International Journal of Computational Methods
© World Scientific Publishing Company

A Data-driven Approach for Traffic Near-miss Anticipation

Xuehuai Shi
School of Computer Science and Engineering, Nanjing University of Science and Technology,
No. 200 Xiaolingwei street, Xuanwu District
Nanjing, 210094, China shixuehuaireal@163.com
http://<https://sxhsine.github.io/Biography/>

Yong Qi *
School of Computer Science and Engineering, Nanjing University of Science and Technology,
No. 200 Xiaolingwei street, Xuanwu District
Nanjing, 210094, China 790815561@qq.com

Qia Wang
School of Computer Science and Engineering, Nanjing University of Science and Technology,
No. 200 Xiaolingwei street, Xuanwu District
Nanjing, 210094, China 1747058726@qq.com

Received (26 October 2018)
Revised (Day Month Year)

We propose a data-driven approach for traffic near-miss anticipation (Figure 11). Through (1) positive-favored loss for early anticipation (PF-LEA), (2) traffic near-miss feature portraits generation model (TNFP-GM), (3) early anticipation model based on QRNN-Inception and (4) a method for optimizing TNFP-GM and QRNN-Inception simultaneously, our proposed approach outperforms some conventional methods on several databases. Additionally, we supplement more crowd-sourced dash-cam videos to DAD for better understanding scenes in early anticipation. In our experimental results, our method can achieve (+6.0%@MAP, +55.8frames@ATTC), (+2.7%@MAP, +15.8frames@ATTC), (+1.8%@MAP, +6.5frames@ATTC) in comparison with several conventional brilliant approaches on DAD, and achieve (55.9%@MAP, 86frames@ATTC), (53.8%@MAP, 90.3frames@ATTC) on SDAD-1 and SDAD-2.

Keywords: Incidents Anticipation; Deep Learning.

1. Introduction

Self-driving cars and advanced driver assistance system (ADAS) have achieved great success due to the development of computer science and traffic science in both academic and industrial fields. In autopilot technology, the paramount objective is to

*corresponding author

deliver the guests in the car to a destination safely, which shows the importance of understanding traffic near-miss scenes. However, to date, traffic near-miss scenes (the scenes that lead vehicles to the dangerous situations that accidents may happen) are not valued, such cars have few opportunities to learn these scenes.

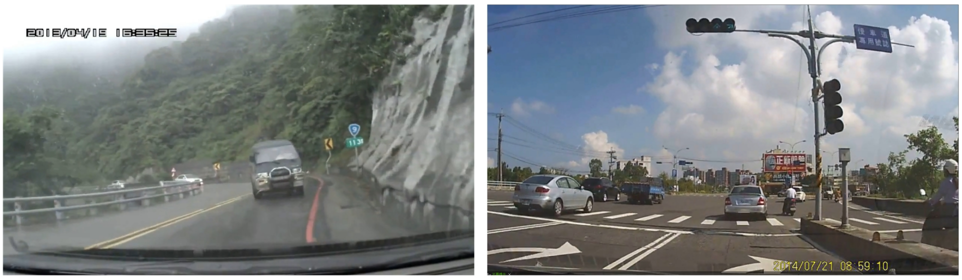
Traffic near-miss incidents are described as the dangerous situations that the final protective measures are threatened and the collisions had been narrowly (See Figure 1) avoided among road users. In this paper, we divided the type of traffic near-miss incidents into two types (an incident with/without own vehicle). For anticipating traffic near-miss incidents, the key challenge is how the vehicles in road users can figure out the near-miss incidents and tide over the situations safely. The vital technologies for achieving this target include environment sensing and anticipation in videos. For the technology of environmental sensing, problems relate to object detection and semantic segmentation. For example, visual simultaneous localization and mapping (vSLAM) and 3D induction detection and ranging are hot issues in many races for self-driving cars. For the technology of anticipation in videos, problems relate to ambiguous event determination and early event detection. For example, among the test scenarios of self-driving cars, pedestrian avoidance, signal intersection, non-motorized crossing are the most adaptive scenarios.

In this paper, we propose a data-driven approach for anticipating traffic near-miss incidents based on the dashcam videos (see Figure 11). The key criterions to evaluate the traffic anticipation models are timeliness and accuracy. As the result of our contributions, we found that our approach can achieve (+6.0%@MAP, 55.8frames@ATTC), (+2.7%@MAP, 15.8frames@ATTC), (+1.8%@MAP, 6.5frames@ATTC) in comparison with several conventional brilliant approaches on DAD, and achieve (55.9%@MAP, 82.3frames@ATTC), (53.7%@MAP, 90.3frames@ATTC) on SDAD-1 and SDAD-2.

In summary, the contributions of our study are as follows:

Technical contribution: We propose a data-driven approach to anticipate traffic near-miss incidents. Firstly, we detect road users in dashcam videos based on faster R-CNN [1]. Secondly, we combine the quasi-recurrent neural network [2] with Inception [3] named QRNN-Inception for improving predictive power. For further advancing QRNN-Inception performance, we employ a Positive-favored method which allows QRNN-Inception to change penalty weights adaptively as the near-miss scenario (positive scenario) is less than the other. Thirdly, we propose an ensemble approach based on QRNN-Inception and the traffic near-miss feature portraits generation model (TNFP-GM) inspired by CAM [4]. After the initialization of QRNN-Inception and TNFP-GM, the ensemble approach would optimize QRNN-Inception based on the result of object detection and TNFP-GM, then tune the framework based on the outcome of object detection and QRNN-Inception. Finally, QRNN-Inception explores the risk-index, and TNFP-GM draws the near-miss feature portraits based on the dashcam videos. The results of the data-driven approach for anticipating traffic near-miss incidents show that it produces a better performance with the state-of-art.

Database contribution: VSLab has introduced the crowd-sourced dashcam videos dataset [5], we divided the traffic near-miss scenarios into the incidents with own vehicle (the collisions happened between the Dashcam owner and the other road users) and without own vehicle (the crashes happened between the other road users without the Dashcam owner) based on the brilliant work of VSLab. Then we tag each frame in the dashcam videos into positive frames and negative frames. The results of experiments in two categories indicate the model shows more predictive power in the database with own vehicle than the database without own vehicle.



(a) A Traffic Near-miss Incident with Own Vehicle

(b) A Traffic Near-miss Incident without Own Vehicle

Fig. 1. Traffic Near-miss Incidents with/without Own Vehicle

2. Related Work

Since our work focus on traffic near-miss incidents anticipation, it is closely related to scene segmentation and anticipation in videos. Herein, we first discuss related work of scene segmentation in computer vision. Then, we motion the recent works of anticipation in videos. Finally, we detail different large-scale traffic near-miss datasets.

2.1. *Scene Segmentation*

Scene segmentation is a fundamental issue in computer vision because it is the basis of many practical applications. According to different ascension formation, the topic can be divided into two parts. One part is to propose a framework for scene segmentation; the other part is to extract features for scene segmentation. As well, the topic can be categorized into two directions as the data type includes RGB, grayscale images and RGBD.

In the part of **the framework for scene segmentation**: For **analyzing RGB**, several standard segmentation frameworks have been proposed in [6, 7, 8, 9]. In [10, 11, 12], a symmetric attention mechanism has been employed for object locations detection in the environment. For **analyzing RGBD**, [13, 14, 15, 16, 17, 18, 19, 20] discover a new aspect for semantic scene segmentation. For example, a batch of RGBD images can be merged into one single point and be labeled with Markov Field (MRF) in [20]. In [16], SIFT features and 3D locations are merged into a Conditional Random Field (CRF). By applying depth gradient and spin normal descriptors, [13] details more predictive power. [19] proposed a CNN for feature extraction from RGBD images.

