# An Attention-guided Multistream Feature Fusion Network for Early Localization of Risky Traffic Agents in Driving Videos

Muhammad Monjurul Karim, Zhaozheng Yin, *Senior Member, IEEE*, and Ruwen Qin*, *Member, IEEE*

***Abstract*—Detecting dangerous traffic agents in videos captured by vehicle-mounted dashboard cameras (dashcams) is essential to ensure safe navigation in complex environments. Accident-related videos are just a minor portion of the driving-related big data, and the transient pre-accident process is highly dynamic and complex. Besides, risky and non-risky traffic agents can be similar in their appearance. These make risky traffic agent localization in the driving video particularly challenging. To this end, this paper proposes an attention-guided multistream feature fusion network (AM-Net) to localize dangerous traffic agents from dashcam videos ahead of potential accidents. Two Gated Recurrent Unit (GRU) networks use object bounding box and optical flow features extracted from consecutive video frames to capture spatio-temporal cues for distinguishing risky traffic agents. An attention module, coupled with the GRUs, learns to identify traffic agents that are relevant to an accident. Fusing the two streams of global and object-level features, AM-Net predicts the riskiness scores of traffic agents in the video. In supporting this study, the paper also introduces a new benchmark dataset called Risky Object Localization (ROL). The dataset contains spatial, temporal, and categorical annotations of the accident, object, and scene-level attributes. The proposed AM-Net achieves a promising performance of 85.59% AUC on the ROL dataset. Additionally, the AM-Net outperforms the current state-of-the-art for video anomaly detection by 3.5% AUC on the public DoTA dataset. A thorough ablation study further reveals AM-Net's merits by assessing the impact of its constituents.**

***Index Terms*—accident prediction, early risky object localization, autonomous vehicle, multi-modal, attention, deep learning, dashcam**

## I. Introduction

Autonomous driving and Advanced Driver Assistance Systems (ADAS) have made rapid progress in recent years [1]. Although research is progressing positively towards a vision of more comfortable and safer driving experiences, there are still concerns about traffic accidents. From 2014 to March 17, 2023, 564 autonomous vehicle collisions were reported in California [2]. Moreover, according to the 2018 Global Status Report on Road Safety from World Health Organization, about 1.35 million people are killed in traffic accidents yearly [3]. Developing an intelligent driving function to help drivers or autonomous systems identify and localize risky traffic agents is urgently needed to reduce collisions and fatalities.

Muhammad Monjurul Karim, and Ruwen Qin are with the Department of Civil Engineering, Stony Brook University, Stony Brook, NY 11794, USA.

Zhaozheng Yin is with the Department of Biomedical Informatics, Department of Computer Science, and AI Institute, Stony Brook University, Stony Brook, NY 11794, USA.

* Corresponding author: Ruwen Qin, email: ruwen.qin@stonybrook.edu

Manuscript received MM DD, YYYY; revised MM DD, YYYY.

Therefore, it is an essential task to recognize risky traffic agents that will cause, or be involved in, accidents and localize these agents in the driving video captured by a vehicle-mounted dashboard camera (dashcam), a type of sensor that is both low-cost and widely deployed. Creating the ability to recognize and localize risky traffic agents will provide a valuable reference for the subsequent behavior and motion planning of ADAS. Besides, detecting the presence of risky traffic agents can help drivers reduce the chance of involving in accidents. This capability applies to other applications, such as traffic safety, autonomous driving, and pedestrian protection [4], [5].

Several computer vision studies have addressed a related task that detects anomalous events from a dashcam [6]–[11]. That is, this stream of literature focuses on identifying frames where the risk of a traffic accident is present. However, those studies did not address the critical challenge of localizing risky traffic agents in a driving scene.

Early localization of risky traffic agents, which are likely to involve in a future accident, from a highly dynamic and complex driving scene have three crucial challenges. First, the visual appearance of a traffic agent in the driving scene may not tell much about its risk level because of the very similar visual appearance of some traffic agents. Second, the time window for recognizing risky traffic agents is short. Last but not the least, the long-term temporal dependency of traffic agents underlying the accident risk is hard to capture. Several attempts have explored this topic, mainly using deep learning methods [12]–[16]. While these studies have laid a solid foundation for risky traffic agent localization, there are still several research needs that must be addressed.

One of the main research needs is the identification and utilization of features most relevant to risky traffic agents. Despite the various features used, including appearance, motion, size, and shape, a consensus on the most informative features has yet to be established. Thus, research is needed to identify these features and their contribution to the localization of risky traffic agents. Furthermore, how the selected features are fused would impact their ability to characterize these traffic agents. However, no such guidance exists, particularly for risky traffic agent localization. Another need is a dynamic attention mechanism for differentiating traffic agents. In complex traffic scenes, spatially distributed objects interact, and the interaction evolves. Incorporating the attention mechanism into the deep learning framework will improve the performance of risky traffic agent localization. Furthermore, the ability to localize

the presence of risky traffic agents ahead of time is crucial to avoiding accidents or mitigating their consequences. Current work has yet to create and measure such ability, and one reason for this is the need for the appropriate temporal annotation of risky traffic agents in existing datasets.

To tackle the challenges, the contributions of this paper are as follow:

- A new attention-guided multistream feature fusion network (AM-Net) is proposed for the early localization of risky traffic agents in driving videos. Scene- and object-specific spatio-temporal features are extracted and fused to better capture risk-related cues. A dynamic attention mechanism is further incorporated to selectively attend to critical cues. This framework has led to a new state-of-the-art performance.
- A new dataset, called the Risky Object Localization (ROL) dataset, has been developed to support training the model for the early localization of risky traffic agents and assessment of model performance in terms of correctness and earliness.
- A comprehensive study uses numerical experiments to verify the contribution of feature selection, feature fusion, and attention mechanism to the early localization of risky traffic agents.

The remainder of this paper will further the discussion by presenting the following contents in sequence. Section II summarizes the literature to determine state-of-the-art. After that, Section III delineates the proposed AM-Net, followed by the development of the ROL dataset in Section IV. Then, Section V presents experimental studies to demonstrate and verify the merits of the contributed network and dataset. In the end, Section VI summarizes research findings, limitations, and important future work.

## II. The Literature

This paper is built on studies that contribute to risky traffic agent localization, either directly or indirectly. The related literature is summarized below.

### A. Anomaly Detection in a Video

A topic related to risky traffic agent localization in a driving scene is video anomaly detection, which is about finding abnormal events in the video. Video anomaly detection is often formulated by profiling the normal behavior and measuring the spatial-temporal feature consistency. Deep learning-based methods can improve video anomaly detection by creating a more accurate normal video profile ( [15], [17]–[20]). For example, Hasan et al. [17] developed a 3D Convolutional feed-forward Auto-Encoder (ConvAE) to model regular video frames. Motivated by this, Chong et al. [18] used a Convolutional Long Short Term Memory Auto-Encoder (ConvLSTMAE) to simultaneously take advantage of both Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) in modeling normal appearance and motion patterns. Georgescu et al. [19] proposed a self-supervised and multitask learning approach for anomaly detection in videos, which can also be applied to risky traffic agent localization. Liu et al. [15] developed a Future Frame Prediction (FFP) method, and an observed deviation against the prediction indicates abnormality. In those studies, normal situations are usually stable scenes, which limits their applicability to driving videos captured by cameras in rapid motion.

### B. Traffic Accident Anticipation Using Dashcam Videos

Traffic accident anticipation in dashcam videos has become a research focus recently. Unlike video surveillance systems, dashcam videos capture moving traffic agents that not only rapidly move but appear and disappear quickly in the scene. Different advanced methods have been developed to learn the spatio-temporal patterns of traffic agents to provide an overall riskiness score for the scene, including an LSTM predictor [6], reinforced learning [8], graph neural network [11], generative adversarial network [21], and a dynamic attention network [9]. Although they only predict a risky event in the temporal domain, a risky event is always caused by risky traffic agents in the driving scene. As a result, these studies have established a strong methodological foundation for localizing risky traffic agents in the spatial domain.

