Global Feature Aggregation for Accident Anticipation

Mishal Fatima, Muhammad Umar Karim Khan, and Chong-Min Kyung

Korea Advanced Institute of Science and Technology

Motivation

- Approximately 1.35 million people die in road crashes each year, on average 3,700 people lose their lives every day on the roads.[1]

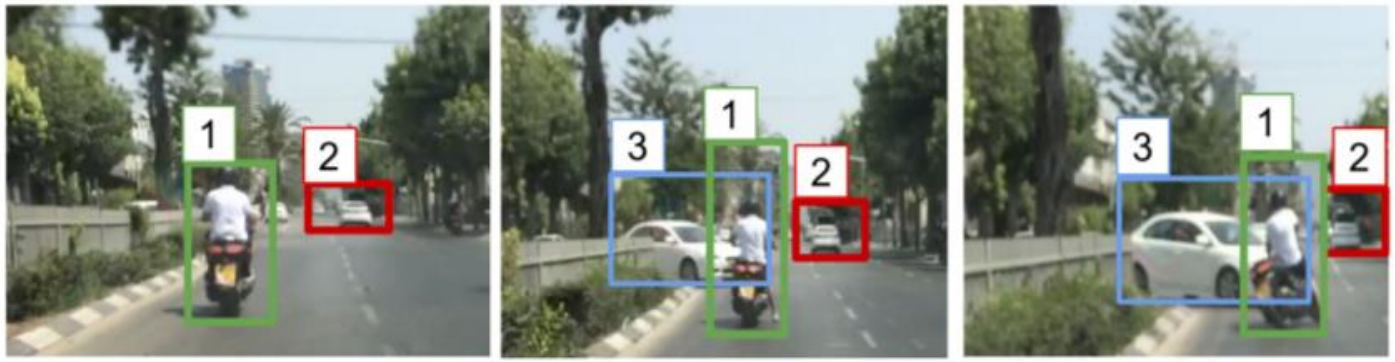
- These accidents include vehicles colliding with one another, with animals or pedestrians, and with road signs.

Dashcam footage showing an accident.

[1] https://www.asirt.org/safe-travel/road-safety-facts/

Challenges

- Wide variety of vehicles can cause accidents in real life.
- Less generalization ability of neural network for all kinds of accidents.
- It is thus important to model relationship between appearance features of different objects.



Different objects and their interactions.

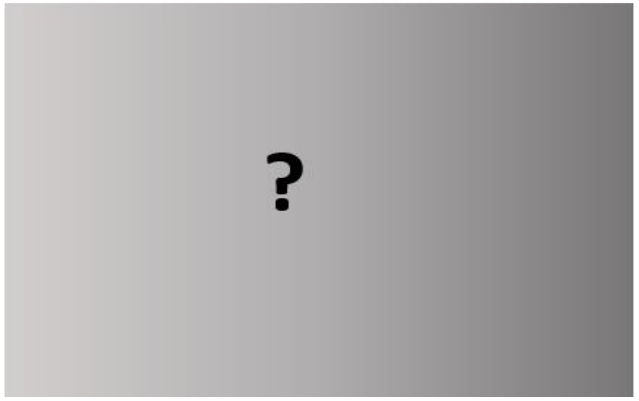
Detection and Anticipation: Difference

Action Detection.

Action Anticipation.

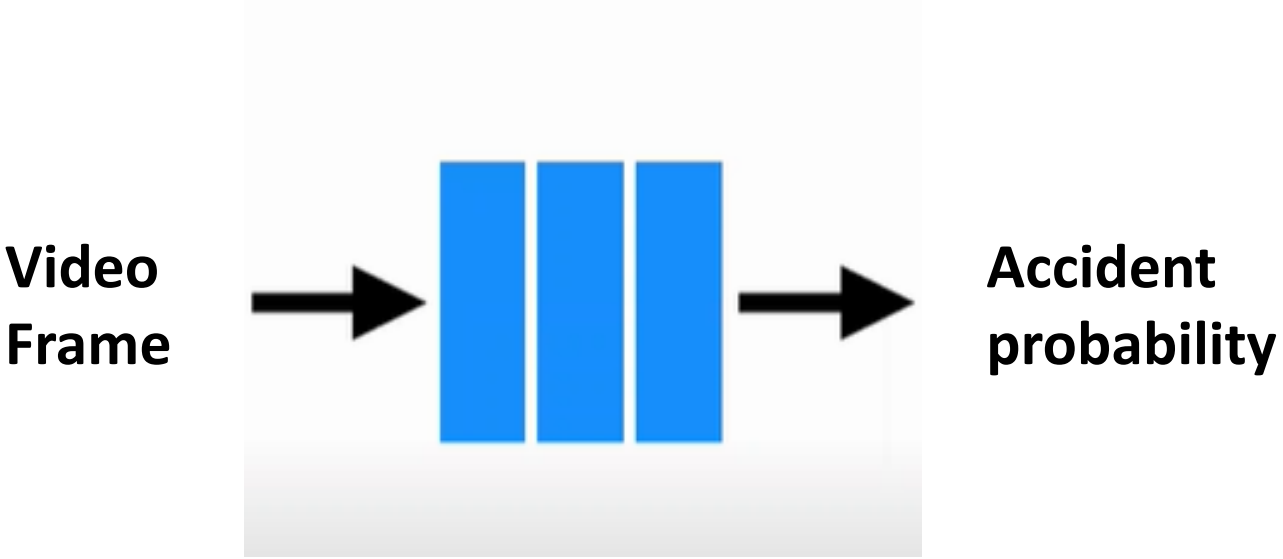
Detection and Anticipation: Difference

Action Detection.



Action Anticipation.

Accident Anticipation: Problem Formulation



Contributions

- We propose a novel **Feature Aggregation Block** that takes into account inter-object interactions and use it for the application of road accidents anticipation.