In the part of **feature extraction for scene segmentation**: To date, one direction of the topic is to **apply contextual modeling** to enhance scene segmentation [21, 22, 23, 24, 25]. For multi-scale context information extraction, [21] proposed an atrous spatial pyramid pooling (ASPP). For long-range context information extraction, [22, 23, 24] proposed RNN-based models. For multi-region information, [25] proposed multiple pooling. The other direction is to **aggregate in a multi-scale way** due to single-scale features cant perform robust segmentation [26, 27, 28, 29]. For corresponding features fusion, [26, 27] employed a multi-resolution input approach. For aggregating results, [28] employed multi-scale patches. [29] employed a new method for improving the performance of recurrent convolutional neural networks by images of different size afferent different layers.

2.2. *Anticipation in Videos*

Anticipation in videos is a challenging work in computer vision as the future event will be predicted from ambiguous information. There are many vital issues in the field of video anticipation, such as danger detection [5, 30, 31], trajectory prediction [32, 32, 33], and motion planning [34], etc., we only discuss early event anticipation and novel loss functions.

As one of the most critical issues, **early event anticipation** has been widely studied [37,38,39,40,41,42]. For early action detection, a probabilistic model was employed in [35]. For real-time action prediction, [36] constructed a traffic-specified database. For the traffic context information extraction, the transitional action was defined in [37]. For more predictive power in early action detection, [38] employed a spatial attention mechanism. A method using CRF with information on human poses and object coordinates was applied in [39], [40] trained CNN to extract feature for action anticipation in a self-supervised manner. The other critical issue in video anticipation is the **novel loss functions** for better predictive power. In contrast to the widely studied early event detection problem, such work focuses on developing unique loss functions to improve the model predictive power [38, 41, 42] or to promote the generalization performance of the model in the early event detection such as future appearance [43] and future features [40]. The positive-favored loss function proposed in this paper focus on generalization performance of the model as well, which allows QRNN-Inception to change penalty weights adaptively as the near-miss scenario (positive scenario) is less than the other.

2.3. *Traffic Databases*

There are several new databases for traffic near-miss anticipation, such as KITTI [44], Dashcam Accident Dataset (DAD) [5], and Near-miss Incident Database (NIDB) [30].

KITTI is a vision benchmark dataset, which supports multiple autopilot tasks such as multi-object detection, environment segmentation, visual ranging, and real-time tracking, etc. The videos in KITTI were collected from urban roads, highways and suburban roads in multiple cities. **NIDB** is a traffic near-miss database which includes a quantity of traffic incident videos. There are several classes in NIDB, includes low/high risk for bicycles, pedestrians, and vehicles, as well as a background class. **DAD** is one of the first crowd-sourced datasets for anticipating accidents, containing accidental events on the collected data.

However, many near-miss incidents in these databases happened between the others (e.g., the other two vehicles collided without own vehicle). Thus, we divided the videos in DAD into two categories: near-miss incidents with own vehicle and without own vehicle. We future tag each frame in the DAD into positive frames and negative frames for better learning performance on self-driving cars in traffic near-miss situations.

3. Overview

Figure 11 provides an overview of our work in traffic near-miss anticipation. Our input is a video with 100 frames; each frame is an RGB image which shape is (height, width)=(720,1280). In each frame, the model will output the risk index of the current moment from the early anticipation model named QRNN-Inception and the traffic near-miss feature portraits from TNFP-GM.

Our approach works in a data-driven manner. The present moment risk index is learned from the result of scene segmentation of the original video and the result of the traffic near-miss feature portraits from TNFP-GM. Similarly, the traffic near-miss feature portraits are learned from the result of scene segmentation of the original video and benefit from the output of current moment risk index if the early anticipation model works well. We propose a framework to optimize TNFP-GM and QRNN-Inception simultaneously which we detail in segment 4. The structure consists of three key components: (1) scene segmentation of the traffic near- miss videos, (2) traffic near-miss incidents anticipation and (3) traffic near-miss feature portraits generation. In algorithm 1, we summarize the pipeline of the framework for traffic near-miss incidents anticipation.

3.1. *Scene Segmentation of the Traffic Near-miss Videos*

For the promptness in real-time traffic near-miss incidents anticipation, we propose a scene segmentation model based on fast R-CNN ([1]). It takes each frame in the video and a set of object proposals as input and outputs a set of object classes

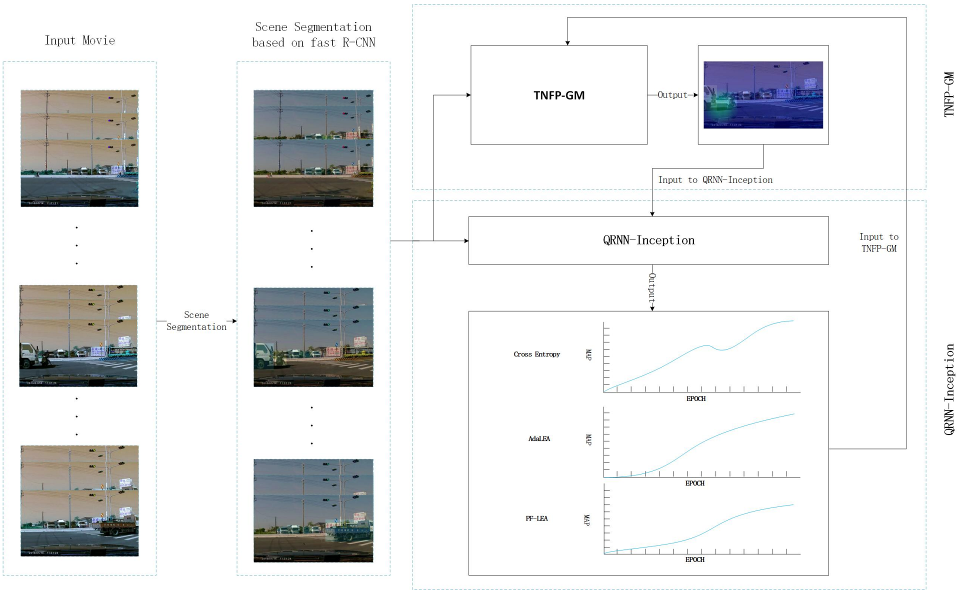


Fig. 2. Pipeline of Traffic Near-miss Incidents Anticipation

and the corresponding four values ((r,c,h,w) that specifies its top-left corner (r,c) and its height and width (h,w)). Figure 3 provides the architecture of this section. Firstly, it produces a convolutional feature map through several convolutional layers and max-pooling layers. Secondly, it generates a fixed length feature vector by the hyper-parameters H and W using the ROI pooling layer for each object proposal. Finally, the probabilities of the set of object classes and the corresponding four values are produced by two sibling output layers after a sequence of fully connected layers (FCs).