### C. Risky Traffic Agent Localization in a Driving Scene

Several recent studies have put forward various approaches for the localization of risky traffic agents. Ohn-Bar et al. [22] developed a deep spatio-temporal importance prediction model that assigns riskiness scores to objects in a driving scene. Kozuka et al. [12] proposed a weakly supervised method for forecasting pedestrian-involved risky regions, whereas Zeng et al. [13] introduced a soft-attention mechanism to provide agent-centric riskiness scores for different traffic agents. An RNN has been employed explicitly to model nonlinear interactions among agents. However, these techniques have not made use of motion and location-scale features, which are essential for capturing the dynamic nature of traffic agents and their spatial relationships. Moreover, restricting the number of candidate objects in each frame is not a practical solution in many driving scenarios where the number of traffic agents can vary considerably.

Li et al. [14] investigated the causal impact of risky traffic agents on the driver's behavior by eliminating candidate traffic agents from the input video stream. Although it appears straightforward, it is an extremely complex problem to solve in real-world applications due to the presence of numerous known and unknown casualties. Kim et al. [23] introduced a domain adaptation technique to train a deep neural network for the identification of dangerous vehicles using synthetic data. Their approach is impressive in domain adaptation; however, their CNN-based method suffers from capturing long-term spatio-temporal relationships of accident-relevant cues. Malla et al. [24]'s DRAMA system performs joint risk localization and captioning in driving scenes. DRAMA employs a captioning approach to provide context for the identified risky traffic agents and their impact on the driving scene. Another line of research [25]–[27] formulated this problem as an important object identification problem by imitating human gaze behavior and predicting a pixel-level attention map that

serves as a proxy for risk. Nonetheless, human attention maps may not always be accurate for capturing risky agents.

Other methods, such as [16], [28], are trained to perform a related task such as trajectory prediction problem, where higher inconsistencies between the ground truth and the predicted trajectories are considered as risk. Specifically, Yao et al. [16] proposed a trajectory-based technique for localizing risky traffic agents, using future trajectory prediction to identify inconsistencies in agent behavior. Kim et al. [28] estimated agent importance by forecasting pixel-level attention heat maps. However, these methods are vulnerable when a traffic agent suddenly appears in the scene.

Although impressive, the methods outlined above exhibit some limitations in terms of identifying and capturing the most significant accident-relevant cues from video sequences to determine risky traffic agents. Furthermore, these approaches do not sufficiently explore the potential of feature fusion to enhance localization capability. Notably, existing work in the literature lacks the capacity to measure how early the models can localize risky traffic agents.

### D. Risky Object Localization Datasets

In response to the increased focus on deep learning-based anticipation of risky events in traffic videos, several large-scale datasets have emerged. Chan et al. [6] curated the Dashcam Accident Dataset (DAD), consisting of 620 video clips depicting on-road accidents, with the last ten frames of each video clip containing the accident. Yao et al. [29] developed the A3D dataset, comprising 1,500 videos annotated with the starting and ending times of accidents. Bao et al. [7] created a car accident dataset consisting of 1,500 videos, which includes annotations of environmental attributes and accident causes to support traffic accident anticipation. However, these datasets do not directly apply to the problem of localizing risky traffic agents due to the absence of object-level risk annotation.

You et al. [30] developed the CTA dataset, a benchmark dataset comprising 1,935 crash videos, which includes cause and effect events for different accidents, including their temporal intervals. This dataset contains labels for the cause of a crash and their potential effect after the crash. Fang et al. [31] collected driver attention on 2,000 crash videos to construct the DADA-2000 dataset, which demonstrated that driver attention can support future accident prediction problems.

In addition, Kim et al. [23] developed a synthetic dataset called GTA-crash, which was collected from the GTA5 game to reduce the cost of accident video collection. However, synthetic data may not capture the distribution of real driving scenes. Recently, Fang et al. [32] developed the largest dataset with 2.19 million video frames named CAP, which provides factual text descriptions before the accident and driver attention maps to support different transportation research problems.

Yao et al. [16] collected 4,600 videos and annotated risky traffic agents contributing to an accident with their bounding boxes in videos, when annotators judged the accident to be inevitable. It should be noted that the subjective nature of the annotation approach in many of these datasets may introduce some variability in the annotation quality. Additionally, the annotations in many of these datasets either do not include the beginning time of accidents or consider risky object appearance time as the accident or anomaly beginning time. However, the beginning time of accidents is necessary information to assess the earliness of a model's ability to localize risky traffic agents.

## III. Methodology

This paper proposes an Attention-Guided Multistream Network (AM-Net) to address identified gaps in current studies on risky traffic agent localization. Fig. 1 illustrates the model, which reads frames of an input video to output the riskiness score $s_{t,i}$ of traffic agent $i$ detected in frame $t$. Detailed description of the proposed AM-Net is below.

### A. Feature Extraction and Aggregation

AM-Net uses a pretrained object detector YOLOv5 [33] to detect traffic agents in each frame and provide the bounding boxes of the detected objects in the frame. $M$ is the number of detected objects, which may vary from one frame to another. Because the temporal association of the same object in successive frames is critical information for risky object localization, a multi-object tracker DeepSort [34] is used to make the association of objects in different frames.

The change in pixel-level motion among video frames is an important clue to find objects with unusual movement. Therefore, AM-Net uses a pretrained RAFT model [35] to extract the optical flow image for every frame. Then, regions in the flow image, determined by the detected objects' bounding boxes, are cropped to become object-level flow images. Then, the feature extractor ResNet50 [36] turns the frame-level flow image into a global feature vector, $\boldsymbol{g}_t(\in \mathbb{R}^D)$, and any of those at the object level into another feature vector, $\boldsymbol{o}_{t,i}(\in \mathbb{R}^D)$. Here, $D$ is the dimension of those feature vectors. Then, $\boldsymbol{o}_{t,i}$ is concatenated with $\boldsymbol{g}_t$ to become the overall flow feature vector for the $i$th object, $\boldsymbol{p}_{t,i}(\in \mathbb{R}^{2D})$:

$$\boldsymbol{p}_{t,i} = [\boldsymbol{o}_{t,i}; \boldsymbol{g}_t], \tag{1}$$

which captures the object's motion feature in the driving scene.

After further passing a fully connected layer $\phi$, the flow feature vector $\boldsymbol{p}_{t,i}$ becomes a lower dimensional feature vector, $\boldsymbol{f}_{t,i} \in (\mathbb{R}^{2d})$:

$$\boldsymbol{f}_{t,i} = \phi(\boldsymbol{p}_{t,i}; \boldsymbol{\theta}_0), \tag{2}$$

where $\boldsymbol{\theta}_0$ are learnable parameters of the fully connected layer.

Changes in bounding boxes' location and scale in successive frames capture the spatial dynamics of traffic agents over time. Therefore, for any object $i$ in frame $t$, its bounding box's location $(x_{t,i}, y_{t,i})$ and its scale in width $w_{t,i}$ and height $h_{t,i}$ are encoded as a feature vector $\boldsymbol{b}_{t,i}(\in \mathbb{R}^4)$:

$$\boldsymbol{b}_{t,i} = [x_{t,i}; y_{t,i}; w_{t,i}; h_{t,i}]. \tag{3}$$

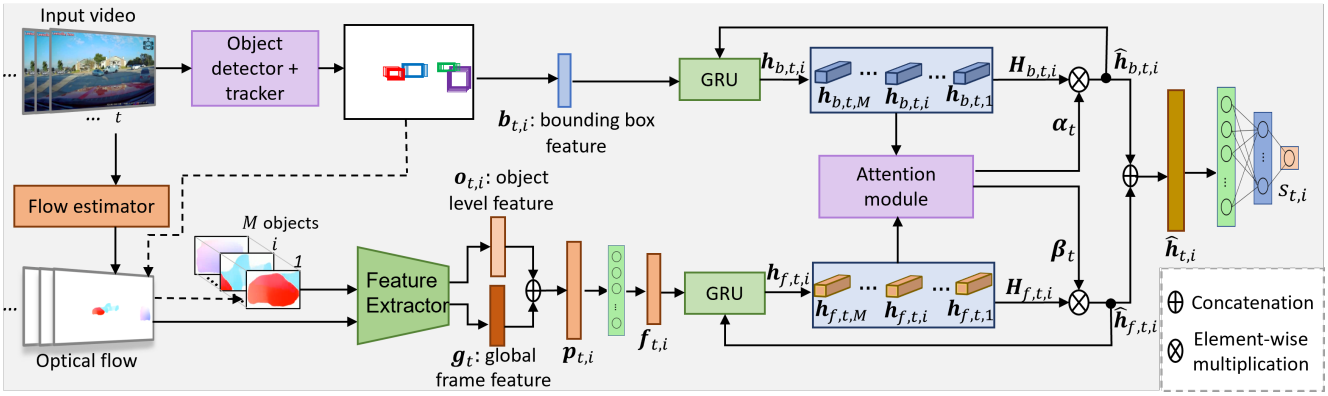



Fig. 1. Overview of the proposed AM-Net framework. The network has two branches: one for detecting and tracking bounding boxes using an object detector and tracker, and then encoding them with a GRU encoder. The other branch estimates optical flow, crops objects based on bounding box information, and extracts features for both the cropped objects and global frame. These two types of features are concatenated and fed into another GRU encoder to update the hidden representation. After that, the two GRUs are coupled with an attention module to get weighted hidden representations from the previous $M$ frames. Finally, concatenated hidden representations are passed through FC layers to produce riskiness scores for each object.

### B. Spatio-temporal Relational Learning with GRUs

Two Gated Recurrent Units (GRUs) respectively encode the extracted bounding box feature vector and flow feature vector of any detected object into their hidden representations and update them over time. Fig. 1 shows that, one GRU takes the bounding box feature vector of object $i$ in frame $t$, $\boldsymbol{b}_{t,i}$, and the weighted hidden representation of the same object in the last frame, $\widehat{\boldsymbol{h}}_{b,t-1,i}$, to update its hidden representation:

$$\boldsymbol{h}_{b,t,i} = GRU(\boldsymbol{b}_{t,i}, \widehat{\boldsymbol{h}}_{b,t-1,i}; \boldsymbol{\theta}_1), \tag{4}$$

where $\boldsymbol{\theta}_1$ are learnable parameters of the GRU, and $\boldsymbol{h}_{b,t,i} \in \mathbb{R}^{N_b}$. Here, $N_b$ is the number of hidden states of the GRU at the bounding box branch. In parallel, the second GRU takes the flow feature vector of the object, $\boldsymbol{f}_{t,i}$, and its weighted hidden representation in the past frame, $\widehat{\boldsymbol{h}}_{f,t-1,i}$, to update the hidden representation:

$$\boldsymbol{h}_{f,t,i} = GRU(\boldsymbol{f}_{t,i}, \widehat{\boldsymbol{h}}_{f,t-1,i}; \boldsymbol{\theta}_2), \tag{5}$$

where $\boldsymbol{\theta}_2$ are the learnable parameters, and $\boldsymbol{h}_{f,t,i} \in \mathbb{R}^{N_f}$. $N_f$ represents the number of hidden states of the GRU at the optical flow branch. The weighted hidden representations in eqs. (4) and (5) are introduced below.

### C. Attention Module

Objects in the driving scene have unequal contributions to a traffic accident. Therefore, learnable attentions should be distributed among the detected objects. Denote $\boldsymbol{H}_{b,t} \in \mathbb{R}^{N_b \times M}$ as the hidden representations of the $M$ objects' bounding box features in frame $t$:

$$\boldsymbol{H}_{b,t} = [\boldsymbol{h}_{b,t,1}, \ldots, \boldsymbol{h}_{b,t,i}, \ldots, \boldsymbol{h}_{b,t,M}]. \tag{6}$$

Here, the number of detected objects $M$ is a variable that can vary over time. Since the number of objects is not a fixed number, the spatio-temporal relational learning is not biased to unrelated features.

The dynamic spatial attention weights for the hidden representation of bounding box features, $\boldsymbol{\alpha}_{b,t} (\in \mathbb{R}^M)$, are computed as

$$\boldsymbol{\alpha}_{b,t} = \text{softmax}(\tanh(\boldsymbol{H}_{b,t}^{\mathrm{T}}) \boldsymbol{w}_b), \tag{7}$$

where $\boldsymbol{w}_b (\in \mathbb{R}^{N_b})$ are learnable parameters. Then, $\boldsymbol{\alpha}_{b,t}$ is used to turn $\boldsymbol{H}_{b,t}$ into a weighted aggregation, $\widehat{\boldsymbol{H}}_{b,t} (\in \mathbb{R}^{N_b \times M})$:

$$\widehat{\boldsymbol{H}}_{b,t} = [\widehat{\boldsymbol{h}}_{b,t,1}, \ldots, \widehat{\boldsymbol{h}}_{b,t,i}, \ldots, \widehat{\boldsymbol{h}}_{b,t,M}] = \boldsymbol{\alpha}_{b,t}^{\mathrm{T}} \odot \boldsymbol{H}_{b,t}. \tag{8}$$

where, $\odot$ represents the element-wise product.

The same attention mechanism is applied to the hidden representation of flow features,

$$\boldsymbol{H}_{f,t} = [\boldsymbol{h}_{f,t,1}, \ldots, \boldsymbol{h}_{f,t,i}, \ldots, \boldsymbol{h}_{f,t,M}]. \tag{9}$$

to get their flow feature attention weights $\boldsymbol{\alpha}_{f,t} (\in \mathbb{R}^M)$:

$$\boldsymbol{\alpha}_{f,t} = \text{softmax}(\tanh(\boldsymbol{H}_{f,t}^{\mathrm{T}}) \boldsymbol{w}_f), \tag{10}$$

where $\boldsymbol{w}_f (\in \mathbb{R}^{N_f})$ are learnable parameters of the attention module. $\boldsymbol{\alpha}_{f,t}$ is applied to $\boldsymbol{H}_{f,t}$ to obtain the weighted flow hidden representations:

$$\widehat{\boldsymbol{H}}_{f,t} = [\widehat{\boldsymbol{h}}_{f,t,1}, \ldots, \widehat{\boldsymbol{h}}_{f,t,i}, \ldots, \widehat{\boldsymbol{h}}_{f,t,M}] = \boldsymbol{\alpha}_{f,t}^{\mathrm{T}} \odot \boldsymbol{H}_{f,t}. \tag{11}$$

### D. Riskiness Score Prediction

As Fig. 1 shows, for any object $i$, the two attention-weighted hidden representations, $\widehat{\boldsymbol{h}}_{b,t,i}$ and $\widehat{\boldsymbol{h}}_{f,t,i}$, will respectively flow into the corresponding GRUs to update the object's hidden representations in the next frame $t+1$, as discussed in Section III-B. Meanwhile, they are concatenated to become the overall hidden representation of the object, $\widehat{\boldsymbol{h}}_{t,i} (\in \mathbb{R}^{N_b+N_f})$:

$$\widehat{\boldsymbol{h}}_{t,i} = [\widehat{\boldsymbol{h}}_{b,t,i}; \widehat{\boldsymbol{h}}_{f,t,i}]. \tag{12}$$

$\widehat{\boldsymbol{h}}_{t,i}$ is decoded by two fully-connected layers $\phi$ to output the scores of positive and negative classes, which are further normalized by the softmax operation to find the riskiness score of object $i$ in frame $t$, $s_{t,i}$,

$$s_{t,i} = \text{softmax}(\phi(\phi(\widehat{\boldsymbol{h}}_{t,i}; \boldsymbol{\theta}_3); \boldsymbol{\theta}_4), \tag{13}$$

