(\mathbf{g}^p_c, \mathbf{g}^r_c)+1)h_c, \quad (11)$$

where M is the window size for smoothing, and h_c is a smoothing parameter.

IV. EXPERIMENTS AND DISCUSSIONS

A. Dataset

In this section, we will evaluate the performance of the proposed method. In this work, we utilize two challenging datasets as the comparison benchmark, i.e., the AnAn Accident Detection (A3D) dataset [3] and DADA-2000 [14] previously collected by ourselves. The videos in A3D and DADA-2000 were collected from various websites, where the camera setting or alignment in each video is different.

Following the work of [3], we utilize Honda Egocentric View Intersection (HEV-I) [52] with 79991 frames in 230 video sequences to train our model, and adopt 254 videos of A3D (with 20,839 frames) for testing. For our DADA-2000 dataset, we sampled 212 videos (with 80,071 frames) to test, where the data splitting of DADA-2000 is released in the website,[1] and most of the testing data is the same with our previous work [53] with tiny tuning for balancing accident categories. Specially, we take 8-10 sequences for each accident category, and some typical frames in each accident category are demonstrated in Fig. 7. Totally, this work has 26 kinds of accidents in the testing set of DADA-2000. For the ground-truth, the temporal labels for the accident frames are set as 1 and 0 for the frames in the normal driving situations. Notably, for the accident videos involved collision objects, the temporal accident window starts once the collision object appears in the field of vision, which covers a time window from "*anomaly to accident*" (A2A). Therefore, this annotation pursues an early detection of traffic accident actually. We also annotated the spatial location of the object to be involved in the traffic accident. The attribute characteristics of the testing datasets in this work is demonstrated in Table. I.

[1]https://github.com/JWFangit/LOTVS-DADA.

B. Implementation Details

For optimizing Eq. 8, Adam optimizer is adopted with the learning rate of 10^{-4} and λ_f, λ_l, $\lambda_{\mathcal{G}}$, $\lambda_{\mathcal{D}}$ are set as 1, 1, 0.1 and 1, respectively. For the training of Eq. 8, we take an alternative training process (with 51 alternations), where each alternative training step contains 200 iterations of the generator and 100 iterations of discriminator for all the training data. M and h_c are set as 10 and 1 in this work. In addition, for the consistency measurement, we input T video frames. Based on our previous investigations [37], [53], inputting more frames for video frame prediction problem will cost more computation resources while show little performance gain, even with a performance degradation because of the complex scene variation in longer time window. Hence, similar to [37], the best T is set as 4. The experiments are conducted on a platform with one NVIDIA RTX2080Ti GPU with 11GB RAM. Because there are many models that share the same weight parameters, the weight size of the proposed method is 15.89M, and the testing efficiency of the proposed method can achieve $7.5\,fps$ after exhaustive evaluation.

C. Evaluation Setups

1) Metrics:

a) Area under ROC curve (AUC): Following the video anomaly detection [4], we first employ the area under the standard frame-level Receiver Operating Characteristic curve (ROC) to evaluate the performance, which focuses on the computing the True Positive Rate (TPR) and the False Positive Rate (FPR). The larger AUC prefers a better performance.

b) Success rate curve: Beside AUC, we also want to evaluate the success rate of each method by computing the sequence ratio, where the AUC values larger than the range from 0 to 1. This metric can give the overall performance comparison on the method's adaptation ability for different driving scenarios.

c) Average precision (AP): In addition, we also want to evaluate AP for traffic accident detection. AP evaluates the correctness of detection, and is computed by counting the average of the detection precision of all testing videos, and the precision is calculated by the ratio between the correctly detected accident frames with all the detected accident frames (consisting correctly and wrongly detected positives). For one detection method, the positive and negative labels of the detected frames are determined by a pre-defined threshold. In this work, if the occurrence degree of traffic accident larger than 0.5, it is positive, and vice versa for the negative.

2) Competitors: For the superiority evaluation, we select three methods representing the state-of-the-art to make comparison with the proposed method. Firstly, the unsupervised traffic accident detection work proposed by Yao *et al.* [3] is chosen as the baselines, which determines the accident by predicting the object location (named as OLP-TAD).