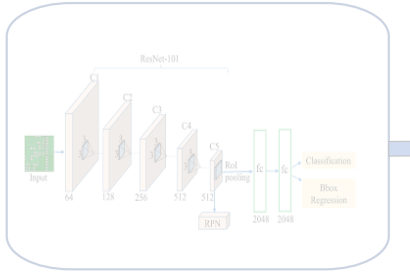
- Provides superior Average-Time-to-Accident (ATTA) compared with other approaches.

- The mean Average Precision (mAP) is also comparable to the state-of-the-art methods.

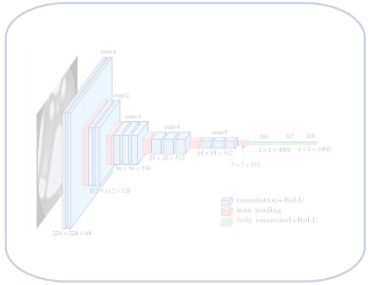
Proposed Approach

Proposed method

Dashcam Footage

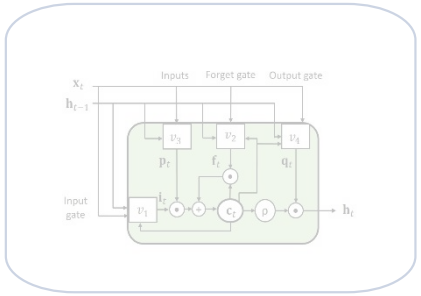


Object Detection



Feature Extraction

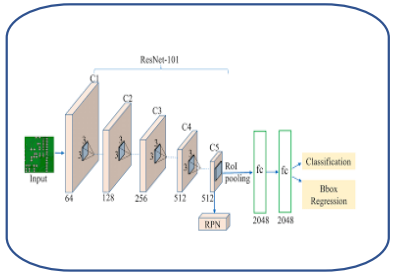
Feature Aggregation



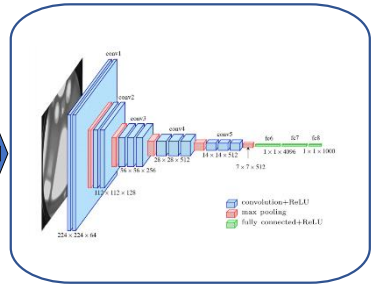
LSTM

Proposed method

Dashcam Footage

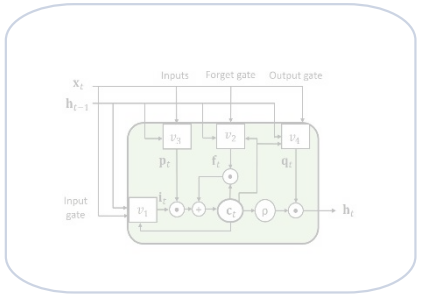


Object Detection



Feature Extraction

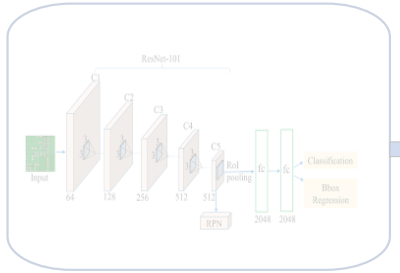
Feature Aggregation



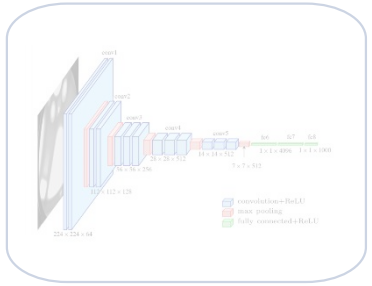
LSTM

Proposed method

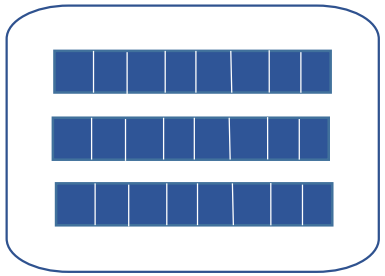
Dashcam Footage



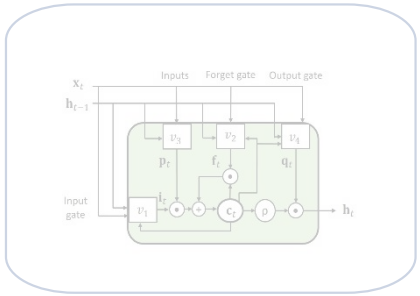
Object Detection



Feature Extraction



Feature Aggregation



LSTM