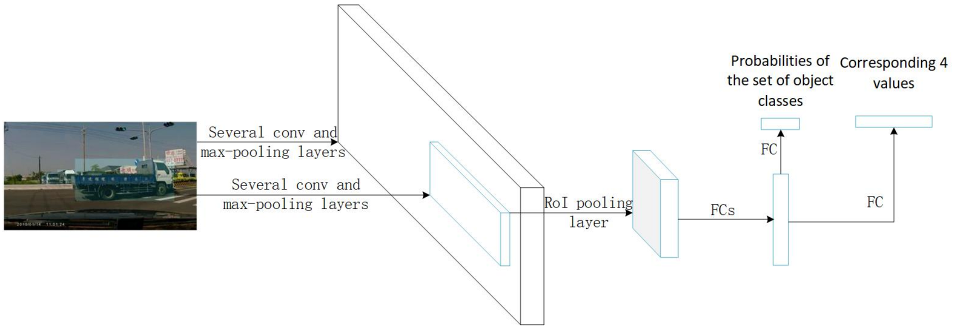

Several conv and max-pooling layers
Several conv and max-pooling layers
RoI pooling layer
FCs
FC
FC
Probabilities of the set of object classes
Corresponding 4 values

Fig. 3. Pipeline of Scene Segmentation in Traffic Near-miss Anticipation

Algorithm 1 Framework of the Data-driven Approach for Traffic Near-miss Anticipation $TNA(D)$

```
Require:
    $D = \{x^{(n)}, y^{(n)}\}_{n=1}^{N}$: dataset with $N$ traffic near-miss examples
    $(x^{(i)} = \{x_1^{(i)}, ..., x_{100}^{(i)}\}, y^{(i)} = \{y_1^{(i)}, ..., y_{100}^{(i)}\})$
    $s_f(D)$: the Scene Segmentation function
    $s_c(D)$: the Convolutional Feature Maps (CFMs) function of Scene Segmentation function
    $Q(input)$: the Early Anticipation Model based on QRNN-Inception
    $G(input)$: the Traffic Near-miss Feature Portraits Generation Model
    $warmup(model)$: the model convergence testing function
    $C$: the number of epoches for training $Q$ and $G$
Ensure:
    Function $TNA(D)$
    init $s_f(D)$
    $s_{f_w} \leftarrow s_f$
    while $s_{f_w}$ is not $warmup(s_{f_w})$ do
        $min\ loss_D(s_{f_w})$
    end while
    from $s_{f_w}$ get $s_{c_w}$
    init $Q, G$
    $Q, G \leftarrow loss_{s_{f_w}, s_{c_w}, D, C}(Q, G)$
    return $Q, G$
```

3.2. *Traffic Near-miss Incidents Anticipation*

In this section, we propose an early anticipation model named QRNN-Inception based on Quasi-recurrent neural networks [2] and inception architecture [3]. It takes the result of scene segmentation model for each frame in the video and the output of TNFP-GM as input and outputs the risk index of the current moment. Figure 4 details the architecture of this section. Instead of several convolutional and fo-pool layers in QRNN, we employ inception architecture for the information extraction in traffic near-miss incidents anticipation model, and the inception unit outputs three sibling branches into the fo-pooling layer. Inception is applied independently to the input as in standard googLeNet. Then we get the hidden state and the current state from the fo-pool layer, and the FCs producing softmax probability estimates the risk index of the present moment using current state.

3.3. *Traffic Near-miss Feature Portraits Generation*

For the traffic near-miss feature portraits generation model (TNFP-GM), we care more about the size and relative location of the detected objects, generate a set of weights for attention drawing. We would like to ignore the influence of relative

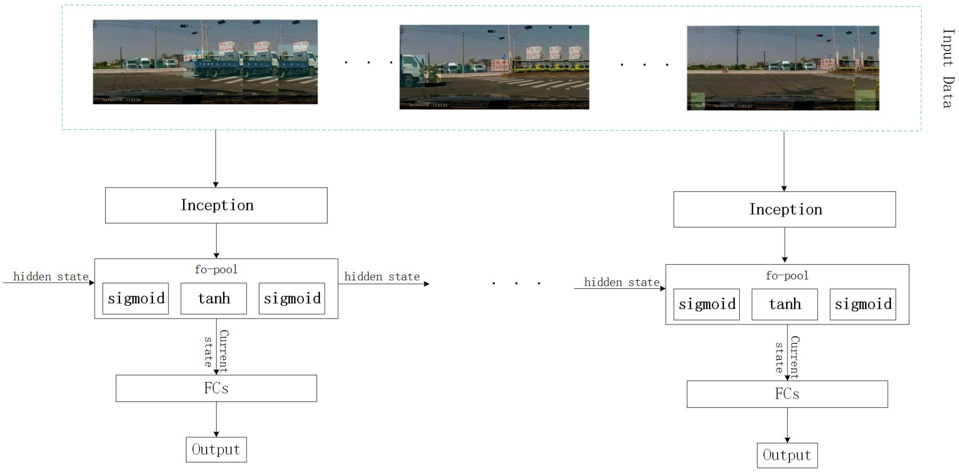


Fig. 4. Architecture of QRNN-Inception in Traffic Near-miss Anticipation

location overlapping at the beginning and slowly changing over time as the image input is an RGB. As the promptness must be taken into account for traffic near-miss anticipation, the architecture of this section is simple. Most work of this section is based on class activation maps [4] and recurrent neural network. The architecture of this section is shown in Figure 5. We intercept the convolutional feature maps from the output of scene segmentation, which is a tensor includes information about the location and size of several objects. We perform global average pooling on the convolutional feature maps and use those as features for a fully-connected layer that produces the desired output. The model will be trained using softmax loss.

4. Loss Function

The target of the proposed traffic near-miss incidents anticipation model is to output the risk rate at each frame that represents probability an accident will occur in the future. Section 3 details the overview of the framework, we explain the loss function of the framework and the details of each section in the aspect of optimization in this section, which is one of the major contributions of this study.

4.1. *Overall Loss Function*

For accident anticipation, we need to optimize the anticipation model and the portraits generation model simultaneously, overall loss function shown in Formula 1:

$$loss_{Q,G} \leftarrow loss_{s_f(D),G}(Q) + loss_{s_c(D),Q}(G) \tag{1}$$

Q is the shortened form of QRNN-Inception and G is TNFP-GM for short, $D = \{x^{(n)}, y^{(n)}\}_{n=1}^{N}$ is dataset with N examples($(x^{(i)} = \{x_1^{(i)}, ..., x_{100}^{(i)}\}, y^{(i)} =$

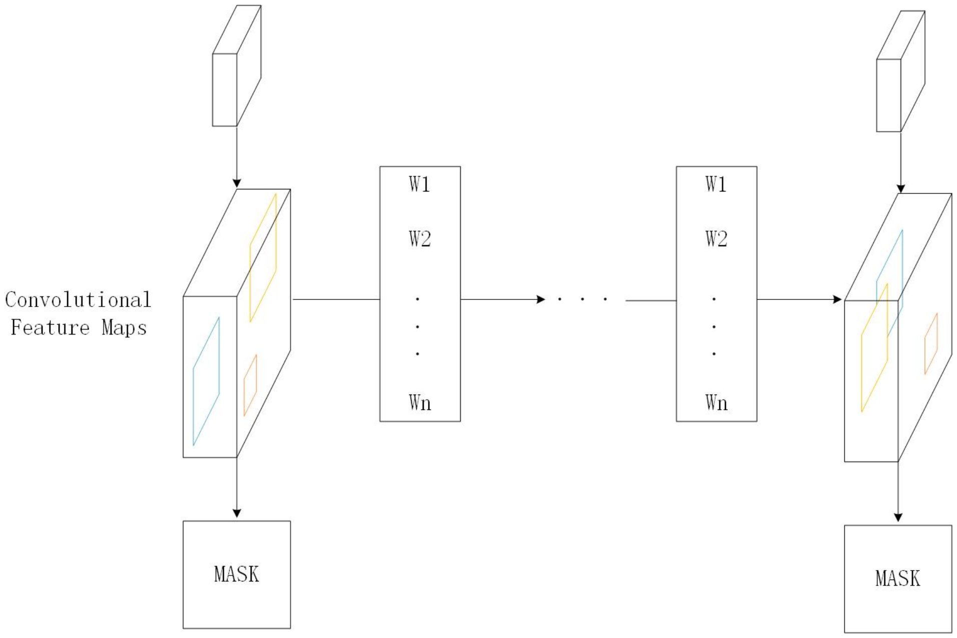


Fig. 5. Architecture of TNFP-GM in Traffic Near-miss Anticipation

$\{y_1^{(i)}, ..., y_{100}^{(i)}\}))$, $s_f(D)$ represents the Scene Segmentation function, $s_c(D)$ is the Convolutional Feature Maps (CFMs) function from Scene Segmentation.