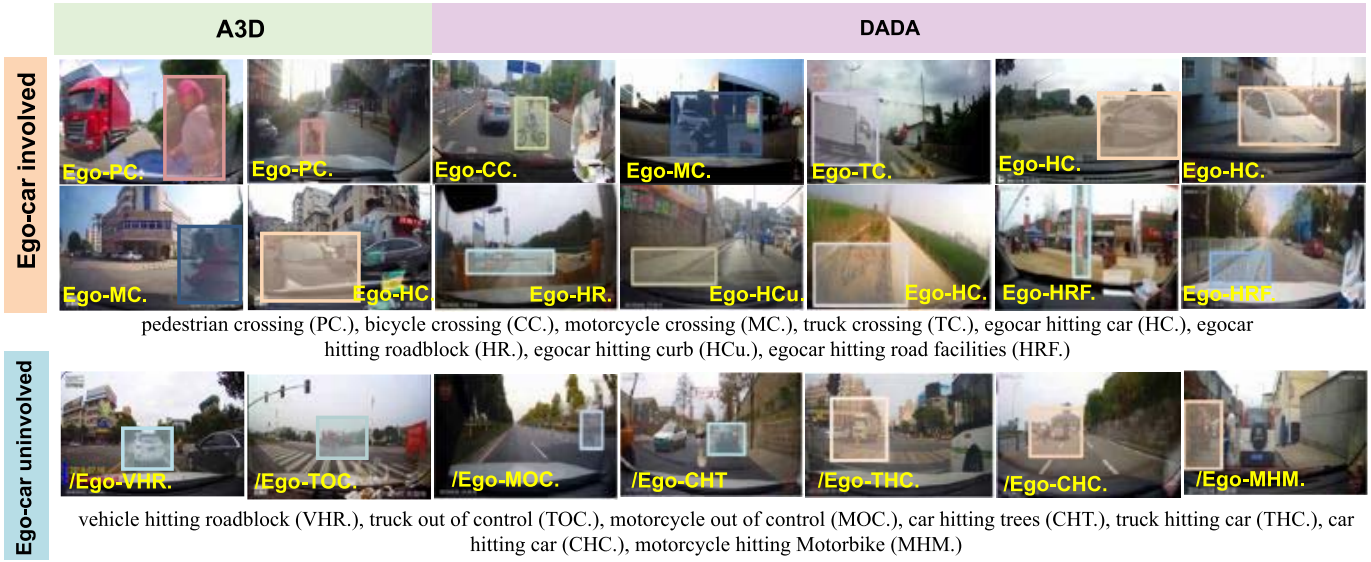


Fig. 7. Typical sample frames of A3D and DADA-2000 dataset, where the accident categories of the videos are noted in this figure. "Ego-" and "/Ego-" denotes the categories which involves the ego-car or does not involve the ego-car.

Secondly, since this work takes the frame prediction as a main component, we also take two competitive frame prediction methods, i.e., frame prediction by generative adversarial network [4] (Liu-FP-TAD), multi-branch deep mask learning based frame prediction (DeepMask-FP-TAD) [37] to verify the performance of traffic accident detection.

In order to provide a fair comparison, all of the methods here are re-trained with the same training dataset, i.e., Honda Egocentric View Intersection (HEV-I) [52] with 79991 frames.

D. Ablation Studies

In order to check the function of different modules in the proposed method, we open an ablation study, and three kinds of settings are evaluated: 1) the whole **SSC-TAD**^{amc} model which combines the appearance, motion and context consistency measurement; 2) the traffic accident detection by only frame prediction (**SSC-TAD**^{a}); 3) the traffic accident detection by fusing location prediction and frame prediction (**SSC-TAD**^{am}).

In terms of frame prediction, following the same criterion of [4], the Peak Signal to Noise Ratio (PSNR) value is taken to check the quality of the predicted frames, and the difference of PSNR values of two adjacent frames are used to determine the occurrence degree of traffic accident. The smaller value indicates a similar appearance of two frames and is set as normal, and vice versa for the frame step in the accident.

As for object location prediction, we propose to adopt the Intersection over Union (IOU) to determine the occurrence degree of traffic accident. IOU computes the overlapping rate between the object bounding boxes in the true future frame and the predicted ones with the same time step. Then the occurrence degree of traffic accident is calculated by averaging all the overlapping rates of the bounding boxes, and the larger IOU pursues a higher degree of normal driving and vice versa for accident frames. For the performance evaluation, the same AUC and AP metrics are used for all the baselines.

1) Component Testing of SSC-TAD: In this work, there are some components in the whole architecture of SSC-TAD^{amc}. The main ones associate with the optical flow path, including the choice of Conv-LSTM (Eq. 2) and Gated-RNN (GRU in this work) (Eq. 5). In addition, the DeepSort in the temporal object association is another main component in object location prediction. In order to check the role of these components, we take the current version of the whole architecture of SSC-TAD^{amc} as a baseline and progressively remove or replace some components gradually. Then, we re-train the modified model with the same training setting, and present the performance comparison of these components on the same testing set of A3D and DADA-2000. We change GRU in Eq. 5 into Long Short Term Memory (LSTM) with the same 256-dimension of hidden states. Conv-LSTM is related with the optical flow image encoder path. If there is no optical flow, Conv-LSTM is omitted automatically. As for the DeepSort component, we change it with Hungarian algorithm [54] which is popular in multi-object temporal association. The detailed results are demonstrated in Table. II.

From the results in Table. II, we find optical flow with Conv-LSTM is the most important component in this work. GRU is better than LSTM but with the smallest performance gap. Besides, the parameters in GRU is fewer than the ones of LSTM. For the temporal object association component, DeepSort is superior to the traditional Hungarian which only takes the object location (without object appearance) into account, and is easily to introduce the index switch issue. In summarize, the components in our work is a best version.

E. Overall Performance Evaluation

The overall performance of different methods are demonstrated in Table. III.

TABLE II

COMPONENT ROLE CHECKING FOR THE WHOLE ARCHITECTURE OF SSC-TAD WITH THE SAME TESTING SETS OF A3D AND DADA-2000. TOA.: TEMPORAL OBJECT ASSOCIATION; HG.: HUNGARIAN; DS.:DEEPSORT

Optical flow	Gated RNN		TOA.		AUC(%)
Conv-LSTM	LSTM	GRU	HG.	DS.	A3D/DADA
		✓		✓	50.3/52.6 ↓
	✓			✓	51.3/50.3 ↓
✓	✓			✓	59.9/60.3 ↓
✓		✓	✓		53.2/54.6 ↓
✓		✓		✓	**67.8/66.5**

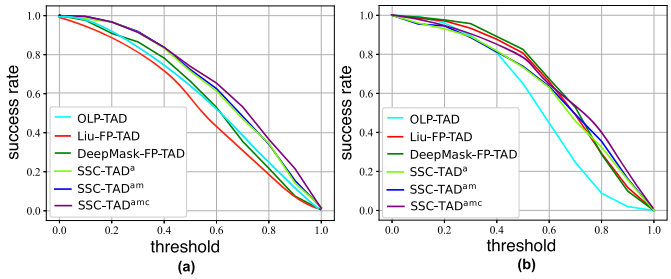


Fig. 8. The success rate of each method on (a) A3D and (b) DADA datasets.

Table. III demonstrates the comparison results. From this table, we can see that the performances of the competitors on our DADA-2000 dataset is weaker than the ones on A3D dataset, especially in AP. This phenomenon verifies that the our DADA-2000 is more challenging than A3D, which is also observed by the large performance difference of AP and AUC values on our DADA-2000. As aforementioned, AP computes the correctness of the approaches, the lower AP means on our dataset means that there are more wrongly detected positives and the traffic accident occurs ruleless in the sequence. From the characteristics of the testing set of A3D and DADA-2000, the average sequence length is about 82 frames and 377 frames, respectively. It implies a more manifest few-shot issue in DADA-2000. Actually, it is more natural in the practical driving situations. In addition, our proposed method generates significant superiority than other state-of-the-art methods in A3D. From the comparison results, we can see that the frame prediction is very helpful for the accident determination, and our visual context consistency measurement can boost the performance manifestly.

