IMPROVED DYNAMIC SPATIAL-TEMPORAL ATTENTION NETWORK FOR EARLY ANTICIPATION OF TRAFFIC ACCIDENTS

Chao Yi[1], Ting-Ji Huang[1], Han-Jia Ye[1], De-Chuan Zhan[1]**

[1] National Key Laboratory for Novel Software Technology, Nanjing University
lebron00821@gmail.com, htj2419@gmail.com, {yehj,zhandc}@nju.edu.cn

ABSTRACT

The proliferation of dashcams has significantly increased the volume of recorded data on road traffic, offering a unique opportunity to develop sophisticated algorithms that can analyze this data and predict potential accidents. In this paper, we introduces a new approach for predicting car accidents in videos by leveraging the Dynamic Spatial-temporal Attention Network (DSTA). We incorporate a frame-level loss and a bag-level loss into our method to aid in the model's learning process. Moreover, given that car crashes often involve continuous processes, we introduce soft labels to smoothen the label transitions in each video frame, thereby helping the model to more accurately identify the accident frame number and minimizing the impact of labeling noise. Additionally, we employ the output of the temporal self-attention aggregation (TSAA) module to enhance the prediction's robustness, which extracts information from all frames of the current video and to avoid interference from individual difficult frames. Our experimental results indicate that our Improved-DSTA (IDSTA) method outperforms the original DSTA method and performs exceptionally well in the AVA dataset. Overall, our proposed approach demonstrates significant potential in predicting car accidents from dashcam videos.

Index Terms— Anticipating Vehicle Accidents, Multi Instance Learning, Autonomous Vehicle

1. INTRODUCTION

Road accidents have been a persistent cause of injuries and fatalities worldwide, emphasizing the need for developing technologies that can effectively anticipate accidents and provide early warnings to drivers. In this regard, dashcams have emerged as a valuable tool that can record the road ahead and provide visual data for analyzing potential hazards and accident-prone situations.

In the realm of computer vision-based accident anticipation, initial studies employed traditional recurrent neural networks that utilize soft attention mechanisms to identify the causal factors present among agents in a given traffic scene [1, 2, 3, 4]. Recent advancements in computer vision and machine learning techniques have brought about significant progress in this field. Numerous studies have explored the use of dashcam videos for early accident anticipation, such as the work by [5], which proposes a Dynamic Spatial-Temporal Attention (DSTA) network for this purpose. In another study, [6] consider normal and anomalous videos as bags and video segments as instances in multip instance learning (MIL) and automatically learn a deep anomaly ranking model that predicts high anomaly scores for anomalous video segments. However, car accidents occur rapidly and may only occupy a few frames in a video, while the rest of the frames may be of little significance. To better distinguish car accidents, it is crucial to enhance the model's ability to capture information from specific frames before and after the event.

As mentioned above, the original frame-level loss of DSTA may not be suitable for the AVA dataset, as it lacks video segments during and after accidents, which makes it difficult to annotate the time of accident occurrence. To address this issue, we propose a novel approach that involves redesigning the frame-level loss by manually annotating the time point at which each car accident video in the training set shows the beginning of abnormal behavior. Furthermore, we introduce the use of soft labels to make label transitions in each frame of a video smoother, thereby reducing the model's training difficulty and minimizing the impact of labeling noise.

To alleviate the issue of imbalanced sample numbers across different categories, we apply weighting to the cross-entropy loss of normal and abnormal frames. Lastly, we design a bag-level loss to assist the model in learning, which takes the maximum value of the GRU outputs for all frames of each video as the anomaly score for that video. Experimental results demonstrate that using the output of the TSAA module with DSTA achieves higher accuracy than using the output of GRU as the prediction result.

In summary, our contributions are

- We propose a novel approach for predicting car accidents in dashcam videos using the Dynamic Spatial-temporal Attention Network, with a redesigned frame-level loss and a bag-level loss to facilitate the model's learning process.
- We introduce the use of soft labels to smoothen the label transitions in each video frame, thereby reducing

*Equal Corresponding Author

the model's training difficulty and minimizing the impact of labeling noise.