Feature Aggregation Block-Explained

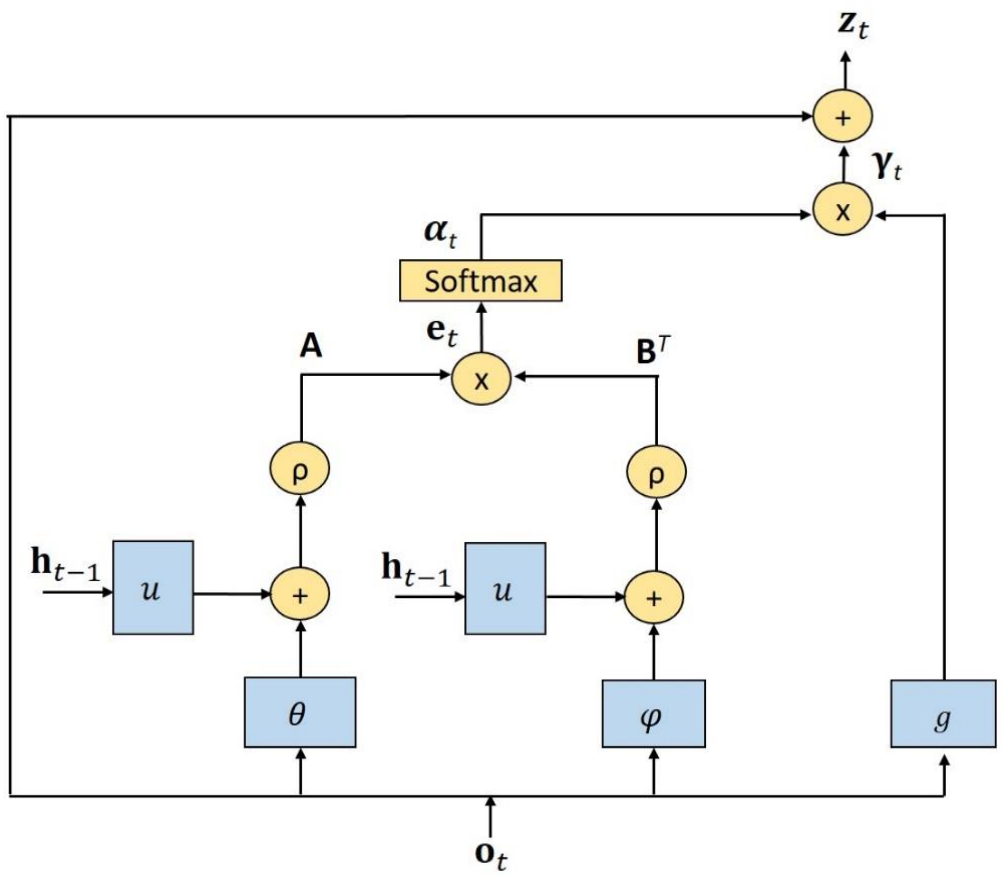
The FA block globally aggregates features over a frame.

In order to detect accidents, it is important that the network understands the global context surrounding an object in a given frame.

The main purpose of FA block is to comprehend object interactions in the neural network.

It has further two component

- Appearance comparison
- Feature refinement



Appearance Comparison: Computes appearance relationship between objects in a given frame.

$$\theta\left(\mathbf{W}_\theta, \mathbf{b}_\theta, \mathbf{o}_t^i\right) = \mathbf{W}_\theta \mathbf{o}_t^i + \mathbf{b}_\theta$$
$$\varphi\left(\mathbf{W}_\varphi, \mathbf{b}_\varphi, \mathbf{o}_t^j\right) = \mathbf{W}_\varphi \mathbf{o}_t^j + \mathbf{b}_\varphi$$

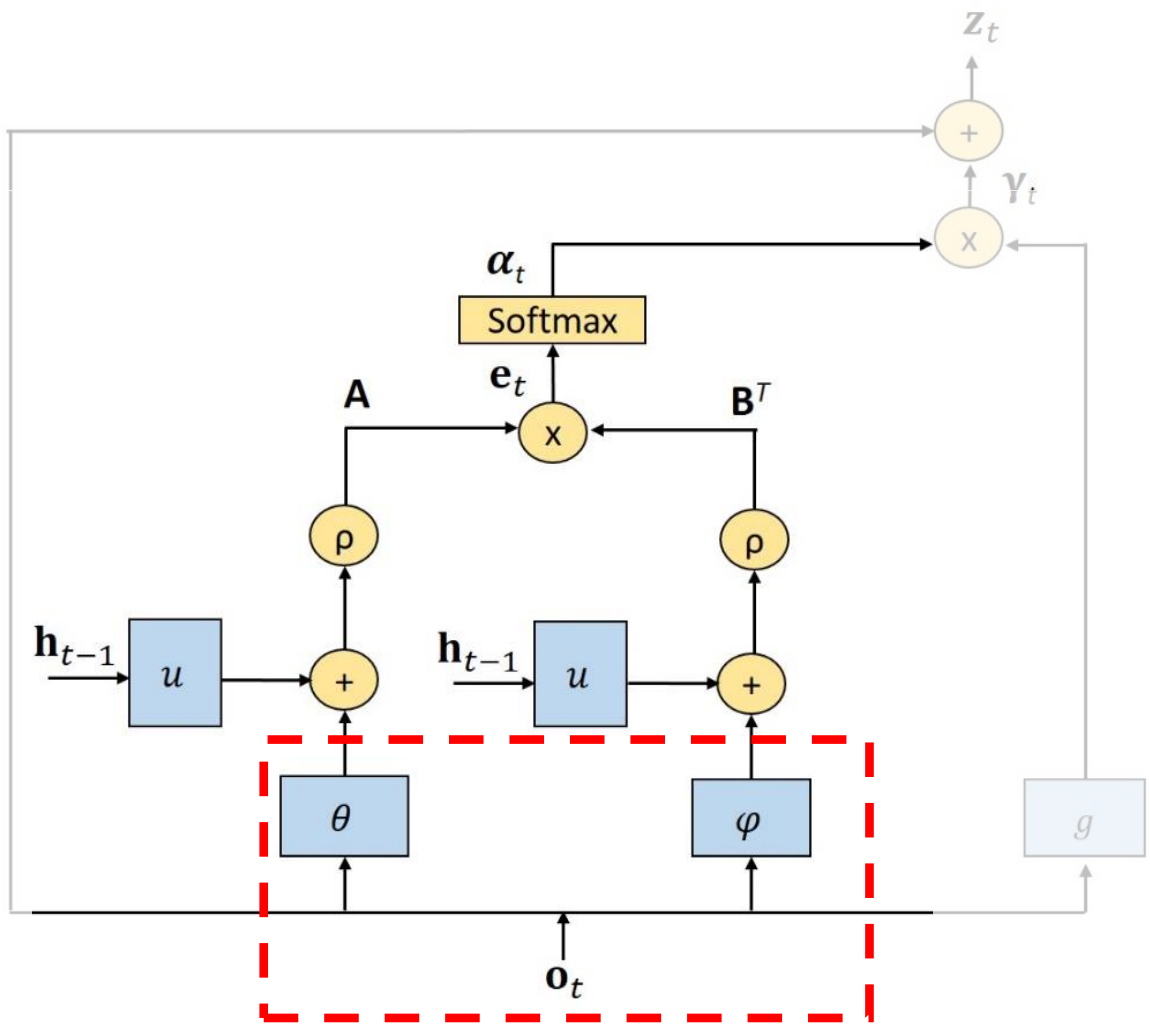
Where

i: Query object

j: All possible objects in a given frame including the query object *i*.