In addition, we also demonstrate the success rate value of different methods on A3D and DADA-2000 datasets. The success rate is computed by the ratio of the sequences whose AUC value is larger than a predefined threshold from 0 to 1. The plots are shown in Fig. 8. From this figure, we can see that our ablation methods generate competitive performance, and our SSC-TAD^{amc} is stable in two datasets. Because of the more challenging situation in DADA dataset, OLP needing object detection modules shows the weakest performance. Besides, some typical sequences with the occurrence degree of traffic accident are presented in Fig. 9, which demonstrates that our SSC-TAD can discriminate the accident or non-accident situation better than other ones.

TABLE III

PERFORMANCE COMPARISON (%) OF THE BASELINES AND THE STATE-OF-THE-ART

Baselines	A3D		DADA-2000	
	AP	AUC	AP	AUC
OLP-TAD [4]	52.4	59.8	30.1	56.1
Liu-FP-TAD [18]	55.5	50.7	37.0	66.6
DeepMask-FP-TAD [37]	59.9	59.0	40.4	**67.5**
SSC-TAD^{a}	61.2	65.4	38.4	64.6
SSC-TAD^{am}	61.5	65.8	40.3	65.2
SSC-TAD^{amc}	**64.1**	**67.8**	**42.7**	66.5

TABLE IV

PERFORMANCE COMPARISON (%) OF THE METHODS IN "EGO-CAR INVOLVED" AND "EGO-CAR UNINVOLVED" SITUATIONS ON DADA-2000 DATASET

Baselines	Ego-car involved		Ego-car uninvolved	
	AP	AUC	AP	AUC
OLP-TAD [4]	31.5	55.7	30.3	56.9
Liu-FP-TAD [18]	42.3	71.3	26.2	57.1
DeepMask-FP-TAD [37]	46.6	**73.0**	27.9	56.3
SSC-TAD^{a}	**48.8**	69.7	29.5	57.4
SSC-TAD^{am}	45.7	68.4	29.7	58.3
SSC-TAD^{amc}	44.3	67.6	**30.4**	**58.7**

F. Performance Evaluation w.r.t., Ego-Car Involved and Ego-Car Uninvolved Situations

In this work, we further analyze the performance of the methods on two kinds of situations, i.e., the ego-car involved and ego-car uninvolved situations. Ego-car involved situation often demonstrates a hitting behavior with other static or dynamic road participants. In this situation, the video frames appear a large motion change when the accident occurs. As for the ego-car uninvolved situation, there are commonly some other road participants that make a collision and the background of the scene in the captured videos shows stable relatively, and the objects in the collision are very small in sometimes. Therefore, these two kinds of situations have manifest difference.

Table. IV demonstrates this comparison results. It can be seen that the performances of most of methods in the case of ego-car involved are higher than that of ego-car uninvolved. It is because the objects in the traffic accident in the ego-car involved situations usually demonstrate larger resolution than that of ego-car uninvolved. In addition, DeepMask-FP-TAD shows a good performance on the ego-car involved situations, which learned a mask to focus on the image region with large motion change and can find the large motion change effectively. As for the ego-car uninvolved situations, because of the relative small motion change of the surrounding objects and the dynamic camera motion, DeepMask-FP-TAD appears poor performance. In contrast, the object location prediction module demonstrates promising performance, e.g., OLP-TAD and our SSC-TAD^{am}. Under this situation, the visual context consistency improves the performance and show the best result for the ego-car uninvolved situations.