- We employ the output of the TSAA module to enhance the prediction's robustness. Experimental results indicate that the proposed IDSTA method outperforms the original DSTA method and performs exceptionally well in the AVA dataset.

We will begin by describing the data collection process and the methodology we will use to develop our algorithms. We will also present the results of our experiments and evaluate the effectiveness of our approach.

2. RELATED WORKS

Anticipating traffic accidents from dashcam videos is a complex problem that requires predicting the likelihood of a future event. While computer vision-based approaches for human action anticipation often rely on appearance features, such as object, activity, and context cues [7, 8, 9, 10, 11], these methods are typically applied to video data captured by static surveillance cameras and may not be suitable for mobile cameras mounted on vehicles. In recent years, various approaches have been proposed to tackle this challenge by leveraging dashcam videos to predict traffic accidents. For instance, Suzuki et al. [1] proposed an adaptive loss function that promotes early anticipation of accidents by assigning penalty coefficients based on the achieved mean time-to-accident during training. Additionally, Yao et al. [12] utilized ego-vehicle motion information to develop an unsupervised approach that predicts the future locations of traffic agents. Likewise, Takimoto et al. [13] integrated physical location data with video data to predict traffic accidents. In another study, [6] considered normal and anomalous videos as bags and video segments as instances in multip instance learning (MIL), automatically learning a deep anomaly ranking model that predicts high anomaly scores for anomalous video segments.

Attention mechanisms have gained considerable attention in recent times due to their ability to focus on relevant features, similar to how humans selectively concentrate on important aspects. Neural networks equipped with attention units can identify and highlight relevant features for accurate predictions. While Chan et al. [4] introduced dynamic soft-attention to traffic accident anticipation, Zeng et al. [2] proposed a soft-attention RNN that models the interaction between traffic agents and identifies risky regions. Additionally, Fatima et al. [3] introduced a feature aggregation block that captures inter-object interactions. Nonetheless, these methods mainly concentrate on learning attentions to spatially distributed agents linked to accidents, paying less attention to the temporal importance of appearance features. Cui et al. [14] integrated both a spatial attention module and a temporal attention module with a GRU [15] to estimate the state-ofhealth of batteries. T Bao et al. [16] used a graph convolutional recurrent neural network (GCRNN) to capture the spatial temporal relations among candidate objects.

Recently, Muhammad et al.[5] proposed a Dynamic Spatial-Temporal Attention (DSTA) network that predicts the probability of a future accident. It selects discriminative temporal segments using a Dynamic Temporal Attention (DTA) module and informative spatial regions using a Dynamic Spatial Attention (DSA) module. However, the original frame-level loss of DSTA may not be suitable for identifying certain accidents well, as it lacks video segments during and after accidents, making it difficult to annotate the time of accident occurrence. To address this, it is necessary to improve the model's ability to capture information from specific frames before and after the event.

3. IMPROVED-DSTA

3.1. Predicting Accidents from Dashcam Video

The problem of predicting accidents from dashcam video is a critical task in the field of computer vision and driving safety. The goal is to develop a machine learning model that can analyze the video feed from a dashboard camera mounted on a vehicle and accurately identify the occurrence of an accident.

The inputs to the model consist of a sequence of video frames captured by the dashcam, usually at a fixed frame rate. Each frame is represented as a matrix of pixel values, with dimensions HxWxC, where H, W, and C denote the height, width, and number of color channels of the image, respectively. The input sequence can vary in length depending on the duration of the driving footage and the frequency of accidents.

The output of the model is the prediction of the accident occurrence and the corresponding frame number t where the accident happens. The accident occurrence is represented as a binary value, where 1 indicates an accident, and 0 indicates no accident. The frame number is an integer value that identifies the specific frame where the accident occurs.

The purpose of this problem is to improve driving safety by providing real-time accident detection and alert systems. Early detection of accidents can help reduce the severity of injuries and prevent further damage to the vehicle or surrounding environment.