For warming up Q, we first set the output of G as a blank mask combine with the result of $scenesegmentation$ as the input of Q, the loss function of this part is shown in Formula 2:

$$min_Q loss_{s_f(D), G \leftarrow blank}(Q) \tag{2}$$

After warming up Q, we need to warming up G using the mature $Q_w \leftarrow argmin_Q(loss_{s_f(D), blank}(Q))$:

$$min_G loss_{s_c(D), Q_w}(G) \tag{3}$$

Then, we individually optimize Q and G with the fixed parameters in each other until convergence (e.g., fixing the parameters in Q to optimize G), details shown in algorithm 2, the details of loss function in D are discussed in section 4.2.

4.2. *Loss Function for QRNN-Inception*

In this section, we would like to discuss the loss function in the QRNN-Inception. As we know the standard of the anticipation model is risk rate output should be the-earlier-the-better as the accuracy of the model is guaranteed. To pursue the standard, our strategy is to neutralize the penalty as the positive frames should get more attention, while the number of positive frames is much less than the negative frames in traffic near-miss incidents database. Whats more, we should increase the punishment cost for those near-miss incidents which are not predicted correctly in time before accidents happened than the near-miss incidents without crashes occurred are not predicted accurately.

Our proposed loss function is based on AdaLEA [30]. In AdaLEA, the losses are divided into two parts, the negative sample which is the standard cross-entropy, the penalties in positive samples are gradually increased as the frame closer to the accident. However, it ignores the fact that the positive frames are much less than the negative frames, which should get more attention. The details of AdaLEA are shown below:

Adaptive Loss for early anticipation:

$$\begin{cases} L_{AdaLEA}(r_t) \;=\; -\sum_{t=1}^{T}\{\alpha y_t log(r_t) + \;(1-y_t)log(1-r_t)\} \\ \\ for\ positive: \\ L^{p}_{AdaLEA}(r_t) \;=\; \sum_{t=1}^{T} -\alpha log(r_t) \\ \\ for\ negative: \\ L^{n}_{AdaLEA}(r_t) \;=\; \sum_{t=1}^{T} -log(1-r_t) \end{cases} \tag{4}$$

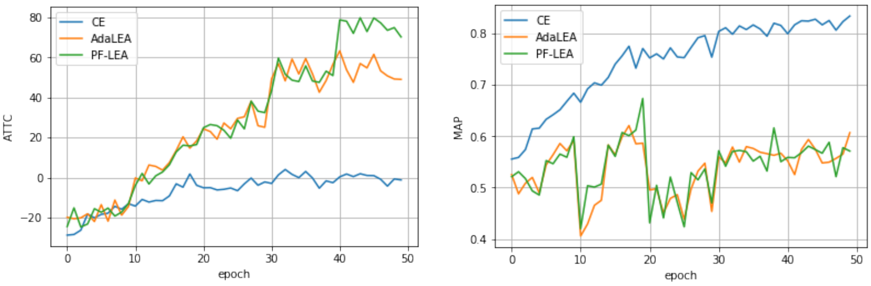
In AdaLEA, T is the starting frame of the annotated accident, $d = Tt$ indicates the time to T from current moment t. $r(t)$ is the risk rate range from 0 to 1. Current epoch is remarked as e, $\Phi(\cdot)$ is a function which represents an ATTC at a training epoch, F is the frame rate of videos, and γ is a hyperparameter.

AdaLEA can adaptively modify the weight value depending on how early the model can anticipate a traffic accident at each learning epoch, but it ignores the bias between positive frames and negative frames, and the importance of the positive frames after the crash happened in the training process of traffic accident anticipation model. Based on the rule of these facts, we propose Positive-Favored Loss for early anticipation (PF-LEA), details are shown below:

Positive-Favored Loss for early anticipation (here, define $L\{X,Y\} = L_{PF-LEA}\{X,Y\}$):

$$\begin{cases} L\{X,Y\} \;=\; \sum_{n=1}^{N} L^{p}(x^{(i)},y^{(i)}) \;+\; L^{n}(x^{(i)},y^{(i)}) \\ \\ for\ positive: \\ L^{p}\{x^{(i)},y^{(i)}\} \;=\; -\{\sum_{t=1}^{T} \alpha y_t^{(i)} \ln(r_t^{(i)}) + \;\sum_{t=T}^{T_{stop}} h(t) \ln(r_t^{(i)})\} \\ \alpha \;=\; exp^{(-max(0,d-F\times\Phi(e-1)-\gamma))} \\ h(t) \;=\; min(A,(t-T)^2-1) \\ \\ for\ negative: \\ L^{n}\{x^{(i)},y^{(i)}\} = \sum_{t=1}^{T} \beta(1-y_t^{(i)}) \ln(1-r_t^{(i)}) \\ \beta \;=\; \frac{n_1}{n_2} \end{cases} \tag{5}$$

In PF-LEA, A and γ are hyperparameters, n_1, n_2 are the number of positve/negative frames from a input video $\{x^{(i)}, y^{(i)}\} \in D$, while $\{x^{(i)}, y^{(i)}\} = \{(x_1^{(i)}, y_1^{(i)}), ..., (x_{100}^{(i)}, y_{100}^{(i)})\}$, the definition of T, $d = Tt$, $r(t)$, e, $\Phi(\cdot)$, F are the same as in AdaLEA. **Figure 6** shows the details about the convergence of accuracy and ATTC (the average time of predict frames to the accident frames) in three different loss functions (**lossFun-accuracy-epoch, lossFun-ATTC-epoch**), it indicates that the anticipation model with PF-LEA can anticipate earlier than the other two loss functions without sacrificing too much accuracy.



(a) ATTC v.s. epoch

(b) MAP v.s. epoch

Fig. 6. Performance of three different functions (Cross entropy, Adaptive loss for early anticipation and positive-favored loss for early anticipation) on DAD.(a) ATTC v.s. epoch on DAD with three different loss functions. (b) MAP v.s. epoch on DAD with three different functions.

5. Experiments

In this section, we evaluate our proposals on the databases with/without own vehicle based on DAD [5].