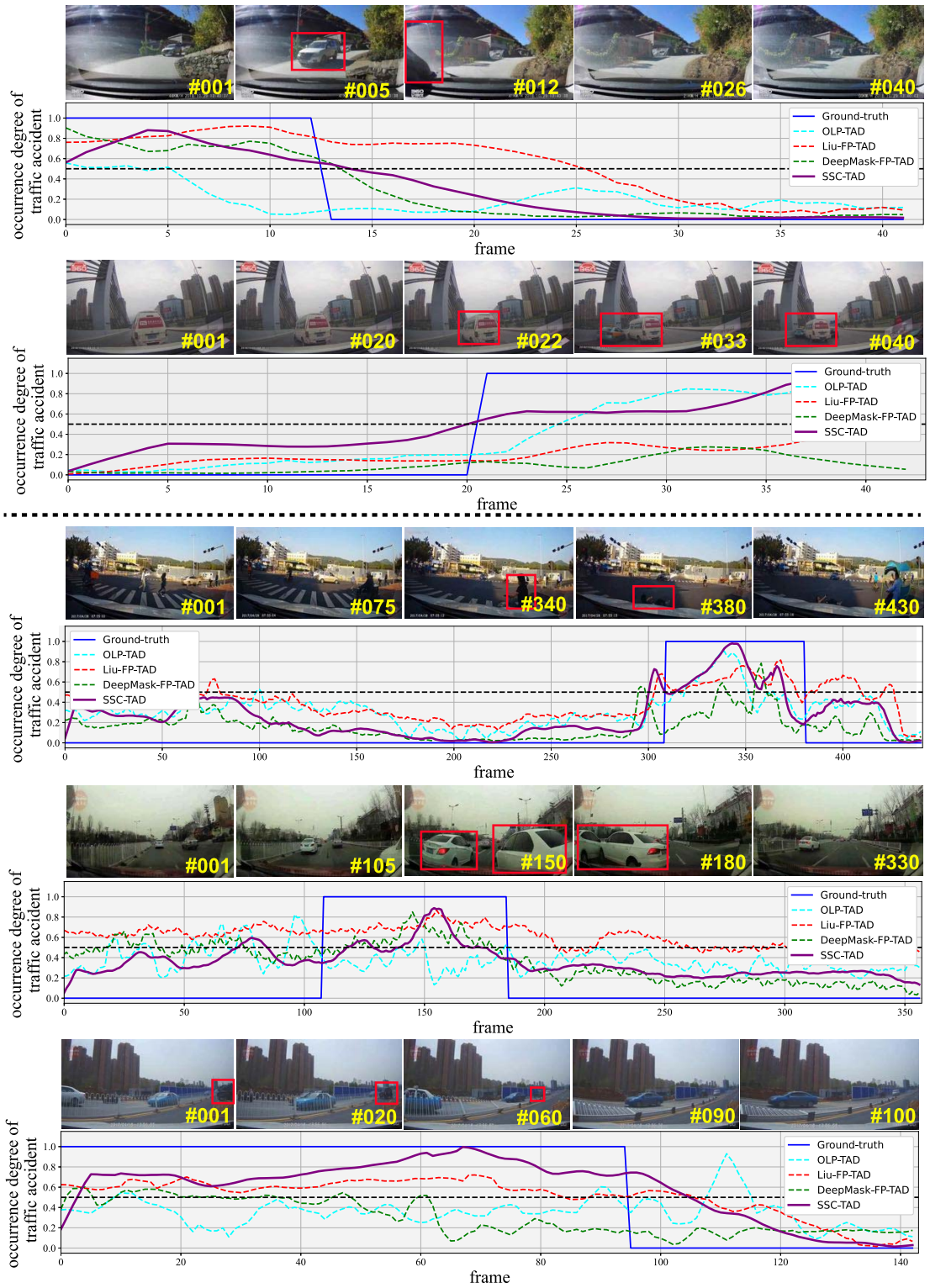


Fig. 9. The occurrence degree curve of OLP-TAD [4], Liu-FP-TAD [18], DeepMask-FP-TAD [37] and our SSC-TAD (the whole model is used here) on some typical sequences of A3D and DADA dataset. The first two rows are taken from A3D dataset and the others are the ones in DADA dataset, where the red boxes mark the object that involves in the accident.

G. Performance Evaluation w.r.t., Different Driving Scenarios

Different driving scenarios (e.g., highway, rural road, urban road) have differing kinds of participants in traffic accidents. Because of the moving characteristics of different road participants, the performance of different methods may be different.

In highway scenes, most accidents are caused by vehicle collisions or out of control, where the vehicles demonstrate high speed, which may cause large scale change of the vehicles to be involved in the accident. Therefore, the distinction of object location prediction is poor than that of the frame prediction strategy, as shown in Table. V. In addition, our

TABLE V

PERFORMANCE COMPARISON (%) OF THE METHODS ON THE DRIVING SCENARIOS OF "*highway*," "*urban road*" AND "*rural road*" IN THE TESTING SET OF DADA-2000 DATASET

Baselines	*highway*		*urban road*		*rural road*	
	AP	AUC	AP	AUC	AP	AUC
OLP-TAD [4]	22.3	52.0	31.8	55.9	31.8	61.7
Liu-FP-TAD [18]	26.6	65.0	39.2	65.8	38.1	71.2
DeepMask-FP-TAD [37]	37.9	71.4	40.2	65.2	38.6	**72.8**
SSC-TAD^{a}	**39.1**	**72.8**	42.2	64.0	37.4	65.0
SSC-TAD^{am}	35.2	71.1	42.0	63.8	38.2	65.1
SSC-TAD^{amc}	35.3	69.4	**44.7**	**65.9**	**43.4**	69.7

methods SSC-TAD^{am} and SSC-TAD^{amc} show a degradation after fusing the object location prediction module, e.g., that the AUC value changes from 72.8% to 71.1% and 69.4%, respectively. As for rural scene scenarios, most accidents occur in low-speed scenes and the frame change is relatively slow, where the participant categories within this kind of scenario often include pedestrians, bicycles or motorcycles and are without the constraint of road rules. In this scenario, the object to be involved in the accident often demonstrate crossing behavior and object location prediction module shows better performance than the one of highway scenario because of the smaller scale change of the objects. It is worth noting that owning to the deep mask constraint for motion change in DeepMask-FP-TAD, it generates the best AUC value (72.8%) in rural scenario. However, our SSC-TAD^{amc} shows best AP value (43.4%) because of the further context relation consistency consideration.

As for the urban road scenario, the driving speed of the vehicles is between the ones of highway scenario and rural road scenario. At this structured road condition, the context relation consistency plays an important role because of the apparent road rules. Under this condition, the traffic accident is usually caused by disobeying traffic rules, and the normal situation presents a relative stable context relation of road scene. Therefore, our SSC-TAD^{amc} demonstrates the best AP value and AUC value.

H. Performance Evaluation w.r.t., Different Behavior Types

As for each traffic accident, the collision could be caused by different behavior type of road participants. In this work, we also compare the performance on different behavior types of the ego-car involved traffic accidents, which is useful for checking the sensitivity of different methods to different behavior types. Here, we group the behavior reason in testing sequences into three large categories, i.e., "*crossing*", "*hitting*" and "*out of control*", and demonstrate the fine-grained behavior types in each group in Table. VI. Totally, fifteen ego-car involved traffic accident categories are compared in this part. Notably, the division criterion for these behavior types is enumerated by the practical relations between the road participants in traffic accidents. For example, the relations between pedestrians and cars in traffic accident mainly are crossing and hitting. The detailed information can be seen from our previous work [14].

By our previous investigation [14] on the spatial occurrence characteristics of "*crossing*" and "*hitting*" (as shown in Fig. 10), crossing behavior usually occurs a horizontal movement in the field of driving vision, and shows a manifest different motion pattern with that of the background region. Therefore, compared with other behavior types, the accident with crossing behavior is easier to be detected than other behaviors and all methods show better performance than that on other behavior types, which is verified by the results in Table. IV. Notably, the spatial occurrence map in Fig. 10 is obtained by summarizing the **center point occupancy** of the "*objects to be involved in accident*" (termed as **crash-object**) in each frame of certain accident category. To be clear for demonstration, the center point of crash-object was expanded with a rectangle neighborhood having the size of 14×14, where the value within it is set as 1 and 0 for vice versa.