In mathematical terms, let X be the input video sequence, where $X = [x_1, x_2, ..., x_n]$ is a sequence of video frames, with xi representing the i^{th} frame. Let Y be the output label, where $Y = [y_1, y_2, ..., y_n]$ is a sequence of binary labels, with $y_i = 1$ indicating the presence of an accident in the corresponding frame x_i, and $y_i = 0$ otherwise. The model learns a mapping function $f(X) \rightarrow Y$ that can accurately predict the occurrence of accidents in the video sequence X.

3.2. Feature Extraction

We performed object detection on images using the pre-trained cascade RCNN network provided by DSTA[5], and retained the top (19-k) bounding boxes with the highest predicted confidence. We used the coordinates of the k bounding boxes provided in the AVA dataset, as well as the (19-k) bounding boxes selected as candidate regions, for a total of 19 bounding boxes. We then cropped the corresponding regions from the image based on the bounding box coordinates. Subsequently, we used a pre-trained VGG-16 network to extract feature vectors from the original image and the cropped regions.

Given that in the task of predicting car accidents, the key objects in the images are generally various types of vehicles, we utilized a pre-trained VGG-16 model on cars-196 to extract features, which is different from some previous works that used pre-trained models on imagenet to extract features. We found that using the pre-trained VGG-16 model on cars-196 for feature extraction achieves better performance than using pre-trained models on imagenet.

3.3. Model Architecture

Our network architecture utilizes DSTA. DSTA extracts spatial information from a single image through a Dynamic Spatial Attention module(DSA) and captures the relationship between different frames in a video through a Dynamic Temporal Attention module(DTA). The future probability of accidents is predicted using a gated recurrent unit(GRU). Due to its excellent ability to extract both spatial and temporal information, DSTA has achieved outstanding performance on datasets such as DAD and CCD. Therefore, we employ DSTA as our network architecture.

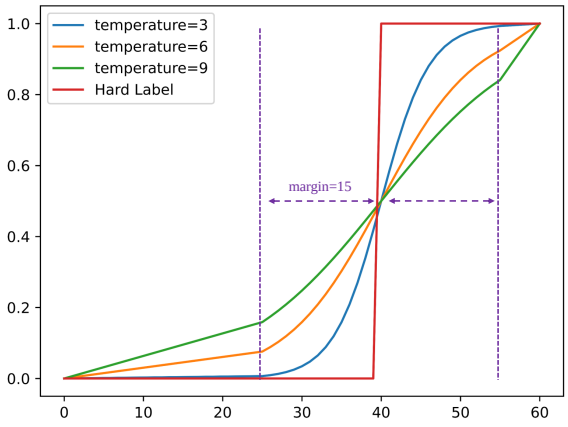

Fig. 1. Comparison of Soft Label and Hard Label. Temperature Controls how soft the label is.

3.4. Soft Label

The AVA dataset contains only video clips before the occurrence of accidents, and does not include video segments during and after the accident, which is different from datasets such as DAD and CCD. Therefore, we cannot annotate the time of accident occurrence (TOA) for each video segment similar to the DAD method (otherwise, TOA would be the last frame of the video), and calculate the frame-level loss based on this. In addition, the original frame-level loss of DSTA decreases as the accident prediction probability of any frame in the accident video increases, that is, this loss hopes that all frames in the accident video have a high accident prediction probability. However, in an accident video, abnormal segments usually only account for a small part of the video. Therefore, the design of the original DSTA loss is not suitable for the AVA dataset.

Based on the characteristics of the AVA task, we have redesigned the frame-level loss. We manually annotated the time point $\hat{t}$ at which each car accident video in the training set shows the beginning of abnormal behavior. The rule for selecting the annotated time point is that the annotator can perceive the abnormal behavior in the current frame at that time. After annotation, we can divide the video into normal and abnormal segments based on $\hat{t}$, and assign a label of 0 (indicating no car accident) to the frames in the normal segment, and a label of 1 (indicating a car accident) to the frames in the abnormal segment, thus obtaining frame-level binary classification label information. Then, we calculate the cross-entropy loss between the output of each frame's GRU and the binary classification label of the current frame, and take the average as the final frame-level loss.