5.1. *Settings*

Database. We supplement more crowd-sourced dashcam videos to DAD and divide the database into two crowd-sourced dashcam videos datasets by the incidents scenarios with or without own vehicle called SDAD-1 and SDAD-2 respectively. SDAD-1 contains 200 positive videos consists of various accidents with own vehicle that include accident scenes for ten frames or more, and 600 negative videos without accident scenes. SDAD-2 consists of 600 positive videos and 1200 negative videos, while the accident scenes in positive videos in SDAD-2 are not included own vehicles. The proportion of training set and testing set on SDAD-1 is 3:1, and so is SDAD-2. The ratio of positive videos on SDAD-1 is 25%, while the percentage of positive videos on SDAD-2 is $\frac{1}{3}$.

Implementation details and evaluation metrics. We get the scene segmentation result from the model based on fast R-CNN, the result concatenate convo-

Algorithm 2 Overall Optimization for Traffic Near-miss Anticipation $loss_{s_f,s_c,D,C}(Q,G)$

```
Require:
    $D = \{x^{(n)}, y^{(n)}\}_{n=1}^{N}$: dataset with $N$ examples
    $(x^{(i)} = \{x_1^{(i)}, ..., x_{100}^{(i)}\}, y^{(i)} = \{y_1^{(i)}, ..., y_{100}^{(i)}\})$
    $Q(input)$: the Early Anticipation Model based on QRNN-Inception
    $G(input)$: the Traffic Near-miss Feature Portraits Generation Model
    $s_f(D)$: the Scene Segmentation function
    $s_c(D)$: the Convolutional Feature Maps (CFMs) of Scene Segmentation function
    $warmup(model)$: the model convergence testing function
    $C$: the number of epoches for training $Q$ and $G$
Ensure:
    Function $loss_{Q,G}(Q,G,D)$
init $Q, G$
$Q_w \leftarrow Q$
while $Q_w$ is not $warmup(Q_w)$ do
    $min\ loss_{s_f,G,D}(Q_w)$
end while
$G_w \leftarrow G$
while $G_w$ is not $warmup(G_w)$ do
    $min\ loss_{s_c,Q_w,D}(G_w)$
end while
$loss_{s_f,s_c,D}(Q_w,G_w) \leftarrow loss_{s_f,G_w,D}(Q_w) + loss_{s_c,Q_w,D}(G_w)$
for $i$ in range($C$) do
    $min_{G_w}\ loss_{s_f,s_c,D,C}(Q_w,G_w)$
    $min_{Q_w}\ loss_{s_f,s_c,D,C}(Q_w,G_w)$
end for
return $Q_w, G_w$
```

lutional feature maps from TNFP-GM and input to the early anticipation model based on QRNN-Inception. In the view of the complexity of QRNN-Inception, we employ a simpler early anticipation model based on QRNN-Inception compared with QRNN-Inception for the time complexity, MAP and TTC on both SDAD-1 and SDAD-2.

In accident anticipation, both accuracy and earliness are required. We employ the MAP, ATTC, AUC, and RECALL to access the performance of models. We calculate precision, TTC, AUC and RECALL for each threshold q, and we would get the MAP, ATTC for the average value of accuracy and TTC from multiple thresholds q.

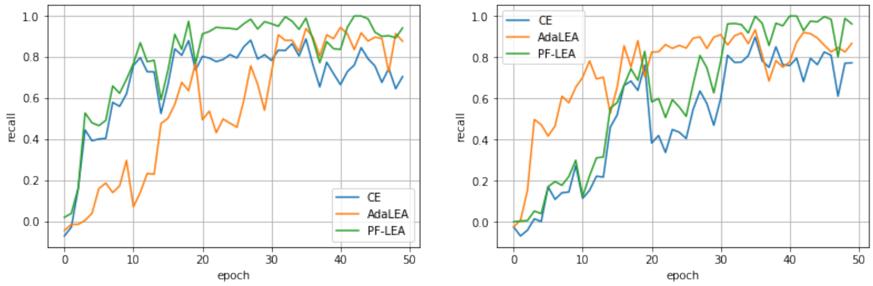
Table 1. The amount of the training set and testing set in SDAD-1 and SDAD-2.

dataset	set	pos/neg	amount
SDAD-1	training set	pos	150
		neg	450
	testing set	pos	50
		neg	150
SDAD-2	training set	pos	450
		neg	900
	testing set	pos	150
		neg	300

5.2. *Exploration*

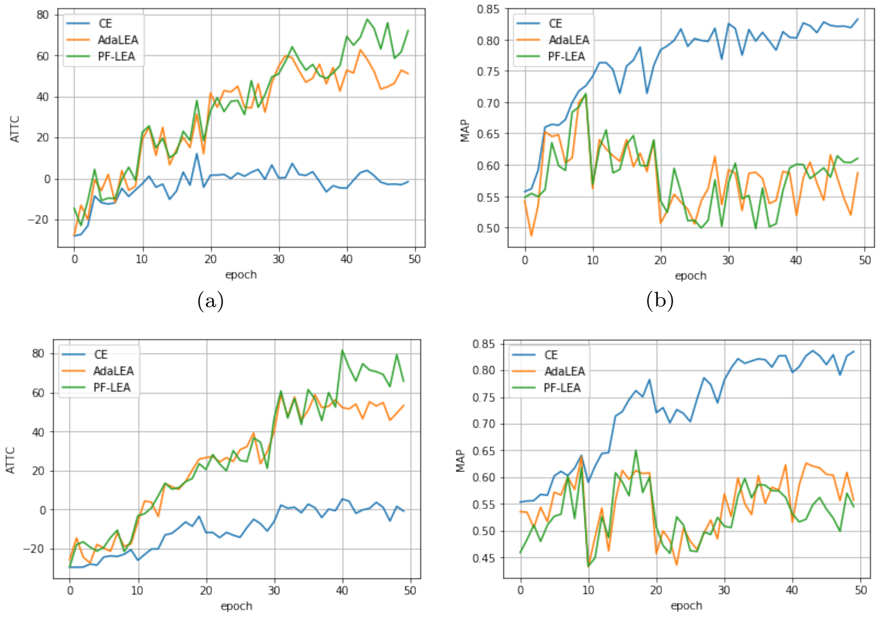
Exploration of PF-LEA. Figure 6 and 8 show the details of cross-entropy (CE), AdaLEA and PF-LEA on DAD, SDAD-1, and SDAD-2. For achieving an equitable evaluation on three loss functions, the experiments of this part employ the same model (QRNN-Inception), and only the loss function was changed. As we can see in Figure 6 (a), Figure 8 (a) and (c), the model employed with CE can not anticipate accidents as it just fit the current situation very well, the model employed with AdaLEA or PF-LEA can predict incidents about 60-80 frames before it happened, and PF-LEA can anticipate the conflicts earlier than AdaLEA without sacrificing too much predictive power. In Figure 6 (b), Figure 8 (b) and (d), the MAP of model using CE as the loss function is much higher than the models using AdaLEA and PF-LEA, the explanation of this phenomenon is that MAP and ATTC are mutually exclusive in some ways, as we anticipating the incidents long before the events happened the period between current frame and the frame accidents happened are considered mispredicted, the ATTC gets higher would lower MAP to some extent. Ultimately, we determined that the PF-LEA is the most advanced approach depending on the ability of early anticipation.

Exploration of TNFP-GM. In this section, we compare the performance of the model armed with TNFP-GM which using the framework we detailed in part 4 with the model without TNFP-GM convolutional feature maps. TNFP-GM is a feature extractor for better anticipating near-miss incidents. Table 2 and Table 3 show the results of QRNN-Inception equipped with TNFP-GM on DAD, SDAD-1, and SDAD-2. As we can see, the TNFP-GM equipped model can improve MAP as well as ATTC on DAD and SDAD-1, increase MAP while availing ATTC a bit on SDAD-2. Generally, the TNFP-GM equipped model can promote the precision of the model; we believe it can contribute to the model understanding traffic near-miss scenes better.