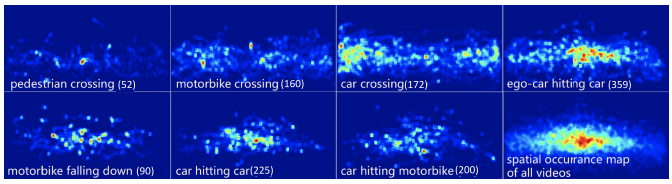


Fig. 10. The spatial occurrence characteristics of different traffic accident categories, where the number in the bracket denotes the number of sequences within the categories. This figure stems from our previous work [14].

In addition, for the crossing behavior, the crash-object demonstrates a large scale change, which makes that the object location prediction may be degraded because of the large-scale change or the object detection bias. As for the hitting behavior, the spatial occurrence of accident often groups into the middle of the vision plane. Under this condition, the frame prediction is easier to find the large motion change. However, in many times, the motion change before accident may be normal. Therefore, the frame prediction based methods (DeepMask-FP-TAD and Liu-FP-TAD) shows a more promising performance than the object location prediction, and our frame prediction version, i.e.,SSC-TAD^{a}, demonstrates 84.9% AUC value, shown in Table. IV. Our SSC-TAD^{amc} shows best average precision value (AP) because of the comprehensive consideration of frame, object location, and context relation consistency. In the future, long-term object location prediction is preferred for hitting accident detection.

As for the "*out of control*" situation, it is caused by the false maneuvering or the bad road condition. The scene motion change is commonly chaotic. Therefore, all the methods demonstrate a mediocre performance, and our SSC-TAD^{amc} shows best because of the fusion of different consistency measurements. There is no moving object in the scene commonly, and the robust perception of road structure change is preferred.

I. Further Analysis of Frame Prediction for TAD

Recently, many works take the frame prediction module in the video anomaly or the accident detection topic. The initial purpose of frame prediction is to generate the future frames as much the same to the ground-truth as possible. However, for the Traffic Accident Detection (TAD) utilized frame prediction module, *is better frame prediction more suitable for TAD?* We think that the answer is not always "yes".

TABLE VI

PERFORMANCE COMPARISON (%) OF THE METHODS ON *fifteen* EGO-CAR INVOLVED ACCIDENT CATEGORIES WITH THE BEHAVIOR TYPES OF "*crossing*", "*hitting*" AND "*out of control*" IN THE TESTING SET OF DADA-2000 DATASET. THE BEHAVIOR TYPES OF "*crossing*" CONTAINS *pedestrian crossing* (PC.), *bicycle crossing* (BC.), *car crossing* (CC.), *truck crossing* (TC.), AND *motorcycle crossing* (MC.). THE BEHAVIOR TYPES OF "*hitting*" CONSISTS OF *egocar hitting pedestrian* (HP.), *egocar hitting bicycle* (HB.), *egocar hitting motorcycle* (HM.), *egocar hitting truck* (HT.), AND *egocar hitting car* (HC.), THE BEHAVIOR TYPES OF "*out of control*" INCLUDES *motorcycle out of control* (MOC.), *truck out of control* (TOC.), *truck hitting trees* (THT.), *egocar hitting trees* (EHT.), AND *motorbike hitting trees* (MHT.)

Baselines	PC.		BC.		CC.		TC.		MC.		Average	
	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC
OLP-TAD [4]	30.9	56.8	38.9	56.4	29.6	61.4	45.0	61.7	38.6	61.3	36.6	59.5
Liu-FP-TAD [18]	43.1	**71.9**	44.5	64.9	29.3	67.8	70.7	83.9	38.9	**65.1**	45.3	70.7
DeepMask-FP-TAD [37]	**49.5**	70.6	51.7	**70.4**	31.7	68.4	71.3	80.6	**41.3**	62.3	49.1	70.5
SSC-TAD^{a}	39.9	71.1	59.5	68.1	36.3	66.3	73.4	81.1	39.9	64.0	49.8	69.5
SSC-TAD^{am}	42.0	71.6	55.8	66.6	35.7	66.7	71.5	81.7	40.6	64.1	49.1	70.1
SSC-TAD^{amc}	43.5	69.2	**59.7**	68.5	**46.9**	**70.3**	**73.7**	**84.2**	39.6	64.7	**52.7**	**71.4**
Baselines	HP.		HB.		HM.		HT.		HC.		Average	
	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC
OLP-TAD [4]	41.1	70.2	42.7	58.3	19.2	53.7	40.0	48.1	30.8	52.1	34.8	56.8
Liu-FP-TAD [18]	46.2	82.1	39.9	61.0	28.8	**67.2**	53.4	73.3	47.4	73.7	43.1	**71.5**
DeepMask-FP-TAD [37]	47.6	77.9	43.7	**65.7**	**34.2**	64.9	54.7	**74.6**	44.1	68.9	44.9	70.4
SSC-TAD^{a}	**57.5**	**84.9**	44.7	59.6	23.4	50.8	**57.3**	74.1	60.7	79.6	48.7	69.8
SSC-TAD^{am}	56.7	82.9	47.1	60.0	25.3	50.5	50.3	71.6	61.2	76.8	48.1	68.4
SSC-TAD^{amc}	55.4	76.9	**47.8**	59.1	29.3	53.9	53.1	71.6	**61.3**	**79.7**	**49.4**	68.3
Baselines	MOC.		TOC.		THT.		EHT.		MHT.		Average	
	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC
OLP-TAD [4]	24.6	51.1	**26.9**	**62.7**	24.7	46.7	21.6	35.7	29.4	55.0	25.3	50.2
Liu-FP-TAD [18]	23.5	56.9	24.8	69.1	24.4	43.4	23.3	46.2	29.1	53.9	25.0	53.9
DeepMask-FP-TAD [37]	29.9	61.1	28.8	65.7	24.0	39.5	**26.0**	**49.4**	28.9	55.4	27.5	54.2
SSC-TAD^{a}	30.4	60.5	23.0	62.9	24.0	43.9	22.0	40.4	35.3	54.4	26.9	52.4
SSC-TAD^{am}	29.3	59.6	24.0	68.2	24.1	42.9	21.9	40.1	35.0	56.4	26.9	53.4
SSC-TAD^{amc}	**30.8**	**61.3**	25.5	66.9	**24.8**	**48.4**	22.1	40.3	**37.8**	**57.1**	**28.2**	**54.8**