However, due to the subjective factors of the annotators, the labeling of the moment t in exceptional cases can be considered noisy in the annotated dataset. Moreover, the change between adjacent frames of a video is often small. The approach of roughly assigning a label of 0 to frame $\hat{t}-1$ and a label of 1 to frame $\hat{t}$ would result in completely opposite labels between similar frames, thereby increasing the impact of labeling noise. To solve this problem, we propose using Soft Labels to make the label transition of each frame in the video smoother, reducing the difficulty of model training and minimizing the influence of dataset noise.

Formula (1) shows the mathematical expression of soft label we used. In the formula, m represents margin, T represents Temperature coefficient, which controls the soft degree of soft label, and x represents the label of current frame. Figure 1 plots hard Lables and Soft labels with different temperatures.

$$y = \begin{cases} \frac{\mathrm{softmax(m\,/\,T)}}{\hat{\mathrm{t}}-\mathrm{m}} \cdot x, & 0 < x \leq \hat{\mathrm{t}}-\mathrm{m} \\ \mathrm{softmax}\left[\frac{x-\hat{\mathrm{t}}}{\mathrm{T}}\right], & \hat{\mathrm{t}}-\mathrm{m} < x < \hat{\mathrm{t}}+\mathrm{m} \\ \frac{[\mathrm{softmax(m\,/\,T)}]\cdot(x-60)-x+\hat{\mathrm{t}}+\mathrm{m}}{\hat{\mathrm{t}}+\mathrm{m}-60}, & \hat{\mathrm{t}}+\mathrm{m} \leq x \leq 1 \end{cases} \tag{1}$$

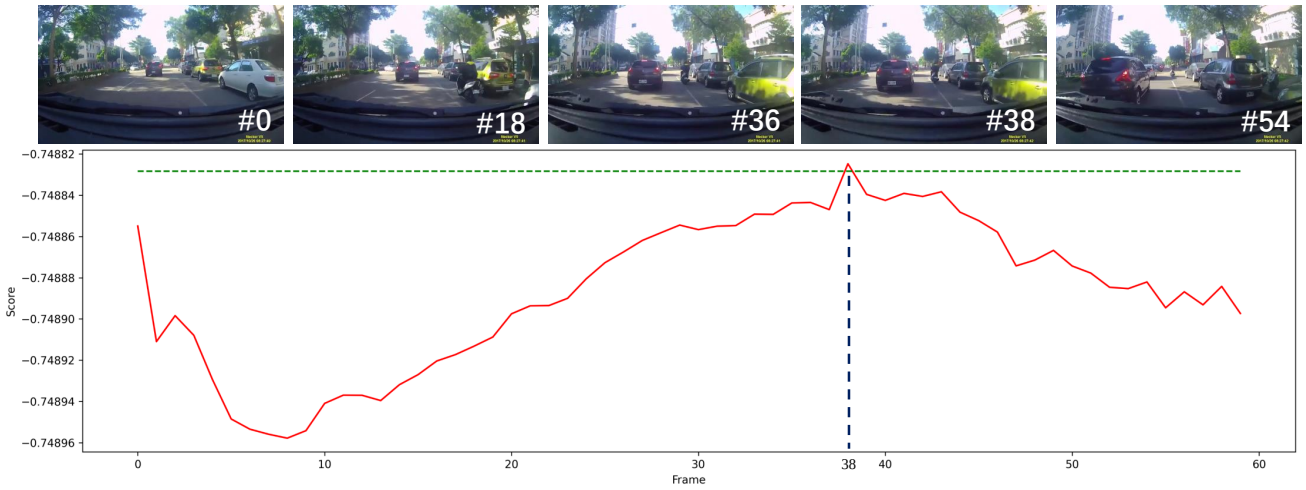

Fig. 2. Prediction with output of GRU may be subject to some outlier.

Considering a car accident video, the abnormal segments typically only account for a small portion of the video, leading to the issue of imbalanced sample numbers across different categories. To alleviate this problem, we applied weighting to the cross-entropy (CE) loss of normal and abnormal frames, assigning a greater weight to the CE loss of abnormal frames.