(a) (b)

Fig. 7. Performance of three different functions (Cross entropy, Adaptive loss for early anticipation and positive-favored loss for early anticipation) on SDAD-1 and SDAD-2.(a) recall v.s. epoch on SDAD-1 with three different loss functions. (b) recall v.s. epoch on SDAD-2 with three different functions.



(a) (b)

(c) (d)

Fig. 8. Performance of three different functions on SDAD-1 and SDAD-2: (a) details the ATTC v.s. epoch on SDAD-1 curves; (b) details MAP v.s. epoch on SDAD-1 curves; (c) details the ATTC v.s. epoch on SDAD-2 curves;; and, (d) details MAP v.s. epoch on SDAD-2 curves.

5.3. *Comparison with state-of-the-arts*

In this section, we discuss the performance of various state-of-art on DAD, SDAD-1, and SDAD-2. According to the methods proposed by [30], we enumerate various

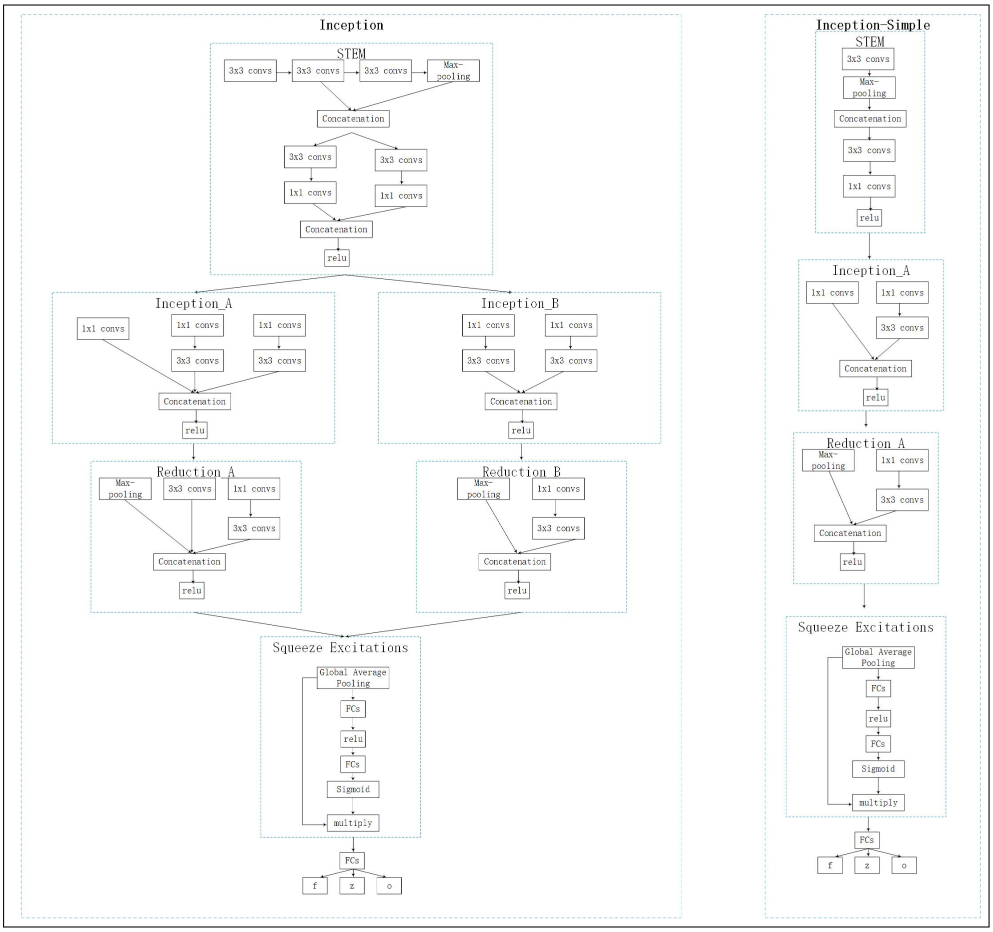
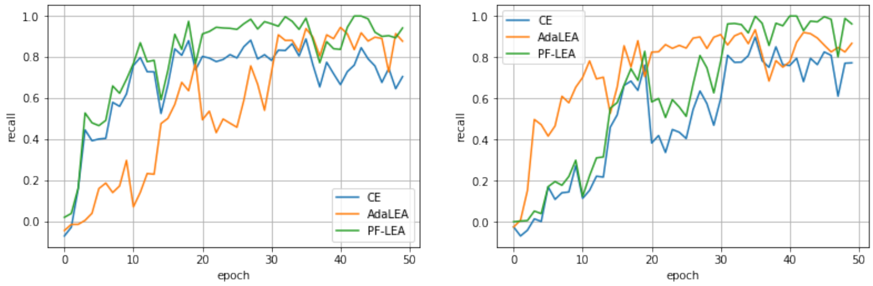


Fig. 9. Structures of Inception and Inception-simple in Traffic Near-miss Anticipation

base deep models (LSTM, QRNN), feature extractors (Inception-simple, Inception, TNFP) and loss functions (EL, AdaLEA, PF-LEA). The structures of Inception-simple and Inception are detailed in Figure 9. With all sorts of these techniques, we could improve the performance of our early anticipation model and compare with all these state-of-the-art methods.

Figure 12 shows the visual comparison with CE, AdaLEA, and PF-LEA on the base model QRNN-Inception. Figure 13 details the early anticipation visual comparison with QRNN-Inception is not employed with TNFP and employed with TNFP. It proves that our proposed PF-LEA and TNFP enabled a system to execute the earliest traffic accident anticipation. Our system anticipated when a car coming in the wrong direction appears at a distance. The quantitative results of traffic near-miss incidents anticipation are detailed



(a) recall v.s. epoch on SDAD-1

(b) recall v.s. epoch on SDAD-2

Fig. 10. Performance of three different functions (Cross entropy, Adaptive loss for early anticipation and positive-favored loss for early anticipation) on SDAD-1 and SDAD-2.(a) recall v.s. epoch on SDAD-1 with three different loss functions. (b) recall v.s. epoch on SDAD-2 with three different functions.