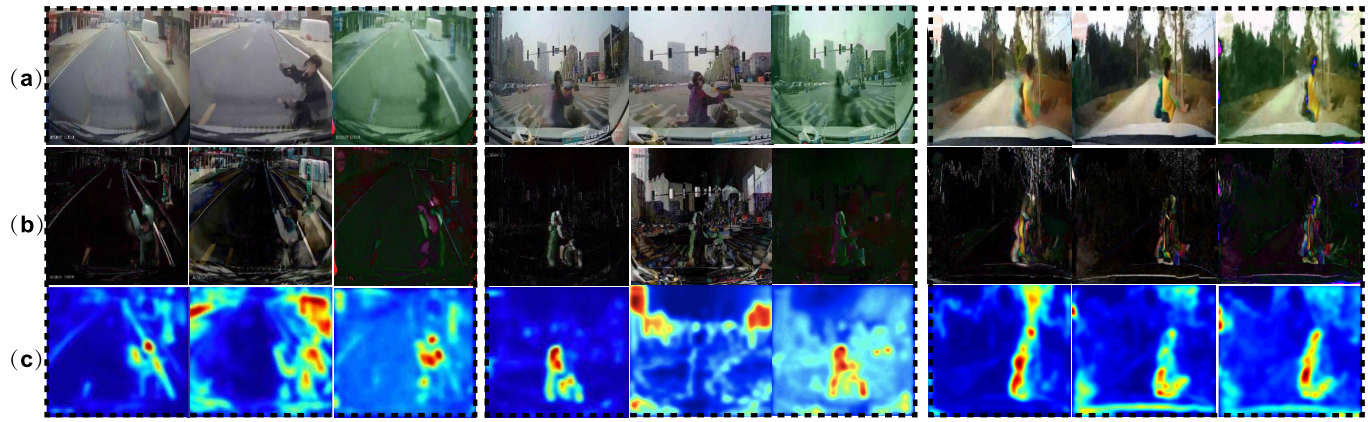


Fig. 11. Some examples for presenting the frame prediction and the anomaly region localization by Liu-FP-TAD [18] (the 1^{st} column in each group marked by the dash boxes), DeepMask-FP-TAD [37] (the 2^{nd} column in each group), and our SSC-TAD^{a} (the third column of each group), where (a) denotes the predicted frame. The images of (b) are generated by computing the absolute value of the subtraction of original frame and the predicted one. The heat maps in (c) are obtained by using a Gaussian kernel to convolute the images of (b), where the Gaussian kernel size is 20 and the variance is 5.

Taking Fig. 11 as an example, from the visual quality of the predicted frame, DeepMask-FP-TAD seems the best, which can be seen from the pedestrian body's texture and the color of the predicted frames. However, for TAD problem, it includes many artifacts after frame subtraction and the anomaly region is not localized well. As for Liu-FP-TAD, although it generates a blurred body region, the bodies of the pedestrians are highlighted, but the anomaly regions are not complete. As for our SSC-TAD^{a} in Fig. 11, although there is a color drift, the localized result of the anomaly region is better than other ones. Actually, in this analysis, we do not want to prove that our method is better than others. In contrast, we want to give a suggestion for the designing of frame prediction module in TAD. *If we want to design a promising frame prediction module for TAD, it needs to make the background region be predicted very well, and vice versa for the target region in the predicted frames*.

J. Some Failure Cases

As shown in Fig. 12, the proposed method still has space to be improved, and generates some typical failure cases. From Fig. 12, we can see that the object scale and light condition are the primary factors for detection failures. Too large or too small scale makes the object location prediction and the frame prediction modules be ineffective for discovering the target object. The dark road environment causes an unclear

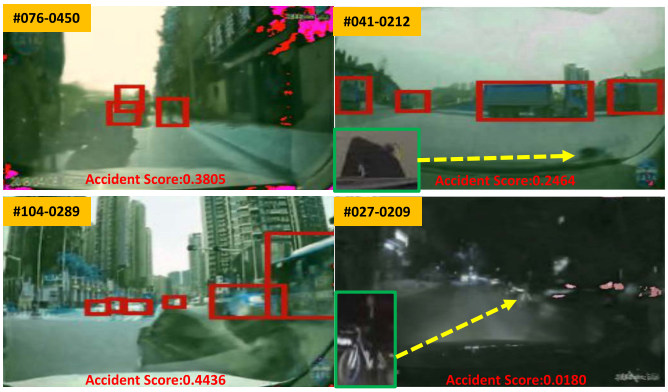


Fig. 12. Some typical failure cases, where the images are all the predicted frames by our work, and the predicted object locations are marked by the red boxes. The green boxes demonstrate the clear crashed objects which are not found by the proposed method. The number in the left-top corner of each picture denotes the sequence index and the frame index (*sequence index-frame index*) in the testing set of DADA-2000.

perception by the cameras. In addition, the incomplete boundary object in the field of vision is easily to be ignored.

We all know that these factors are difficult to be reflected in the practical data collection. Therefore, synthetic data may be useful with the development of open sourced rendering engines. In addition, human assisted vision perception, such as sober drivers' attention, could be involved for the hard situations because of the fast reaction of scene change. In order to ameliorate the perception challenges in dark road environment, edge-computing technique can be introduced by enhancing the imagery quality by the models in the data cloud [55], [56]. Furthermore, the cause-effect association of accidents is promising for an interpretation TAD or traffic accident prediction.