$\mathbf{W}_\theta, \mathbf{W}_\varphi, \mathbf{b}_\theta$ and $\mathbf{b}_\varphi$ are all learnable parameters of fully connected layers.

$\mathbf{o}_t$: Input Features



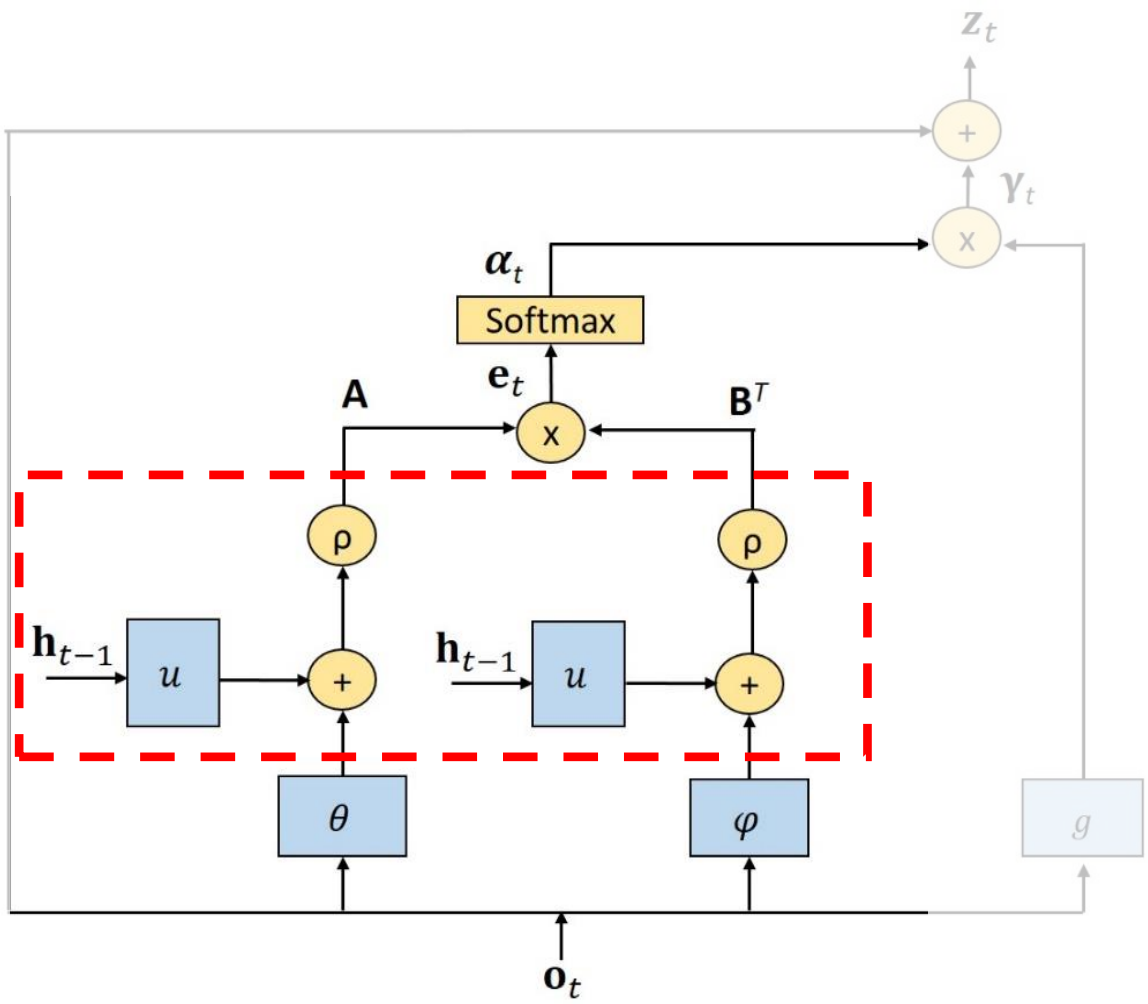
Appearance Comparison: Computes appearance relationship between objects in a given frame.

$$\theta\left(\mathbf{W}_\theta, \mathbf{b}_\theta, \mathbf{o}_t^i\right) = \mathbf{W}_\theta \mathbf{o}_t^i + \mathbf{b}_\theta$$
$$\varphi\left(\mathbf{W}_\varphi, \mathbf{b}_\varphi, \mathbf{o}_t^j\right) = \mathbf{W}_\varphi \mathbf{o}_t^j + \mathbf{b}_\varphi$$

$$u(\mathbf{W}_u, \mathbf{h}_{t-1}) = \mathbf{W}_u \mathbf{h}_{t-1}$$
$$\mathbf{A}^i = \rho[u(\mathbf{W}_u, \mathbf{h}_{t-1}) + \theta\left(\mathbf{W}_\theta, \mathbf{b}_\theta, \mathbf{o}_t^i\right)]$$
$$\mathbf{B}^j = \rho[u(\mathbf{W}_u, \mathbf{h}_{t-1}) + \varphi\left(\mathbf{W}_\varphi, \mathbf{b}_\varphi, \mathbf{o}_t^j\right)]$$

Where

- $\mathbf{W}_u$ is a learnable parameter of fully connected layer.
- ρ is the hyperbolic tangent function.
- $\mathbf{A} \in \mathbb{R}^{N \times D}$ and $\mathbf{B} \in \mathbb{R}^{N \times D}$ show learnable transformations of object features.