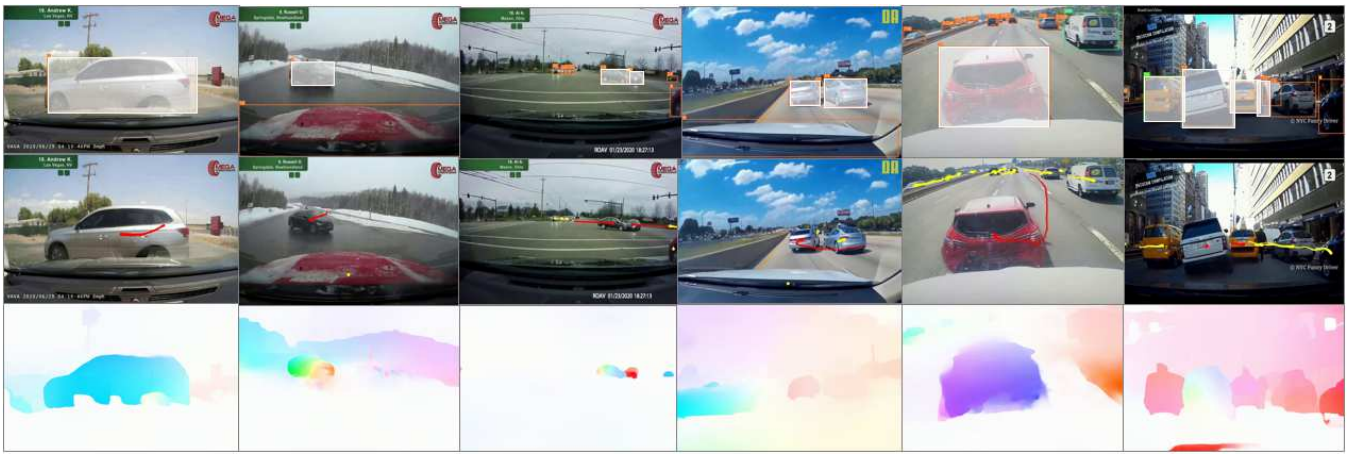


Fig. 2. Sample images from ROL Dataset. The first row shows the spatial annotations of risky traffic agents, in the form of white-shaded bounding boxes. The second row further visualizes traffic agent trajectories, with red curves for risky traffic agents and yellow ones for non-risky traffic agents. The third row are optical flow images of the sample frames.

which is the probability that the object will be involved in an accident soon, and $\boldsymbol{\theta}_3$' and $\boldsymbol{\theta}_4$' are the learnable parameters.

### E. The Loss Function

The goal of the model training process is to fit all the learnable parameters $\boldsymbol{\theta}$'s and $\boldsymbol{w}$'s of the proposed model in Sections III-A∼III-D by backpropagating a differentiable loss function. A training dataset contains videos that each has $T$ frames in total and $M$ objects in each frame. In frame $t$, the risk class of the $i$th object, is indicated by the ground truth label, $l_{t,i}$. $l_{t,i}$ is one if the object is a risky object (i.e., in the positive class) and zero (i.e., in the negative class) otherwise. In real-world driving scenarios, positive and negative classes are imbalanced. Therefore, a weighted cross entropy loss is calculated for each video in the training dataset:

$$\mathcal{L} = -\sum_{t=1}^{T}\sum_{i=1}^{M}\left[\lambda_p l_{t,i}\log(s_{t,i}) + \lambda_n(1-l_{t,i})\log(1-s_{t,i})\right], \tag{14}$$

where $\lambda_p$ and $\lambda_n$ are the weights for the positive and negative classes, respectively. Losses of all the training videos are summed up to get the total loss for optimizing the learnable parameters.

## IV. Dataset Development

This study introduces a new dataset named Risky Object Localization (ROL) dataset, which is available to the public [37]. ROL has 1,000 video clips. Each video clip has 100 frames and contains one or multiple risky traffic agents involved in an accident.

### A. Data Acquisition

To develop the ROL dataset, videos were collected from the crowd sourced online platform YouTube using query terms like "traffic accident" and "road crash". A long video retrieved from the platform often contains multiple accidents edited together. Consecutive frames of the same accident video segment have similar histograms of image pixels, which distinguishes them from other video segments. Using a threshold of 0.75, long videos were segregated into video clips based on their similarity measurement of frame histograms. These video clips may contain irrelevant portions, such as the part too early before an accident or post-accident portions. To address this, collected video clips were trimmed down to 5 seconds each with a frame rate of 20 frames per second (fps), resulting in 100 frames per video. The resolution of video frames is 1,080 $\times$ 720. The collected 1,000 videos were randomly divided into 800 training videos and 200 testing videos.

### B. Data Annotation

The dataset provides detailed annotations for each video clip, including object, accident, and scene information. Annotations are further categorized as either categorical, spatial, or temporal, offering comprehensive insights into the data. TABLE -I shows the hierarchy of the annotations in the ROL dataset.

The object-level annotations and the accident-level temporal annotation are necessary for developing the proposed AM-Net. Other annotations are provided to examine their impact to the AM-Net's performance. Details of the temporal, spatial and categorical annotations of the dataset are described below.

*1) Temporal Annotations:* Temporal information is annotated for both risky traffic agents and accidents in this study. This is because identifying risky traffic agents that may be involved in accidents in subsequent frames is critical for accident prevention. In contrast to the previous methods, such as [16], that rely on annotators' judgment to determine the moment an accident becomes inevitable, this study takes a more objective approach. Specifically, this study annotates the time when a risky traffic agent first appears in the driving scene as the risky object appearance time. The accident beginning time is defined as the moment when a vehicle collides with another static or moving object, which serves as a reference for calculating the earliness of localizing a risky traffic agent. It is worth noting that the accident beginning time of each video clip varies in the ROL dataset.

TABLE I
RISKY OBJECT LOCALIZATION (ROL) DATASET STRUCTURE.

| | Object-level | Accident-level | Scene-level |
|---|---|---|---|
| Categorical | - Risky object class<br>- Traffic agent class | - Ego-vehicle involvement<br>- Manner of collision | - Road system<br>- Related to roadway<br>- Intersection related<br>- Weather<br>- Lighting condition |
| Spatial | - Bounding box | | |
| Temporal | - Risky object appearance time | - Accident beginning time | |

*2) Spatial Annotations:* ROL provides bounding boxes as the spatial annotation of objects in a semi-automated approach that keeps human annotators in the loop. The object detector YOLOv5 [33] detects traffic agents and renders their bounding boxes. The multiobject tracker DeepSort [34] further associates the same object over successive frames as a tracklet and provides a unique tracking ID. By watching videos overlaid with tracking IDs, human annotators identify the tracking IDs of risky objects. Solely relying on the object detector and the object tracker to provide the object-level spatial annotation could miss a few agents, thus lowering the quality of the dataset. In resolving this problem, the paper develops a click-to-select object tracker based on DaSiamRPN [38] as an additional intervention. The tool requires only a few mouse clicks by human annotators to select missing or mis-tracked objects. Although it requires some effort from the annotators, human intervention ensures the best quality of the dataset. Fig. 2 illustrates six sample frames in ROL.

*3) Categorical Annotations:* ROL provides additional annotations of categorical variables, which are summarized in Table II. The traffic agent class is automatically annotated using YOLOv5, while other categorical variables are annotated by human annotators. The annotators also reviewed the automatically annotated traffic agents and re-annotated any mislabeled positive traffic agents. The accident-related and scene-related categorical variables follow the definitions provided by FIRST [39], although they were not required for developing the AM-Net. Nevertheless, they were included in the ROL dataset to broaden the dataset's applications.