V. Conclusion

In this paper, we proposed a new traffic accident detection method by learning the appearance, motion and the context relation consistency within consecutive frames, which absorbed the advantages of frame prediction and location prediction in previous works, and was fulfilled by a multi-task consistency learning framework with a generative and adversarial training strategy. We also designed a novel fusion strategy to fuse the appearance, motion and the context consistency measurement. Based on the extensive experiments on two challenging datasets, i.e, AnAn traffic accident detection (A3D) and DADA-2000 previously collected by ourselves. The superiority is verified by the comparison with several state-of-the-art methods. Furthermore, we analyzed the performance of each method on different behavior types of traffic accidents, different accident categories, and the ego-car involved or uninvolved situations on our DADA-2000 dataset.

References

[1] C. Chen, L. Liu, S. Wan, X. Hui, and Q. Pei, "Data dissemination for industry 4.0 applications in Internet of Vehicles based on short-term traffic prediction," *ACM Trans. Internet Technol.*, vol. 22, no. 1, pp. 1–18, Feb. 2022.

[2] World Health Organization. *Global Health Observatory: Number of Road Traffic Deaths*. Accessed: 2018. [Online]. Available: https://www.who.int/gho/road-safety/mortality/traffic-deaths-number/en/

[3] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised traffic accident detection in first-person videos," in *Proc. IROS*, 2019, pp. 273–280.

[4] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. CVPR*, 2018, pp. 3313–3320.

[5] H. Kataoka, T. Suzuki, Y. Aoki, and Y. Satoh, "Anticipating traffic accidents with adaptive loss and large-scale incident DB," in *Proc. CVPR*, 2018, pp. 54–60.

[6] H. Kataoka, T. Suzuki, Y. Aoki, and Y. Satoh, "Drive video analysis for the detection of traffic near-miss incidents," in *Proc. ICRA*, 2018, pp. 54–60.

[7] F. Chan, Y. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *ACCV*, vol. 2016, pp. 136–153.

[8] E. Jardim, L. A. Thomaz, E. A. B. da Silva, and S. L. Netto, "Domain-transformable sparse representation for anomaly detection in moving-camera videos," *IEEE Trans. Image Process.*, vol. 29, pp. 1329–1343, 2020.

[9] G. Pang, C. Yan, C. Shen, A. van den Hengel, and X. Bai, "Self-trained deep ordinal regression for end-to-end video anomaly detection," in *Proc. CVPR*, 2020, pp. 12170–12179.

[10] Y. Yao, X. Wang, M. Xu, Z. Pu, E. M. Atkins, and D. J. Crandall, "When, where, and what? A new dataset for anomaly detection in driving videos," *CoRR*, vol. abs/2004.03044, pp. 1–4, Apr. 2020.

[11] Y. Seo, M. Defferrard, P. Vandergheynst, and X. Bresson, "Structured sequence modeling with graph convolutional recurrent networks," in *Proc. NeurIPS*, 2018, pp. 362–373.

[12] W. Bao, Q. Yu, and Y. Kong, "Uncertainty-based traffic accident anticipation with spatio-temporal relational learning," in *Proc. ACM Multimedia*, 2020, pp. 2682–2690.

[13] X. Xie, C. Zhang, Y. Zhu, Y. N. Wu, and S.-C. Zhu, "Congestion-aware multi-agent trajectory prediction for collision avoidance," in *Proc. ICRA*, 2021, pp. 13693–13700.

[14] 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," in *Proc. ITSC*, 2019, pp. 4303–4309.

[15] X. Huang, P. He, A. Rangarajan, and S. Ranka, "Intelligent intersection: Two-stream convolutional networks for real-time near-accident detection in traffic video," *ACM Trans. Spatial Algorithms Syst.*, vol. 6, no. 2, pp. 10:1–10:28, 2020.

[16] K. K. Santhosh, D. P. Dogra, and P. P. Roy, "Anomaly detection in road traffic using visual surveillance: A survey," *ACM Comput. Surv.*, vol. 53, no. 6, p. 119, 2020.

[17] R. V. H. M. Colque, C. Caetano, M. T. L. de Andrade, and W. R. Schwartz, "Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 673–682, Mar. 2017.

[18] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. CVPR*, 2016, pp. 733–742.

[19] V. Kaltsa, A. Briassouli, I. Kompatsiaris, and M. G. Strintzis, "Swarm-based motion features for anomaly detection in crowds," in *Proc. ICIP*, 2014, pp. 2353–2357.

[20] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. CVPR*, 2009, pp. 1446–1453.

[21] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[22] K.-W. Cheng, Y.-T. Chen, and W.-H. Fang, "Gaussian process regression-based video anomaly detection and localization with hierarchical feature representation," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5288–5301, Dec. 2015.

[23] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. CVPR*, 2018, pp. 6479–6488.

[24] A. Basharat, A. Gritai, and M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," in *Proc. CVPR*, 2008, pp. 1–8.

[25] X. Zhu, J. Liu, J. Wang, C. Li, and H. Lu, "Sparse representation for robust abnormality detection in crowded scenes," *Pattern Recognit.*, vol. 47, no. 5, pp. 1791–1799, 2014.

[26] V. Kaltsa, A. Briassouli, I. Kompatsiaris, and M. G. Strintzis, "Multiple hierarchical Dirichlet processes for anomaly detection in traffic," *Comput. Vis. Image Understand.*, vol. 169, pp. 28–39, Apr. 2018.

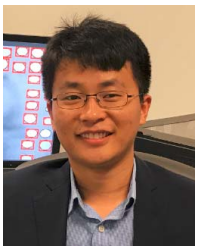
[27] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2014.