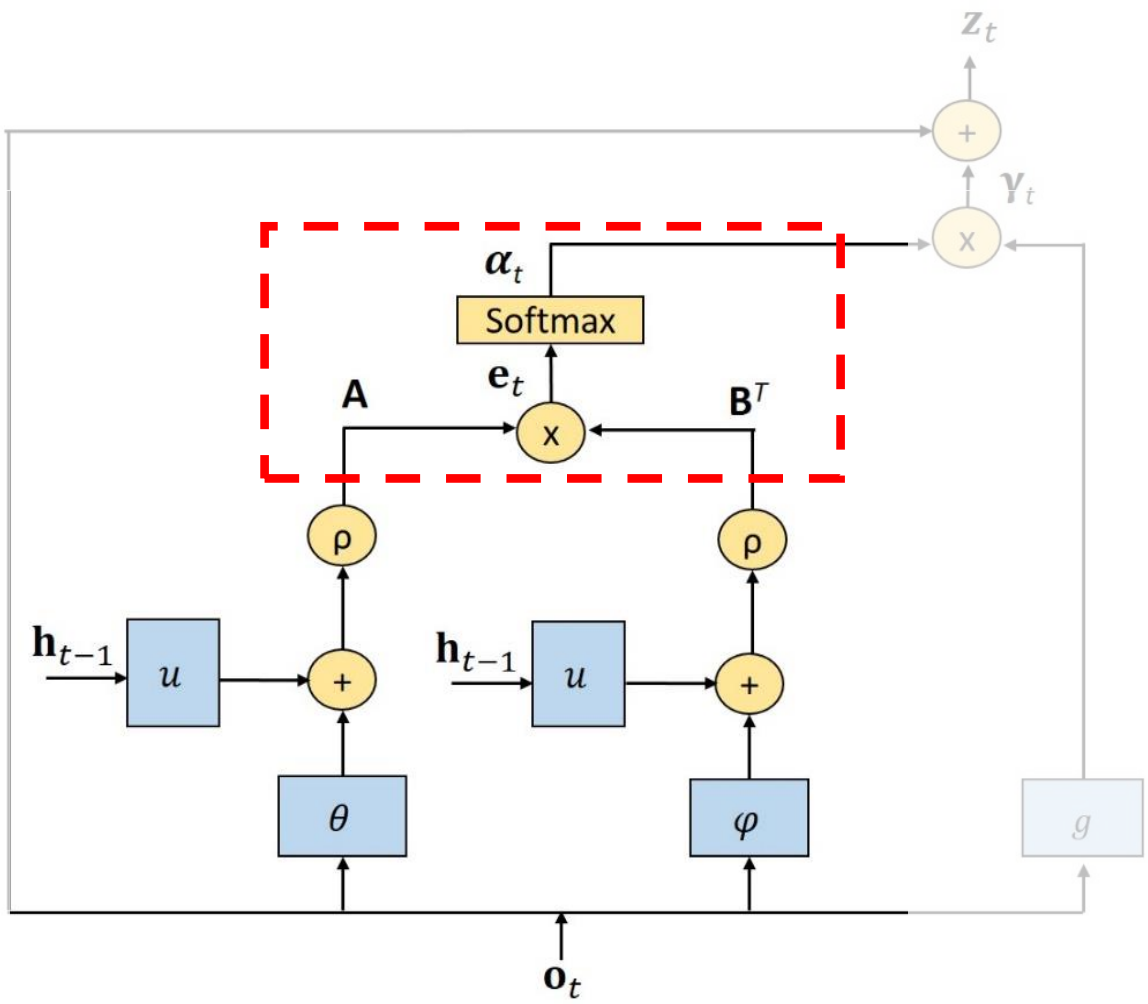
Appearance Comparison: Computes appearance relationship between objects in a given frame.

$$\theta\left(\mathbf{W}_\theta, \mathbf{b}_\theta, \mathbf{o}_t^i\right) = \mathbf{W}_\theta \mathbf{o}_t^i + \mathbf{b}_\theta$$
$$\varphi\left(\mathbf{W}_\varphi, \mathbf{b}_\varphi, \mathbf{o}_t^j\right) = \mathbf{W}_\varphi \mathbf{o}_t^j + \mathbf{b}_\varphi$$

$$u(\mathbf{W}_u, \mathbf{h}_{t-1}) = \mathbf{W}_u \mathbf{h}_{t-1}$$
$$\mathbf{A}^i = \rho[u(\mathbf{W}_u, \mathbf{h}_{t-1}) + \theta\left(\mathbf{W}_\theta, \mathbf{b}_\theta, \mathbf{o}_t^i\right)]$$
$$\mathbf{B}^j = \rho[u(\mathbf{W}_u, \mathbf{h}_{t-1}) + \varphi\left(\mathbf{W}_\varphi, \mathbf{b}_\varphi, \mathbf{o}_t^j\right)]$$