TABLE III presents a comparison between the ROL dataset and other existing traffic accident datasets. It has been observed that several datasets, including DAD, A3D, and CCD, lack important components of accident-level and object-level annotations, such as the manner of collision and spatial locations of objects. The GTA dataset, although a synthetic dataset, offers the object-level annotation, thus enabling the localization of risky objects. However, it lacks accident- and scene-level annotations. On the other hand, the CTA dataset provides both accident- and object-level annotations, but the appearance time of risky objects is subject to the annotator's judgment. The DADA dataset is further extended to form a large CAP dataset. While the DoTA offers annotations at all levels, albeit with certain limitations. Specifically, DoTA lacks collision starting time (i.e., accident beginning time). Additionally, it is worth noting that the (*) marked levels differ from those in the proposed ROL dataset, as discussed in Section IV-B

TABLE II
SUMMARY OF CATEGORICAL VARIABLES AND DISTRIBUTIONS

| Variable | Categories | Dist(%) |
|---|---|---|
| Traffic agent class | Person | 3.0 |
| | Car | 84.1 |
| | Truck | 11.2 |
| | Bus | 0.9 |
| | Motorcycle | 0.5 |
| | Bicycle | 0.2 |
| Risky object class | Positive | 21.3 |
| | Negative | 78.7 |
| Frame class | Positive | 83.03 |
| | Negative | 16.97 |
| Ego vehicle involvement | Ego involved | 45.4 |
| | Non-ego involved | 54.6 |
| Road system | Local | 32.2 |
| | Arterial | 36.8 |
| | Interstate | 30.9 |
| Related to roadway | Roadway | 52.5 |
| | Shoulder & roadside | 12.3 |
| | Median | 2.4 |
| | Intersection related | 29.2 |
| | Others | 3.6 |
| Intersection related | No intersection | 70.8 |
| | 3-way | 5.3 |
| | 4-way | 20.1 |
| Weather | Clear | 76.6 |
| | Cloudy | 12.9 |
| | Rain | 8.5 |
| | Snow | 3.9 |
| Lighting condition | Day | 81.3 |
| | Night | 18.7 |
| Manner of collision | Angle | 36.4 |
| | Sideswipe | 11.6 |
| | Rear-end | 19.2 |
| | Head-on | 6.0 |
| | Others | 26.8 |

## V. RESULTS AND DISCUSSION

Experiments are conducted to verify the effectiveness of the proposed AM-Net and the newly collected ROL dataset. Implementation details, evaluation metrics, and the results are discussed below.

### A. Implementation Details

The proposed AM-Net is built using PyTorch [40]. Model training and testing are performed using an Nvidia Tesla V100 GPU with 32GB of memory. All the input frames are resized

TABLE III
COMPARISON BETWEEN ROL DATASET WITH OTHER DATASETS. INFORMATION OF OTHER DATASETS ARE OBTAINED FROM THEIR RELEASED SOURCES.

| Dataset | #Videos | # Frames | Accident-level | | Object-level | | | Scene-level |
|---|---|---|---|---|---|---|---|---|
| | | | Collision starting time | Manner of collision | Risky object appearance time | Spatial -negative object | Spatial -positive object | |
| DAD [6] | 620 | 62k | ✓ | | | | | |
| A3D [29] | 1,500 | 128k | ✓ | | | | | ✓ |
| CCD [7] | 1,500 | 75k | ✓ | | | | | ✓ |
| GTA [23] | 7,720 | 128k | | | ✓ | ✓ | ✓ | |
| CTA [30] | 1,935 | - | ✓ | ✓* | ✓ | | ✓ | |
| DADA [31] | 2,000 | 658k | ✓ | | | | ✓* | |
| CAP [32] | 11,727 | 2.19m | ✓ | ✓* | ✓* | | ✓ | ✓ |
| DoTA [16] | 4,677 | 732k | | ✓* | ✓* | ✓ | ✓ | ✓ |
| **ROL** | 1,000 | 100k | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

to 224×224 before feeding to the ResNet50 [36] feature extractor. Feature vectors $\boldsymbol{g}_{t,i}$ and $\boldsymbol{o}_{t,i}$ are obtained by applying an average pooling operation to the output of the ResNet50 feature extractor, and their dimension ($D$) is 2,048. A fully-connected layer further reduces the dimension to 256 ($d$). The dimension of flow hidden representations ($N_f$) is 256 and that for bounding box hidden representations ($N_b$) is 32. A learning rate of 0.001 is used to train the AM-Net on the newly developed ROL dataset, and ReduceLROnPlateau is used as the learning rate scheduler. Adam optimizer is used to optimize the network for 30 epochs and the best model is selected. The positive to negative class ratio in ROL is $0.27:1$. Therefore, the class weights for the negative class ($w_n$) and positive class ($w_p$) are selected as 0.27 and 1, respectively.

TABLE IV
ABLATION STUDY ON THE ROL DATASET.

| Model | RGB | | Bbox | Flow | | Att. | AUC | mTTA |
|---|---|---|---|---|---|---|---|---|
| | O | G | | O | G | | (%) | (s) |
| 1 | ✓ | ✓ | | | | | 80.30 | 2.36 |
| 2 | | | ✓ | | | | 81.19 | 2.22 |
| 3 | | | | ✓ | ✓ | | 82.07 | 2.02 |
| 4 | ✓ | ✓ | ✓ | | | | 83.21 | 2.11 |
| 5 | ✓ | ✓ | | ✓ | ✓ | | 81.87 | 2.13 |
| 6 | ✓ | ✓ | ✓ | ✓ | ✓ | | 84.54 | 1.95 |
| 7 | | | ✓ | ✓ | ✓ | | 84.53 | 1.92 |
| 8 | | | ✓ | ✓ | | | 84.14 | 1.89 |
| 9 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 84.91 | 2.19 |
| 10 | | | ✓ | ✓ | ✓ | ✓ | 85.59 | 2.18 |

O: object level feature; G: global frame level feature; Att: Attention; Bbox: Bounding box feature

### *B. Evaluation Metrics*

The model performance evaluation focuses on two aspects: the correctness in localizing risky traffic agents in videos and the earliness of risky traffic agent detection. To evaluate the correctness, this study uses Area under the Receiver Operating Characteristic Curve (AUC). AUC can measure the ability of AM-Net to differentiate risky and non-risky traffic agents.

To measure the earliness of the prediction, mean Time-to-Accident (mTTA) is used. Time-to-Accident (TTA) is defined as the first time when a riskiness score $s_{t,i}$ goes across a threshold value $\bar{s}$. That is,

$$\text{TTA} = \max\{\tau - t | s_{t,i} \geq \bar{s}, 0 \leq t \leq \tau, \forall\, i\}, \qquad (15)$$

where $\tau$ is the accident beginning time. TTA is dependent of the selection of a threshold value $\bar{s}$. mTTA, the average of TTA values at different threshold values, is calculated as an earliness metric independent of the threshold value selection.

### *C. Evaluation of the Proposed Model Architecture*

An ablation study is conducted to evaluate the contribution of different components of AM-Net. In addition to the bounding box and flow features described in Section-III, this ablation study includes the RGB appearance feature to assess its effect in localizing risky traffic agents. The RGB appearance feature is integrated into the AM-Net in the same manner as the flow feature. Ten models are trained and tested on ROL to compare the correctness and earliness metrics of AM-Net and nine variants. The results are summarized in TABLE IV.

**Single feature**: Model #1 from the table utilizes object-level (O) and global frame-level (G) appearance features that can be directly extracted from RGB images as the input for localizing risky traffic agents. This model achieves 80.30% AUC with 2.36 second mTTA, setting a benchmark for evaluating the proposed AM-Net (model #10) and its variants (models #2∼9). Model #2 uses the bounding box (Bbox) feature as the sole input, which increases AUC to 81.19% but reduces mTTA by 0.14 seconds. This suggests that changes in bounding box location and scale are more discriminative than appearance features for detecting risky traffic agents in traffic videos. Similarly, Model #3 uses flow features as the input, achieving 82.07% AUC with 2.02 second mTTA. The comparison between Models #1, #2, and #3 shows that flow features in standalone is stronger compared to the other two features in increasing the localization accuracy.

**Feature fusion**:Model #4, which combines the bounding box feature with appearance features, achieves a higher AUC than using any individual feature type (Models #1 and #2). However, it does not result in a longer mTTA. Similar results are observed for Model #5, which fuses the flow features with appearance features. When all three types of features are combined in Model #6, the AUC increases to 84.54%, but the mTTA decreases to 1.95 seconds. In general, Models #4∼#6 provide evidence that feature fusion is effective in enhancing localization accuracy.

Model #7 removes the appearance feature input from Model #6, yielding a similar result in terms of both AUC and mTTA.