3.5. Multi Instance Learning Loss

As mentioned earlier, frame-level loss is noisy, and if only frame-level loss is used for training, it may still affect the performance of the model. However, multi-instance learning only requires bag-level labels in the training data, i.e., it does not require frame-level labels. As some previous works[6] have described, a video can be seen as a bag, and each frame or a segment of the video can be seen as an instance, so the original car accident prediction problem can be naturally regarded as a binary classification problem of multi instance learning. Therefore, we designed a bag-level loss to assist the model in learning. Specifically, in the training process, we take N videos from the current batch and divide them into two categories: normal videos and car accident videos. Then we take the maximum value of the GRU outputs for all frames of each video as the anomaly score for that video. Afterwards, we calculate two types of gap loss: (1) overall gap loss: calculate the gap between the average anomaly scores of normal videos and car accident videos in the current batch; (2) hard example gap loss: calculate the gap between the maximum anomaly score of normal videos and the minimum anomaly score of car accident videos in the current batch. We hope to increase these gap to enhance the model's ability to distinguish difficult samples.

The formula of this loss is shown in (2), where $\boldsymbol{O} \in R^{N \times F}$ represents the GRU output anomaly score of all F frames of N videos in a batch.

$$L_{gap}(\boldsymbol{O}) = \max\left(0, 1 - \min\left(\{O_i\}_{y_i=1}\right) + \max\left(\{O_i\}_{y_i=0}\right)\right) + \max\left(0, 1 - \frac{\sum_{i,y_i=1}(O_i)}{n_p} + \frac{\sum_{i,y_i=0}(O_i)}{n_n}\right), \quad (2)$$

In which $O_i = \max_j\left(\boldsymbol{O_{ij}}\right)$.

Meanwhile, we also partition the video segments outside the soft label margin into normal and abnormal segments based on the annotated abnormal time point $\hat{t}$. Then, we calculate the frame-level gap loss by taking the minimum value of the abnormal segments and the maximum value of the normal segments.

$$L_{gap\_frame}(\boldsymbol{O}) = \sum_{i,y_i=1} \max\left(0, 1 - \max\left(\{\boldsymbol{O_{ij}}\}_{j \geq t_i+m}\right) + \max\left(\{\boldsymbol{O_{ij}}\}_{j<t_i-m}\right)\right) \quad (3)$$

In which $\hat{t}_i$ represents the annotated abnormal time point of ith video.

3.6. TSAA Prediction

In our experiments, we found that using the output of the TSAA module with DSTA as the prediction result achieves higher accuracy than using the output of GRU as the prediction result, which differs from the original DSTA method that uses GRU for prediction. One possible reason we analyzed is that the original DSTA method uses GRU to output the probability of a future car accident for each frame in a video, and when the probability of a certain frame exceeds a threshold, the video is considered to have a car accident in the future. This method is susceptible to interference from individual difficult frames, resulting in incorrect results. However, DSTA's

TSAA module extracts information from all frames of the current video through an attention module to predict whether a car accident will occur in the future, thus possessing better robustness. Figure 2 shows an example where TSAA correctly predicted a negative sample and GRU incorrectly predicted a positive sample. Only some frames in the sample have high abnormal scores, and GRUs tend to produce false results in such cases.

4. RESULTS

We calculated the F1-score and test accuracy on our self-labeled test set to compare the performance of different methods precisely. As Shown in Table 1, the experimental results demonstrate that our IDSTA method outperforms the original DSTA method.

Table 1. Performance on AVA test dataset.

	freeway	
Model	F1-Score	Accuracy
DSTA[5]	0.504	0.545
IDSTA	**0.614**	**0.630**
IDSTA (without soft label)	0.513	0.551
IDSTA (without gap loss)	0.526	0.584
IDSTA (without gap-frame loss)	0.480	0.558
	road	
DSTA[5]	0.438	0.623
IDSTA	**0.612**	**0.701**
IDSTA (without soft label)	0.518	0.623
IDSTA (without gap loss)	0.536	0.623
IDSTA (without gap-frame loss)	0.523	0.630