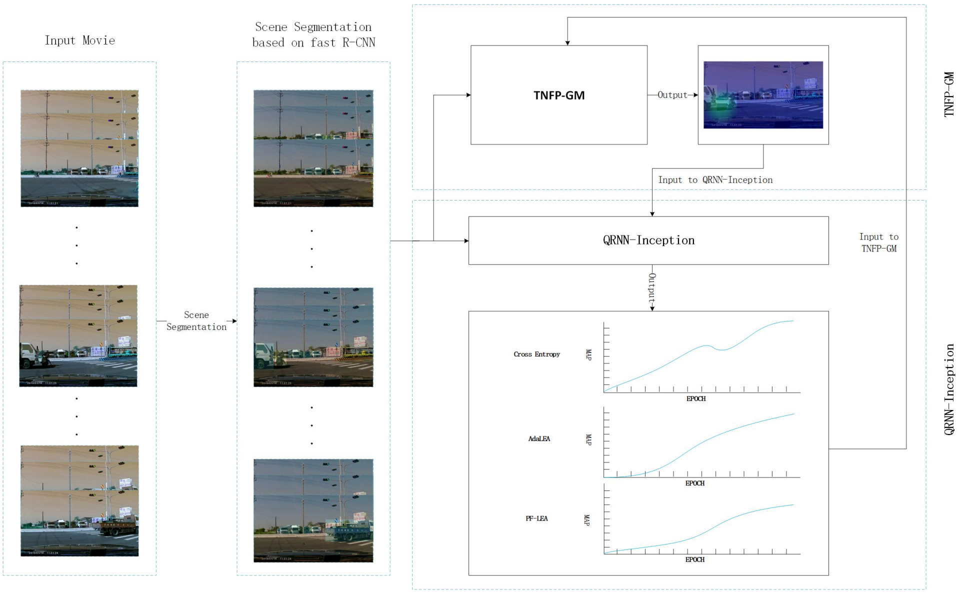


Fig. 11. Pipeline of Traffic Near-miss Incidents Anticipation

in Table 2 and Table 1 on DAD, SDAD-1, and SDAD-2. Generally, our proposed approach (QRNN+Inception+TNFP+PF-LEA) can achieve the best performance. On DAD, it achieves (+6.0%@MAP, 55.8frames@ATTC), (+2.7%@MAP, 15.8frames@ATTC), (+1.8%@MAP, 6.5frames@ATTC) in comparison with several conventional brilliant approaches. On SDAD-1, it achieves (55.9%@MAP, 82.3frames@ATTC), and achieves (53.7%@MAP, 90.3frames@ATTC) on SDAD-2. In traffic near-miss anticipation, we would rather do the wrong near-miss warning

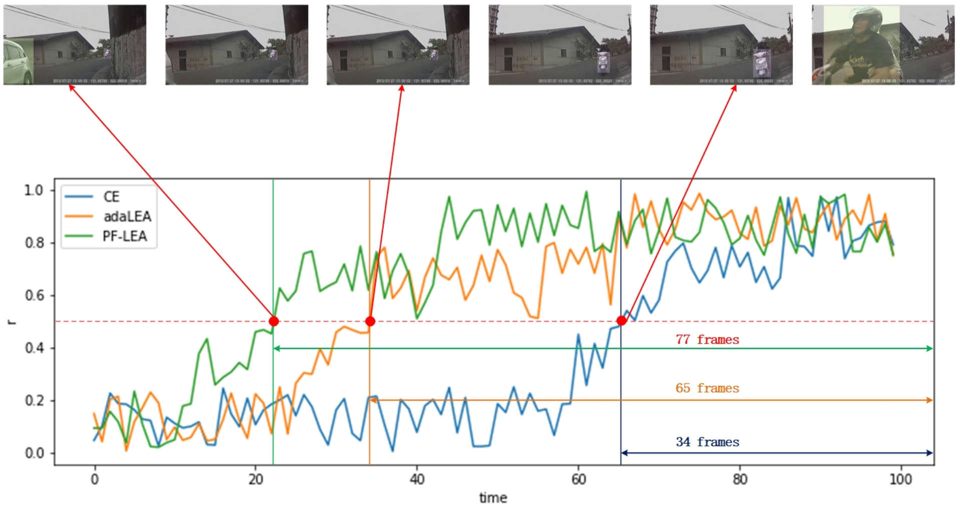


Fig. 12. Visual comparison among CE, AdaLEA and PF-LEA based on QRNN-Inception without TNFP on DAD.

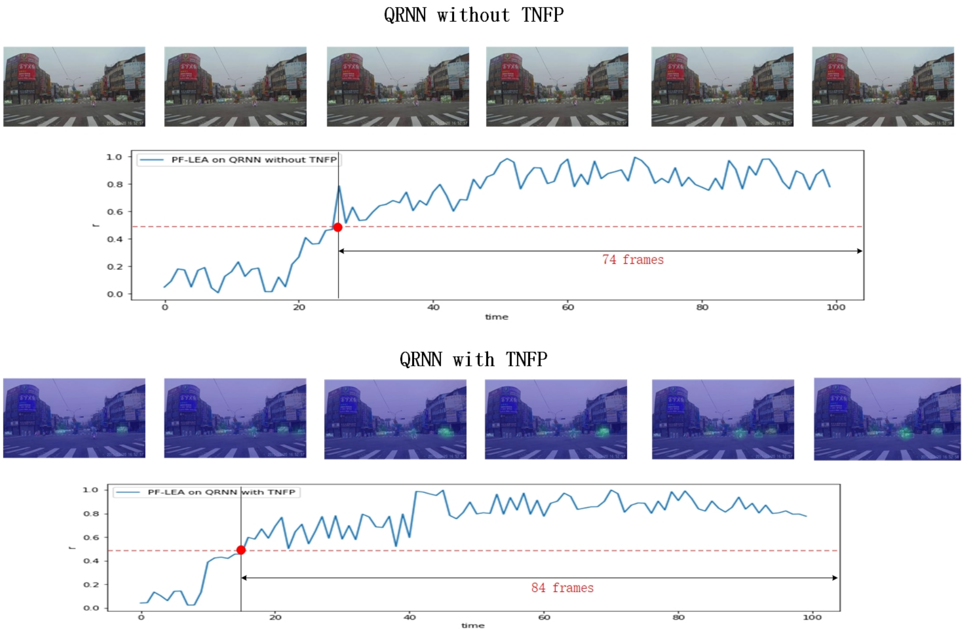


Fig. 13. Visual comparison of model with or without TNFP on DAD.

Table 2. Results of risk anticipation on DAD: the results of some other state-of-art are cited from [5], [45] and [30].

	Chan16[5]	**Zeng17[45]**	**Suzuki18[30]**	**LSTM**	**QRNN-Inception-simple**	**QRNN-Inception**	**Our final approach**
LSTM	✓	✓		✓			
QRNN			✓		✓	✓	✓
Inception-simple				✓	✓		
Inception						✓	✓
EL	✓	✓					
AdaLEA			✓				
PF-LEA				✓	✓	✓	✓
TNFP							✓
MAP(%)	48.1	51.4	52.3	51.2	52.3	53	53.2
ATTC(n-frames)	32	72	81.3	83.7	81.4	83.2	87.8

Table 3. Results of risk anticipation on SDAD-1 and SDAD-2.

	LSTM			**QRNN-Inception-simple**			**QRNN-Inception**			
LSTM	✓	✓	✓							
QRNN				✓	✓	✓	✓	✓	✓	✓
Inception-simple				✓	✓	✓				
Inception							✓	✓	✓	✓
CE	✓			✓			✓			
AdaLEA		✓			✓			✓		
PF-LEA			✓			✓			✓	✓
TNFP										✓
MAP on SDAD-1(%)	73.4	52.1	51.8	87.2	54.6	54.1	90.2	55.6	55.4	55.9
ATTC on SDAD-1(n-frames)	2.8	76.1	80.3	-1.6	79.5	80.6	0.7	79.7	82	86
MAP on SDAD-2(%)	71.6	49.7	49.6	86.5	53.1	53.4	90.1	53.4	53.1	53.8
ATTC on SDAD-2(n-frames)	2.1	72.2	75.8	0.8	69.0	72.1	-0.4	81.3	87.6	90.3

with a lower MAP than miss one near-miss accident with a higher MAP. Inspired by the idea, Figure 10 details the recall v.s. epoch curves of three different loss function on the same base model on SDAD-1 and SDAD-2 to evaluate the performance of early anticipation. It shows that PF-LEA can almost predict all of the positive frames without sacrificing too much predictive power in comparison with AdaLEA, while CE does not pay much attention to positive frames.