Model#8 further removes the global frame-level flow feature from Model #7, resulting in lower AUC and mTTA. The global frame-level flow feature captures the motion of the traffic scene due to the ego-vehicle movement and the motion of other objects in the same frame, which are important clue for identifying risky traffic agents in the frame. The comparison of those four models (#6∼#8) suggests that fusing bounding box feature and flow features (both the object level and the global frame level) is the most suitable design of feature fusion. Appearance features, while are effective in other tasks, may lower the ability to differentiate traffic agents by their riskiness classes.

**Attention**: Model # 9 and Model #10 show the effectiveness of the proposed attention module. Model # 9 adds the attention module to Model # 6 where all the features (i.e. appearance, bbox, and flow) have been used. This model achieved 84.91 % AUC with a much longer mTTA of 2.19 second. Finally, Model #10 (i.e., the proposed AM-Net) adds the proposed attention module to Model #7, which achieves a promising combination of performance, 85.59% AUC and 2.18 seconds mTTA. In both Models, the attention module significantly improves the mTTA indicating that it effectively addresses the limitation of bounding box feature and flow features in localizing risky traffic agents earlier.

### *D. Comparison to State-of-the-Art Models*

This study compares the proposed AM-Net (i.e., #10 in TABLE IV) with existing models [15]–[18] on DoTA [16]. To make this comparison, the AM-Net model, which was initially trained on ROL, is fine-tuned using 300 randomly selected videos from the DoTA training dataset. Specifically, the final fully connected layers were fine-tuned for three epochs. The comparison is made from the perspective of video anomaly detection. That is, they are compared on the metric of frame-level AUC. TABLE V summarizes the comparative study results. The frame-level AUC values of models [15]–[18] are provided by [16]. To compute the frame-level AUC using the output of AM-Net, the highest riskiness score of any detected objects in a frame ($\max_i\{s_{t,i}\}$) is considered as the riskiness score for the frame. The Receiver Operating Characteristic (ROC) curve is attained by calculating the true positive rate and false positive rate at various detecting threshold values. Then, the frame-level AUC is calculated accordingly.

TABLE V
COMPARISON OF THE PROPOSED MODEL WITH EXISTING METHODS ON DOTA [16] TESTING DATASET FOR VIDEO ANOMALY DETECTION

| Method | Features | Frame-level AUC(%) |
|---|---|---|
| ConvAE [17] | Flow (G) | 66.3 |
| ConvLSTMAE [18] | Flow (G) | 62.5 |
| FFP [15] | RGB(G) | 67.5 |
| FOL [16] | RGB(G) + Bbox + Flow(O) + Ego | 73.0 |
| **AM-Net (Ours)** | Bbox + Flow (O+G) + Att. | **76.5** |

TABLE V shows that ConvAE and ConvLSTMAE, using the global frame-level flow feature as the only input, achieve 66.3% and 62.5% AUC, respectively. FFP, which uses the global frame-level appearance feature, achieves 67.5% AUC. FOL effectively boosts up the performance to 73.0% by fusing the bounding box, ego motion, and object-level optical flow feature with the global frame-level appearance feature. AM-Net achieves the highest AUC of 76.5%, about 3.5% higher than FOL. The improvement of AM-Net over FOL can be attributed to the choice of features and the method of video anomaly detection. The comparison also indicates that the bounding box and flow features, along with the attention mechanism, are the most important information that captures the spatio-temporal pattern of risky traffic agents. It is worth noting that the proposed attention mechanism significantly contributes to the early localization of risky traffic agents, as shown in the ablation study (Table-IV). However, the earliness of localization on DoTA is skipped in this comparison because DoTA's testing dataset lacks the annotation of accident beginning time. In this study, AM-net was also compared with the RGB-B method, proposed by Kim et al. [23], on the ROL dataset. The RGB-B method was both trained and tested on the ROL dataset, and yielded an AUC of 64.66% with 1.85 seconds of mTTA, while AM-Net outperforms it with an AUC of 85.59% and 2.18 seconds of mTTA.

TABLE VI
PERFORMANCE BY MANNER OF COLLISION: AM-NET VS. RGB-B [23]

| Manner | AUC (%) | | mTTA (s) | | TTA_0.3 (s) | |
|---|---|---|---|---|---|---|
| | AM-Net | RGB-B | AM-Net | RGB-B | AM-Net | RGB-B |
| Angle | 83.84 | 63.11 | 1.91 | 1.64 | 2.22 | 1.96 |
| Sideswipe | 91.17 | 68.80 | 2.20 | 2.68 | 2.98 | 3.05 |
| Rear-end | 90.85 | 67.52 | 2.34 | 1.98 | 3.38 | 2.85 |
| Head-on | 80.99 | 67.71 | 2.07 | 1.44 | 2.29 | 1.47 |
| Others | 83.47 | 62.72 | 2.85 | 2.12 | 3.04 | 1.93 |

### *E. Performance by Manner of Collision*

To evaluate the performance of the proposed AM-Net in localizing risky traffic agents involved in different manner of collisions, this study conducted a comparative study with the RGB-B method proposed by Kim et al. [23]. AUC, mTTA, TTA_0.3 are calculated for each collision type. TTA_0.3 refers to the TTA value when prediction threshold is 0.3. The testing dataset of ROL contains 34.0% angle, 18.5% sideswipe, 10.0% rear-end, 5.5% head-on, and 32.0% other types of collision videos.

Localizing risky traffic agents in sideswipe collisions is particularly challenging due to the short time for localization. However, AM-Net achieves the highest AUC of 91.17% in this category. In contrast, RGB-B approach only obtained an AUC of 68.80%. AM-Net also achieves a high AUC of 90.85% in rear-end collisions whereas RGB-B achieved 63.11% AUC. For angle collisions, AM-Net achieves AUC of 83.84%, while for head-on collisions, it achieves an AUC of 80.99%. It should be noted that the low AUC for detecting risky traffic agents in head-on collisions (80.99%) may be due to the rarity of these collisions in the collected dataset (6%). On testing videos of other collision types, AM-Net achieves 83.47% AUC. In contrast, in these three categories, AUC obtained by the RGB-B method is below 68%.

AM-Net achieves impressive early localization results, even in angle collisions where risky traffic agents can be localized