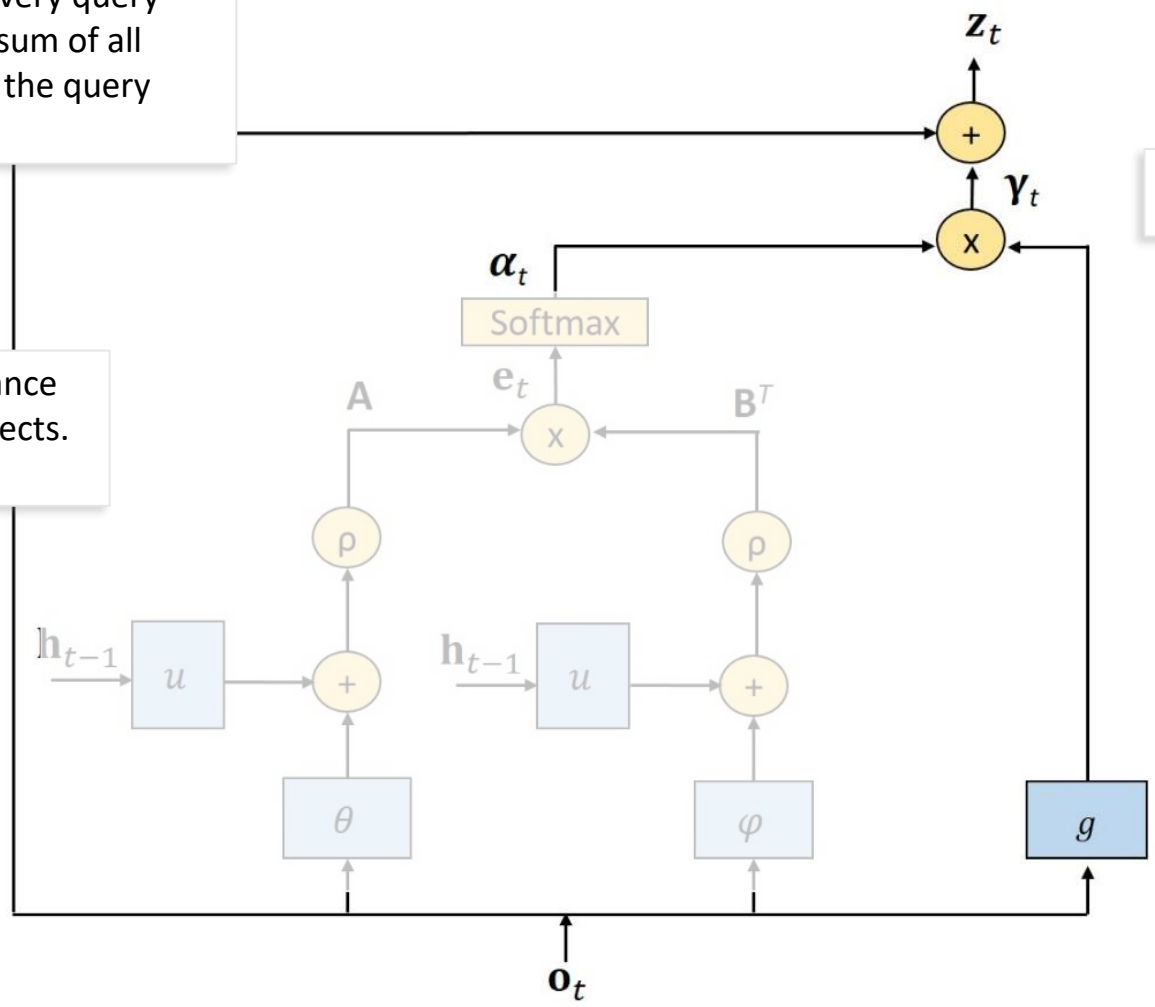
Dot Product Similarity: $e_t^{ij} = \mathbf{A}^i (\mathbf{B}^j)^T$

$$\alpha_t^{ij} = \frac{\exp\left(e_t^{ij}\right)}{\sum_j \exp\left(e_t^{ij}\right)}$$



Feature Refinement: The FA block strengthens the features of every query object by adding a weighted sum of all objects present in a frame to the query object.

The weights indicate appearance relationship between the objects.