5. CONCLUSION

In this work, we proposed the Improved-DSTA (IDSTA) method for car accident anticipation, which effectively distinguishes between abnormal and normal frames in a car accident video. To overcome the challenges posed by the rapid occurrence of car accidents and the limited number of abnormal frames, we introduced a redesigned frame-level loss and bag-level loss, along with soft labels, to help the model's learning process. Our contributions enhance the model's ability to capture information from specific frames before and after the event, resulting in a more robust and effective approach for car accident anticipation. The experimental results on the AVA dataset demonstrate that our proposed IDSTA method outperforms the original DSTA method and achieves high accuracy. Our approach holds the potential to be applied to other datasets and contribute towards the development of better accident anticipation systems.

6. ACKNOWLEDGMENTS

This work is supported by National Key R&D Program of China (2022ZD0114805), NSFC (61773198, 61921006, 62006112), Collaborative Innovation Center of Novel Software Technology and Industrialization, NSF of Jiangsu Province (BK20200313).

7. REFERENCES

[1] Tomoyuki Suzuki, Hirokatsu Kataoka, Yoshimitsu Aoki, and Yutaka Satoh, "Anticipating traffic accidents with adaptive loss and large-scale incident db," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3521–3529.

[2] Kuo-Hao Zeng, Shih-Han Chou, Fu-Hsiang Chan, Juan Carlos Niebles, and Min Sun, "Agent-centric risk assessment: Accident anticipation and risky region localization," in *Proceedings of the 17th IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, 2017, pp. 2222–2230.

[3] Mishal Fatima, Muhammad Umar Karim Khan, and Chong-Min Kyung, "Global feature aggregation for accident anticipation," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 2809–2816.

[4] Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun, "Anticipating accidents in dashcam videos," in *Proceedings of the 13th Asian Conference on Computer Vision*, Taipei, Taiwan, 2016, pp. 136–153.

[5] Muhammad Monjurul Karim, Yu Li, Ruwen Qin, and Zhaozheng 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.

[6] Waqas Sultani, Chen Chen, and Mubarak Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the 18th IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, US.

[7] Raúl Quintero Mínguez, Ignacio Parra Alonso, David Fernández-Llorca, and Miguel Ángel Sotelo, "Pedestrian path, pose, and intention prediction through gaussian process dynamical models and pedestrian activity recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 5, pp. 1803–1814, 2018.

[8] Mingze Xu, Mingfei Gao, Yi-Ting Chen, Larry S Davis, and David J Crandall, "Temporal recurrent networks for online action detection," in *Proceedings of the International Conference on Computer Vision*, Seoul, Korea, 2019, pp. 5532–5541.

[9] Hema S Koppula and Ashutosh Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE Transactions on Pattern Anal-*

ysis and Machine Intelligence, vol. 38, no. 1, pp. 14–29, 2015.

[10] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba, "Anticipating visual representations from unlabeled video," in *Proceedings of the 16th IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 98–106.

[11] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei, "Peeking into the future: Predicting future person activities and locations in videos," in *Proceedings of the 19th Conference on Computer Vision and Pattern Recognition*, Long Beach, California, USA, 2019, pp. 5725–5734.

[12] Yu Yao, Mingze Xu, Yuchen Wang, David J Crandall, and Ella M Atkins, "Unsupervised traffic accident detection in first-person videos," in *Proceedings of the 32th International Conference on Intelligent Robots and Systems*, Macau, China, 2019, IEEE, pp. 273–280.

[13] Yoshiaki Takimoto, Yusuke Tanaka, Takeshi Kurashima, Shuhei Yamamoto, Maya Okawa, and Hiroyuki Toda, "Predicting traffic accidents with event recorder data," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Prediction of Human Mobility*, 2019, pp. 11–14.

[14] Shengmin Cui and Inwhee Joe, "A dynamic spatial-temporal attention-based gru model with healthy features for state-of-health estimation of lithium-ion batteries," *IEEE Access*, vol. 9, pp. 27374–27388, 2021.

[15] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning*, 2014.

[16] Wentao Bao, Qi Yu, and Yu 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.