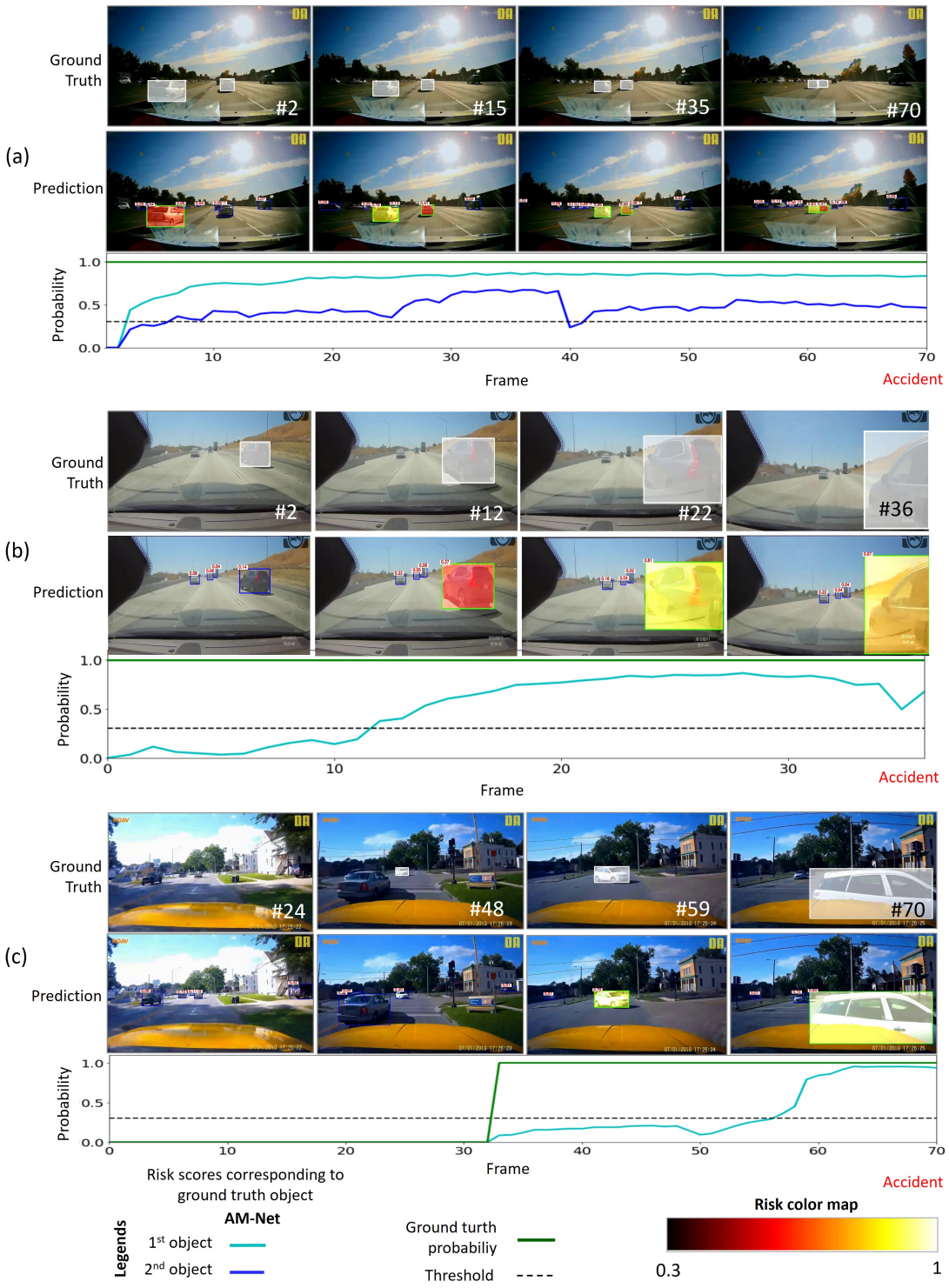



Fig. 3. Examples of risky traffic agent localization in driving scenes.

with an mTTA of only 1.91 seconds. This is notable given the sudden changes in motion and trajectory that characterise this type of collisions. For the rest of collision types, AM-Net is also effective at early localization, with lead times of at least 2 seconds prior to the accident. Notably, AM-Net outperforms the RGB-B method in terms of mTTA for all categories except sideswipe. The TTA_0.3 results show a similar pattern to the mTTA values for both AM-Net and RGB-B. In terms of the most effective early localization, AM-Net performs best for rear-end collisions, with an average lead time of 2.34 seconds before the accident occurs, when excluding the ”other” category.

The experimentation with respect to different manners of collision confirms that AM-Net can achieve promising perfor-

mance in localizing risky traffic agents, particularly in angle, sideswipe, and rear-end collision types. As per NHTSA report [39], 66% of reported collisions with body injury or property damage pertain to these three manners of collision. Correctly localizing risky traffic agents ahead of time would allow for taking preventive actions to avoid the accident.

### F. Performance by Ego Vehicle Involvement

In determining whether the proposed AM-Net performs differently on the early localization of risky traffic agents in ego vehicle-involved versus non-involved accidents, a test was conducted to compute the AUC and mTTA on both categories. AM-Net achieves a high AUC of 91.23% with a relatively short mTTA of 2.02 seconds on ego vehicle-involved accidents. For ego non-involved accidents, the proposed method achieves a slightly lower AUC of 83.78% and a longer mTTA of 2.79 seconds. It is worth noting that localization of risky objects in ego vehicle-involved accidents is less challenging in terms of prediction accuracy, as the objects appear relatively closer to the ego vehicle. In contrast, traffic agents appear relatively far in the third-party non-ego vehicle footage, making the accurate localization more difficult. Overall, the proposed AM-Net achieves very promising performance in both categories, indicating its potential to improve traffic safety.

### G. Performance in Negative Videos

To understand the effectiveness of the proposed method in distinguishing between risky and non-risky traffic agents in completely negative videos, a small scale experiment was conducted. Specifically, the experiment collected 60 negative videos from the DAD dataset, which were combined with 60 positive videos from the ROL dataset, to form a test dataset of 120 video. Then, AM-Net achieved an AUC of 87.04% on this dataset. Moreover, AM-Net achieved an AUC of 86.05% on the 60 positive videos of this dataset. These results confirm that proposed AM-Net has learned enough cues from the ROL dataset to differentiate between risky and non-risky traffic agents in both positive and negative videos.

### H. Qualitative Evaluation

Fig. 3 illustrates three representative examples of risky traffic agent localization in ROL dataset to demonstrate the effectiveness of the proposed AM-Net. These examples illustrate the scenarios of complex scenes with small objects, ego vehicle-involved accidents, and suddenly appearing risky traffic agents. Video demonstrations of the results are available at [37]. In each example, white-shaded bounding boxes highlight the true risky traffic agents. The green curve in the bottom row represents the true riskiness score of these agents, which is one as long as they appear in the frame. AM-Net localizes risky traffic agents with colored bounding boxes, and a threshold value of 0.3 is set to determine if a predicted object is a risky traffic agent for the illustration purpose. The reddish black to yellowish white color scheme of bounding boxes shown in Fig. 3 corresponds to the range [0.3, 1]. The riskiness score of the localized risky traffic agents is indicated by curves in colors other than green.

The video of example (a) contains two risky traffic agents, which are two vehicles far away from the ego vehicle. Since they appear in the video from the beginning, the ground truth riskiness score (i.e., the green line) remains constant at one throughout the timeline of the video sequence. The video contains a risky situation where the two vehicles are approaching each other from the side, resulting in a collision starting at frame #70. The accident scene is complex because the two risky traffic agents are very small due to the far distance from the ego vehicle. AM-Net successfully assigns the highest riskiness scores to those two traffic agents from a very early stage. The color of the first risky traffic agent's bounding box (in the middle row of the example) quickly turns from red to yellow, and the riskiness score (the cyan curve in the bottom row of the example) increases gradually. For the second risky traffic agent, its riskiness score (the dark blue curve) is a little lower than that of the first traffic agent. It is mainly because of the relatively lower velocity and relatively smaller size of the second vehicle than the first one. Upon reaching the 0.3 threshold value, the network successfully maintains the riskiness score above the threshold value almost throughout the timeline.

In example (b), the ego vehicle collides with another vehicle (the risky traffic agent) that cuts into its lane from the side at frame #36. In this example also, the risky traffic agent appears in the video from the beginning. AM-Net assigns the highest riskiness score to the correct vehicle in the traffic.

In example (c), the risky traffic agent is not in the video until frame #33 when the vehicle is coming to the front of the ego vehicle from the opposite direction, resulting in an angle collision at frame #70. Despite the risky traffic agent appearing briefly, AM-Net accurately assigns the highest riskiness score to this vehicle and did not generate any false positive detection.

## VI. Conclusion

This paper introduced a novel deep learning framework named Attention-guided Multistream feature fusion Network (AM-Net) for the early localization of risky traffic agents from dashcam videos. AM-Net extracts spatial and motion information and learns spatio-temporal features of traffic agents from successive frames. By fusing multistream features at object and global frame levels and differentiating attention to different agents, AM-Net effectively localizes risky agents before accident occurrences. An ablation study justified the input selection and verified the effectiveness of the proposed mechanism. Experimental evaluation showed that AM-Net outperforms state-of-the-art. The paper also introduced ROL, a benchmark dataset containing object-, accident-, and scene-level annotations, which can fuel multidisciplinary research on transportation safety enhancement.

**Limitations and future work:** Although the AM-Net has shown promising results in localizing risky traffic agents and anticipating traffic accidents, some limitations still need to be addressed in future research. One of the main limitations is the scarcity of data, which hinders the early localization

of risky traffic agents from head-on collisions. In addressing this issue, future research could explore incorporating prior knowledge from the dataset to allocate more attention to the front-facing vehicles approaching the ego vehicle. Another limitation is the ability of the object detector and tracker in adverse weather conditions, which can cause the tracker to reset and hinder the updating of the hidden representation based on observed trajectory. This problem could be addressed using more robust object detectors and trackers or multiple sensor fusion. In addition, AM-Net currently considers six main types of traffic agents, which may limit its ability to localize risky traffic agents in low-frequent cases where other types of objects, such as road debris or animals, initiate the risk. Future research could therefore expand the types of traffic agents the model considers to include animals and road debris to better anticipate potential accidents caused by such objects. Another promising avenue of the future research is predicting the crash severity level. Further investigation is required to predict more fine-grained severity level of different crash scenarios.