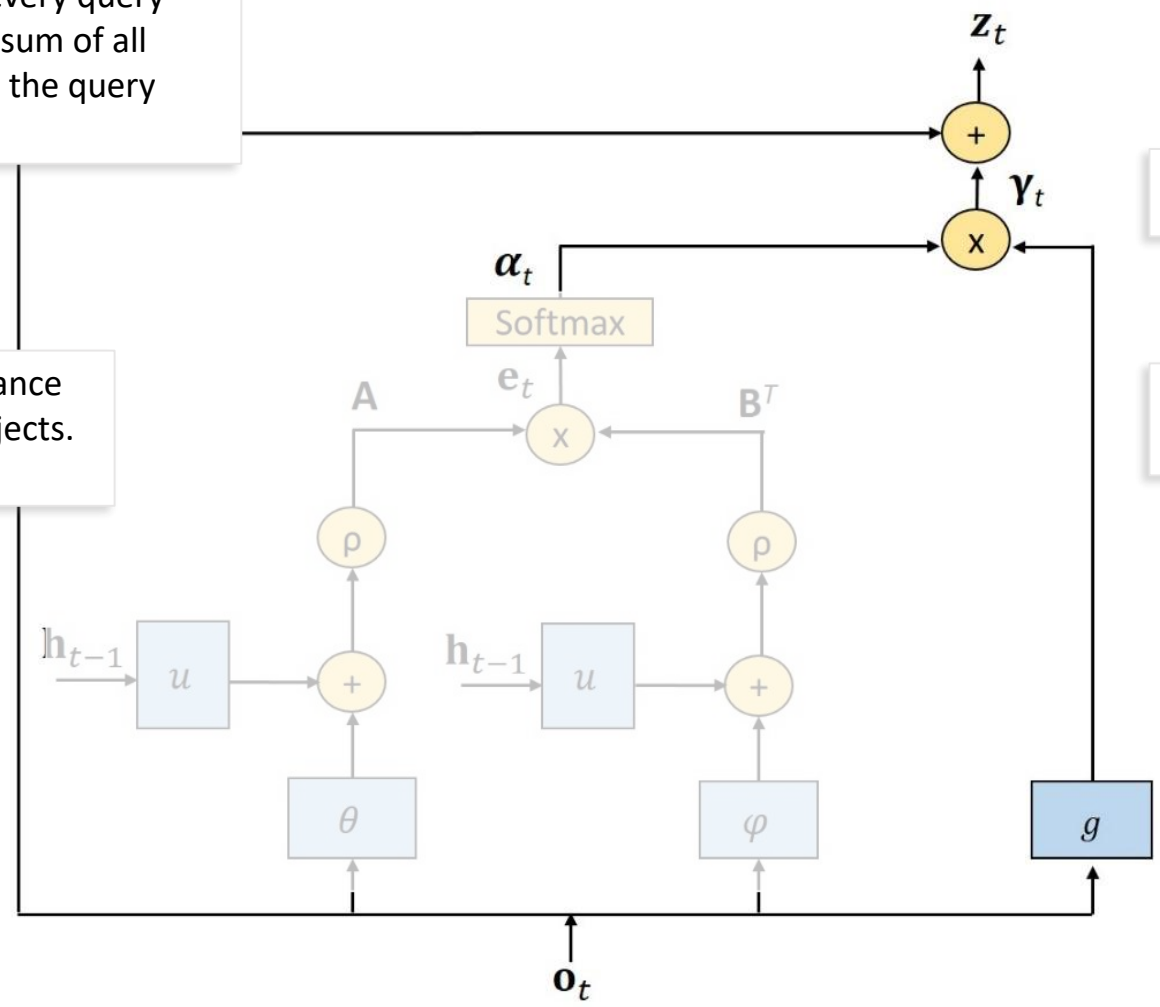


$$\boldsymbol{\gamma}_t^i = \sum_{j=1}^{N} \alpha_t^{ij} g\left(\mathbf{W}_g, \mathbf{b}_g, \mathbf{o}_t^j\right)]$$

$$g\left(\mathbf{W}_g, \mathbf{b}_g, \mathbf{o}_t^j\right) = \mathbf{W}_g \mathbf{o}_t^j + \mathbf{b}_g$$

Feature Refinement: The FA block strengthens the features of every query object by adding a weighted sum of all objects present in a frame to the query object.

The weights indicate appearance relationship between the objects.



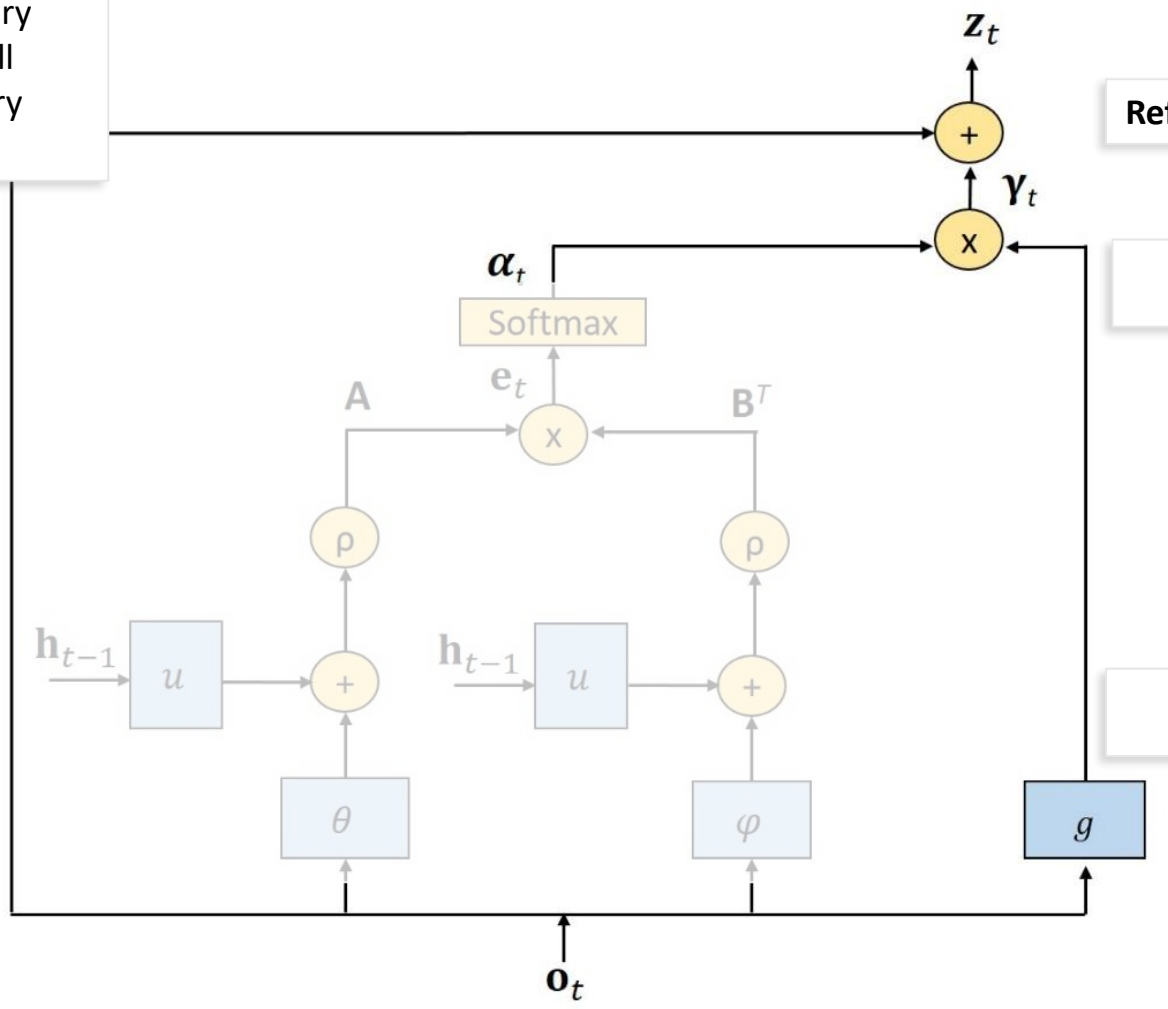
$$\boldsymbol{\gamma}_t^i = \sum_{j=1}^{N} \alpha_t^{ij}\, g\left(\mathbf{W}_g, \mathbf{b}_g, \mathbf{o}_t^j\right)]$$