According to the results of the experiment, the QRNN on behalf of LSTM can improve both MAP and ATTC significantly on all of the databases. What's more, QRNN with Inception-simple or Inception does better with early anticipation. This suggests that QRNN or QRNN with multiple convolutional layers can focus on the direct relationship between frames (e.g., motion feature) and there is a possibility that QRNN is more suitable for analysis on continuous sequential data, such as videos, and Inception units can help better understanding the scenes.

6. Conclusion

We propose a data-driven approach for traffic near-miss anticipation. Inherited from DAD, we supplement more crowd-sourced dashcam videos to DAD and divide the database into two crowd-sourced dashcam videos datasets by the incidents scenarios with or without own vehicle called SDAD-1 and SDAD-2 respectively. We propose a framework ensembled with scene segmentation based on fast R-CNN, traffic near-miss feature portraits generation model (TNFP-GM) based on class activation maps [4] and recurrent neural network, and the early anticipation model based on QRNN-Inception. For loss function in early anticipation model, we present PF-LEA and experiments prove it works well. What's more, we employ a method for optimizing TNPF-GM and QRNN-Inception simultaneously. Finally, our method can achieve (+6.0%@MAP, +55.8frames@ATTC), (+2.7%@MAP, +15.8frames@ATTC), (+1.8%@MAP, +6.5frames@ATTC) in comparison with several conventional brilliant approaches on DAD, and achieve (55.9%@MAP, 86frames@ATTC), (53.8%@MAP, 90.3frames@ATTC) on SDAD-1 and SDAD-2. We believe the model can be improved if we can better understand the scenes of traffic near-miss incidents through the spatial and temporal viewpoint.

Acknowledgment

We thank for the support of National Key RD Program Intergovernmental International Science and Technology Innovation Cooperation Key Project (2016YFE0108000), Jiangsu Province Key Research and Development Program (BE2017163).

References

S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

J. Bradbury, S. Merity, C. Xiong, and R. Socher, "Quasi-recurrent neural networks," *arXiv preprint arXiv:1611.01576*, 2016.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

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.

P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International journal of computer vision*, vol. 59, no. 2, pp. 167–181, 2004.

J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 105–112.

C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM transactions on graphics (TOG)*, vol. 23, no. 3. ACM, 2004, pp. 309–314.

G. Kootstra, N. Bergström, and D. Kragic, "Fast and automatic detection and segmentation of unknown objects," in *Humanoid Robots (Humanoids), 2010 10th IEEE-RAS International Conference on.* IEEE, 2010, pp. 442–447.

E. Potapova, K. M. Varadarajan, A. Richtsfeld, M. Zillich, and M. Vincze, "Attention-driven object detection and segmentation of cluttered table scenes using 2.5 d symmetry," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on.* IEEE, 2014, pp. 4946–4952.

C. L. Teo, C. Fermuller, and Y. Aloimonos, "Detection and segmentation of 2d curved reflection symmetric structures," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1644–1652.

X. Ren, L. Bo, and D. Fox, "Rgb-(d) scene labeling: Features and algorithms," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.* IEEE, 2012, pp. 2759–2766.

N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision.* Springer, 2012, pp. 746–760.

S. Gupta, P. Arbelaez, and J. Malik, "Perceptual organization and recognition of indoor scenes from rgb-d images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 564–571.

N. Silberman and R. Fergus, "Indoor scene segmentation using a structured light sensor," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on.* IEEE, 2011, pp. 601–608.

C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," *arXiv preprint arXiv:1301.3572*, 2013.

S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri, "Geometry driven semantic labeling of indoor scenes," in *European Conference on Computer Vision.* Springer, 2014, pp. 679–694.

S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European Conference on Computer Vision.* Springer, 2014, pp. 345–360.

H. S. Koppula, A. Anand, T. Joachims, and A. Saxena, "Semantic labeling of 3d point clouds for indoor scenes," in *Advances in neural information processing systems*, 2011, pp. 244–252.

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville, "Reseg: A recurrent neural network-based model for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 41–48.

B. Shuai, Z. Zuo, B. Wang, and G. Wang, "Scene segmentation with dag-recurrent neural networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1480–1493, 2018.

W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki, "Scene labeling with lstm recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3547–3555.

H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

G. Lin, C. Shen, A. Van Den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3194–3203.

S. Liu, X. Qi, J. Shi, H. Zhang, and J. Jia, "Multi-scale patch aggregation (mpa) for simultaneous detection and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3141–3149.

P. H. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *31st International Conference on Machine Learning (ICML)*, no. EPFL-CONF-199822, 2014.

T. Suzuki, H. Kataoka, Y. Aoki, and Y. Satoh, "Anticipating traffic accidents with adaptive loss and large-scale incident db," *arXiv preprint arXiv:1804.02675*, 2018.

H. Kataoka, T. Suzuki, S. Oikawa, Y. Matsui, and Y. Satoh, "Drive video analysis for the detection of traffic near-miss incidents," *arXiv preprint arXiv:1804.02555*, 2018.

J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction," in *European Conference on Computer Vision.* Springer, 2014, pp. 618–633.

D. Xie, S. Todorovic, and S.-C. Zhu, “inferring dark matter and dark energy from videos,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2224–2231.

H. Gong, J. Sim, M. Likhachev, and J. Shi, “Multi-hypothesis motion planning for visual object tracking,” in *Computer Vision (ICCV), 2011 IEEE International Conference on.* IEEE, 2011, pp. 619–626.

M. S. Ryoo, “Human activity prediction: Early recognition of ongoing activities from streaming videos,” in *Computer Vision (ICCV), 2011 IEEE International Conference on.* IEEE, 2011, pp. 1036–1043.

H. Kataoka, Y. Satoh, Y. Aoki, S. Oikawa, and Y. Matsui, “Temporal and fine-grained pedestrian action recognition on driving recorder database,” *Sensors*, vol. 18, no. 2, p. 627, 2018.

H. Kataoka, Y. Miyashita, M. Hayashi, K. Iwata, and Y. Satoh, “Recognition of transitional action for short-term action prediction using discriminative temporal cnn feature.” in *BMVC*, 2016.

M. S. Aliakbarian, F. S. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, “Encouraging lstms to anticipate actions very early,” in *IEEE International Conference on Computer Vision (ICCV)*, vol. 1, no. 2, 2017.

H. S. Koppula and A. Saxena, “Anticipating human activities using object affordances for reactive robotic response,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 14–29, 2016.

C. Vondrick, H. Pirsiavash, and A. Torralba, “Anticipating visual representations from unlabeled video,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 98–106.

S. Ma, L. Sigal, and S. Sclaroff, “Learning activity progression in lstms for activity detection and early detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1942–1950.

A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, “Recurrent neural networks for driver activity anticipation via sensory-fusion architecture,” in *Robotics and Automation (ICRA), 2016 IEEE International Conference on.* IEEE, 2016, pp. 3118–3125.

J. Gao, Z. Yang, and R. Nevatia, “Red: Reinforced encoder-decoder networks for action anticipation,” *arXiv preprint arXiv:1707.04818*, 2017.

A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.* IEEE, 2012, pp. 3354–3361.

K.-H. Zeng, S.-H. Chou, F.-H. Chan, J. C. Niebles, and M. Sun, “Agent-centric risk assessment: Accident anticipation and risky region localization,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, no. 4, 2017, p. 6.

[1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44]