## ACKNOWLEDGMENT

Qin and Karim receive support from National Science Foundation (NSF) through the grant ECCS-#2026357. Yin and Karim receive support from NSF through ECCS-#2025929.

## REFERENCES

[1] L. Chen, Y. Li, C. Huang, B. Li, Y. Xing, D. Tian, L. Li, Z. Hu, X. Na, Z. Li, S. Teng, C. Lv, J. Wang, D. Cao, N. Zheng, and F.-Y. Wang, "Milestones in autonomous driving and intelligent vehicles: Survey of surveys," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1046–1056, 2023.

[2] California DMV, "Autonomous vehicle collision reports," 2023, available at:https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/autonomous-vehicle-collision-reports/, accessed March, 2023.

[3] WHO, "Global status report on road safety 2018: Summary," World Health Organization, Tech. Rep., 2018.

[4] U. Alvi, M. A. K. Khattak, B. Shabir, A. W. Malik, and S. R. Muhammad, "A comprehensive study on IoT based accident detection systems for smart vehicles," *IEEE Access*, vol. 8, pp. 122 480–122 497, 2020.

[5] M. Shirpour, N. Khairdoost, M. A. Bauer, and S. S. Beauchemin, "Traffic object detection and recognition based on the attentional visual field of drivers," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 594–604, 2023.

[6] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 136–153.

[7] W. Bao, Q. Yu, and Y. Kong, "Uncertainty-based traffic accident anticipation with spatio-temporal relational learning," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2682–2690.

[8] ——, "DRIVE: Deep reinforced accident anticipation with visual explanation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7619–7628.

[9] M. M. Karim, Y. Li, R. Qin, and Z. Yin, "A dynamic spatial-temporal attention network for early anticipation of traffic accidents," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 9590–9600, 2022.

[10] M. M. Karim, Y. Li, and R. Qin, "Toward explainable artificial intelligence for early anticipation of traffic accidents," *Transportation Research Record*, vol. 2676, no. 6, pp. 743–755, 2022.

[11] T. Wang, K. Chen, G. Chen, B. Li, Z. Li, Z. Liu, and C. Jiang, "GSC: A graph and spatio-temporal continuity based framework for accident anticipation," *IEEE Transactions on Intelligent Vehicles*, pp. 1–13, 2023.

[12] K. Kozuka and J. Carlos Niebles, "Risky region localization with point supervision," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 246–253.

[13] K.-H. Zeng, S.-H. Chou, F.-H. Chan, J. Carlos Niebles, and M. Sun, "Agent-centric risk assessment: Accident anticipation and risky region localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2222–2230.

[14] 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 *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 711–10 718.

[15] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–a new baseline," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6536–6545.

[16] Y. Yao, X. Wang, M. Xu, Z. Pu, Y. Wang, E. Atkins, and D. Crandall, "DoTA: Unsupervised detection of traffic anomaly in driving videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 444–459, 2023.

[17] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 733–742.

[18] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *International Symposium on Neural Networks*. Springer, 2017, pp. 189–196.

[19] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, "Anomaly detection in video via self-supervised and multi-task learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 742–12 752.

[20] S. Haresh, S. Kumar, M. Z. Zia, and Q.-H. Tran, "Towards anomaly detection in dashcam videos," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 1407–1414.

[21] Y. Qiu, T. Misu, and C. Busso, "Unsupervised scalable multimodal driving anomaly detection," *IEEE Transactions on Intelligent Vehicles*, 2022.

[22] E. Ohn-Bar and M. M. Trivedi, "Are all objects equal? deep spatio-temporal importance prediction in driving videos," *Pattern Recognition*, vol. 64, pp. 425–436, 2017.

[23] H. Kim, K. Lee, G. Hwang, and C. Suh, "Crash to not crash: Learn to identify dangerous vehicles using a simulator," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 978–985.

[24] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, and J. Li, "DRAMA: Joint risk localization and captioning in driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1043–1052.

[25] S. Alletto, A. Palazzi, F. Solera, S. Calderara, and R. Cucchiara, "DR(eye)VE: A dataset for attention-based tasks with applications to autonomous and assisted driving," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 54–60.

[26] A. Tawari, P. Mallela, and S. Martin, "Learning to attend to salient targets in driving videos using fully convolutional rnn," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 3225–3232.

[27] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney, "Predicting driver attention in critical situations," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*. Springer, 2019, pp. 658–674.

[28] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2942–2950.

[29] 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 *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9711–9717.

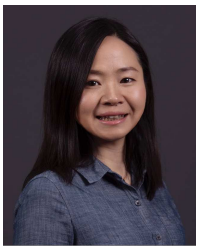
[30] T. You and B. Han, "Traffic accident benchmark for causality recognition," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 2020, pp. 540–556.

[31] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "DADA: Driver attention prediction in driving accident scenarios," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4959–4971, 2021.

[32] J. Fang, L.-L. Li, K. Yang, Z. Zheng, J. Xue, and T.-S. Chua, "Cognitive accident prediction in driving scenes: A multimodality benchmark," *arXiv preprint arXiv:2212.09381*, 2022.
[33] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.
[34] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3645–3649.
[35] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *European Conference on Computer Vision*. Springer, 2020, pp. 402–419.
[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[37] M. M. Karim, D. Racz, G. Liu, Z. Li, Y. Gong, W. Li, M. Incardona, R. Qin, and Z. Yin, "Risky object localization (ROL) in a driving scene dataset," 2022. [Online]. Available: https://github.com/monjurulkarim/ROL_Dataset
[38] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 101–117.
[39] NHTSA, "Fatality and injury reporting system tool (FIRST)," 2022, available at:https://cdan.nhtsa.gov/query, accessed July, 2022.
[40] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, pp. 8026–8037, 2019.

**Muhammad Monjurul Karim** received his B.Sc. degree in Industrial & Production Engineering from Bangladesh University of Engineering and Technology (BUET) and MS degree in Systems Engineering from Missouri University of Science and Technology. He is currently pursuing his Ph.D. in civil engineering concentrating on computer vision at Stony Brook University, NY. His research interest includes solving large-scale visual recognition and prediction problem by developing deep learning approaches.

**Zhaozheng Yin** (Senior Member, IEEE) is a SUNY Empire Innovation Associate Professor in the AI Institute, Department of Biomedical Informatics, and Department of Computer Science at Stony Brook University. He received his bachelor and master degrees from Tsinghua University and University of Wisconsin-Madison, respectively, and his PhD degree in computer science and engineering from Penn State, in 2009. He also worked as a postdoctoral fellow in the Robotics Institute of Carnegie Mellon University during 2009-2011. He has been working on Computer Vision and Machine Learning, with wide applications in Biomedical Image Analysis, Cyber-Physical Systems, Smart Manufacturing, Precision Agriculture, Civil Infrastructure Inspection, Transportation Systems Engineering, and Human Robot Collaboration. He is an associate editor of IEEE Transactions on Curcuits and Systems for Video Technology and Journal of Visual Communication and Image Representation. He has been an area chair of CVPR, ECCV, MICCAI and WACV.



**Ruwen Qin** (Member, IEEE) is an Associate Professor of Civil Engineering at Stony Brook University. She received her B.S. and M.S. degrees in Spacecraft Design from Beijing University of Aeronautics and Astronautics, and her Ph.D. degree in Industrial Engineering & Operations Research from Pennsylvania State University - University Park. Her current research focuses on machine learning methods for human and environmental sensing & learning in cyber-physical-human systems, smart connected systems, and intelligent automation systems. Her research has been sponsored by various funding agencies such as National Science Foundation.