[28] F. T. Liu, K. M. Ting, and Z. Zhou, "Isolation forest," in *Proc. ICDM*, 2008, pp. 413–422.
[29] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proc. ICCV*, 2017, pp. 341–349.
[30] H. T. Tran and D. Hogg, "Anomaly detection using a convolutional winner-take-all autoencoder," in *Proc. BMVC*, 2017, pp. 1–12.
[31] N. Li, F. Chang, and C. Liu, "Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes," *IEEE Trans. Multimedia*, vol. 23, pp. 203–215, 2021.
[32] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Comput. Vis. Image Understand.*, vol. 172, pp. 88–97, Jul. 2018.
[33] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-cascade: Cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1992–2004, Apr. 2017.
[34] B. R. Kiran, D. M. Thomas, and R. Parakkal, "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos," *J. Imag.*, vol. 4 no. 2, p. 36, 2018.
[35] B. Mohammadi, M. Fathy, and M. Sabokrou, "Image/video deep anomaly detection: A survey," *CoRR*, vol. abs/2103.01739, pp. 1–8, Mar. 2021.
[36] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, "Efficient gan-based anomaly detection," in *Proc. ICLRW*, 2018, pp. 1–13.
[37] S. Li, J. Fang, H. Xu, and J. Xue, "Video frame prediction by deep multi-branch mask network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1283–1295, Apr. 2021.
[38] J. T. Zhou, L. Zhang, Z. Fang, J. Du, X. Peng, and Y. Xiao, "Attention-driven loss for anomaly detection in video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4639–4647, Dec. 2020.
[39] Y. Yuan, J. Fang, and Q. Wang, "Incrementally perceiving hazards in driving," *Neurocomputing*, vol. 282, pp. 202–217, Mar. 2018.
[40] Y. Yuan, D. Wang, and Q. Wang, "Anomaly detection in traffic scenes via spatial-aware motion reconstruction," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1198–1209, May 2017.
[41] B. Jiang, Y. Zhang, B. Luo, X. Cao, and J. Tang, "STGL: Spatial-temporal graph representation and learning for visual tracking," *IEEE Trans. Multimedia*, vol. 23, pp. 2162–2171, 2021.
[42] S. Liang, X. Wei, S. Yao, and X. Cao, "Efficient adversarial attacks for visual object tracking," in *Proc. ECCV*, 2020, pp. 34–50.
[43] T. Liu, X. Cao, and J. Jiang, "Visual object tracking with partition loss schemes," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 3, pp. 633–642, Mar. 2017.
[44] M. Kilicarslan and J. Y. Zheng, "Predict vehicle collision by TTC from motion using a single video camera," *IEEE Trans. Intell. Transp.*, vol. 20, no. 2, pp. 522–533, May 2018.
[45] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. CVPR*, 2017, pp. 1647–1655.
[46] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*,2015, pp. 802–810.
[47] N. Wojke, A. Bewley, and D. Paulus, "Simple online and real-time tracking with a deep association metric," in *Proc. ICIP*, 2017, pp. 3645–3649.
[48] J. Chung, C. C. Gülçehre, K. Cho, and Y. Bengio, "Gated feedback recurrent neural networks," in *Proc. ICML*, 2015, pp. 2067–2075.
[49] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, May 2017, pp. 2980–2988.
[50] M. Sohn, "Distance and cosine measures of niche overlap," *Soc. Netw.*, vol. 23, no. 2, pp. 141–165, 2001.
[51] R. W. Schafer, "What is a Savitzky-Golay filter?" *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 111–117, Jul. 2011.
[52] Y. Yao, M. Xu, C. Choi, D. J. Crandall, E. M. Atkins, and B. Dariush, "Egocentric vision-based future vehicle localization for intelligent driving assistance systems," in *Proc. ICRA*, 2019, pp. 9711–9717.
[53] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "DADA: Driver attention prediction in driving accident scenarios," *IEEE Trans. Intell. Transp. Syst.*, early access, Jan. 1, 2021, doi: 10.1109/TITS.2020.3044678.
[54] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics*, vol. 2, nos. 1–2, pp. 83–97, 1955.
[55] Y. Wu, H. Guo, C. Chakraborty, M. Khosravi, S. Berretti, and S. Wan, "Edge computing driven low-light image dynamic enhancement for object detection," *IEEE Trans. Netw. Sci. Eng.*, early access, Feb. 14, 2022, doi: 10.1109/TNSE.2022.3151502.
[56] S. Wan, S. Ding, and C. Chen, "Edge computing enabled video segmentation for real-time traffic monitoring in internet of vehicles," *Pattern Recognit.*, vol. 121, Oct. 2022, Art. no. 108146.

Jianwu Fang (Member, IEEE) received the Ph.D. degree in signal and information processing (SIP) from the University of Chinese Academy of Sciences, China, in 2015. He is currently the Director and an Associate Professor with the Laboratory of Traffic Vision Safety (LOTVS) and the Department of Big Data Management and Application, College of Transportation Engineering, Chang'an University, Xi'an, China. He has published many papers on top-ranked journals and conferences, such as IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, AAAI, ICRA, and ITSC. His research interests include computer vision and pattern recognition and their applications intelligent transportation.

Jiahuan Qiao received the bachelor's degree in automation from the Xi'an University of Posts and Telecommunications in 2018 and the master's degree in pattern recognition and intelligent systems from Chang'an University in 2021. Her research interests include intelligent transportation and driving anomaly detection.

Jie Bai received the bachelor's degree in automation from Chang'an University in 2020, where he is currently pursuing the master's degree in traffic information engineering and control. His research interests include intelligent transportation and pedestrian intent prediction.



Hongkai Yu (Member, IEEE) received the Ph.D. degree in computer science and engineering from the University of South Carolina, Columbia, SC, USA, in 2018. He is currently an Assistant Professor with the Department of Electrical Engineering and Computer Science, Cleveland State University, USA. His research interests include computer vision, machine learning, deep learning, and intelligent transportation systems.

Jianru Xue (Member, IEEE) received the B.S. degree from the Xi'an University of Technology in 1994 and the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xian, China, in 1999 and 2003, respectively. Since 1999, he has been with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, where he is currently a Full Professor. He had worked at Fuji Xerox, Tokyo, Japan, from 2002 to 2003; and visited the University of California, Los Angeles, from 2008 to 2009. His research interests include computer vision, visual localization and navigation, and video coding based on analysis. He and his team won the IEEE ITS Institute Lead Award in 2014. He and his students won the Best Application Paper Award in Asian Conference on Computer Vision 2012.