$\boldsymbol{\gamma}_t^i$ represents global context related to query object *i*.

$$g\left(\mathbf{W}_g, \mathbf{b}_g, \mathbf{o}_t^j\right) = \mathbf{W}_g \mathbf{o}_t^j + \mathbf{b}_g$$

Feature Refinement: The FA block strengthens the features of every query object by adding a weighted sum of all objects present in a frame to the query object.

The weights indicate appearance relationship between the objects.



$$z_t^i = o_t^i + \gamma_t^i$$

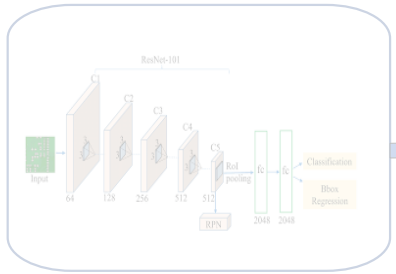
Refined Object Features

$$\gamma_t^i = \sum_{j=1}^{N} \alpha_t^{ij} g\left(W_g, b_g, o_t^j\right)]$$

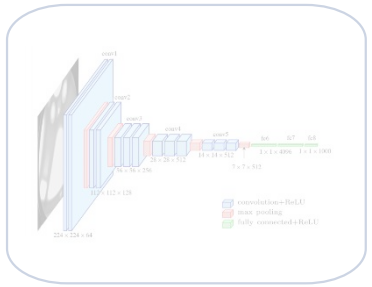
$$g\left(W_g, b_g, o_t^j\right) = W_g o_t^j + b_g$$

Proposed method

Dashcam Footage

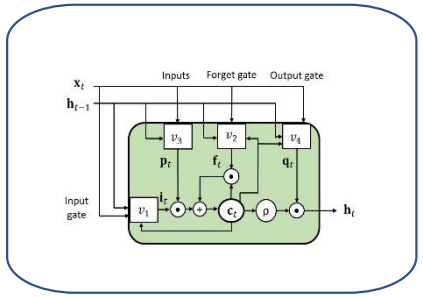


Object Detection



Feature Extraction

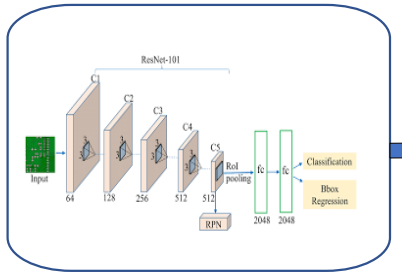
Feature Aggregation



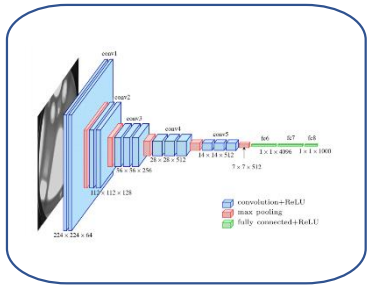
LSTM

Proposed method

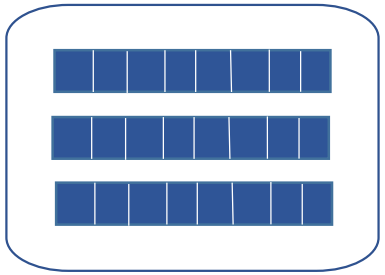
Dashcam Footage



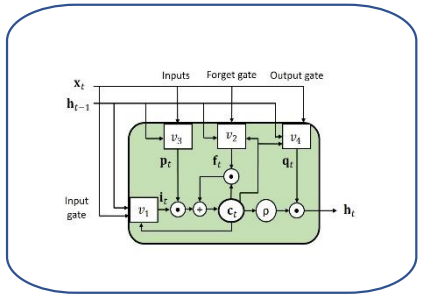
Object Detection



Feature Extraction



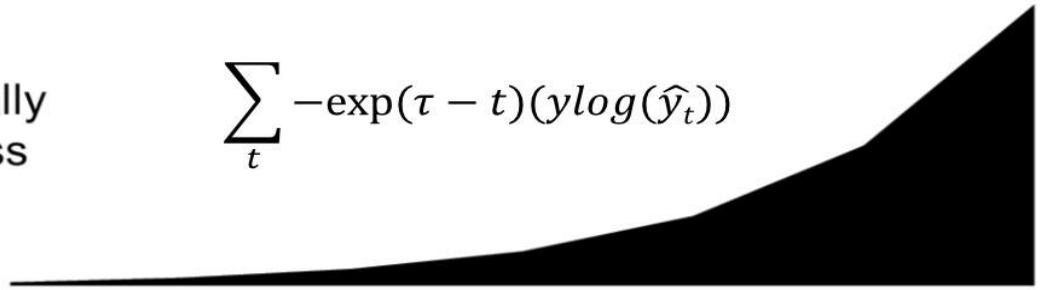
Feature Aggregation



LSTM

Exponentially growing loss

$$\sum_{t} -\exp(\tau - t)(y log(\widehat{y}_t))$$




Experiments and Results

Evaluation Metrics

- **Mean Average Precision (mAP)**
  - The general definition of mean Average Precision is area under the precision recall curve.

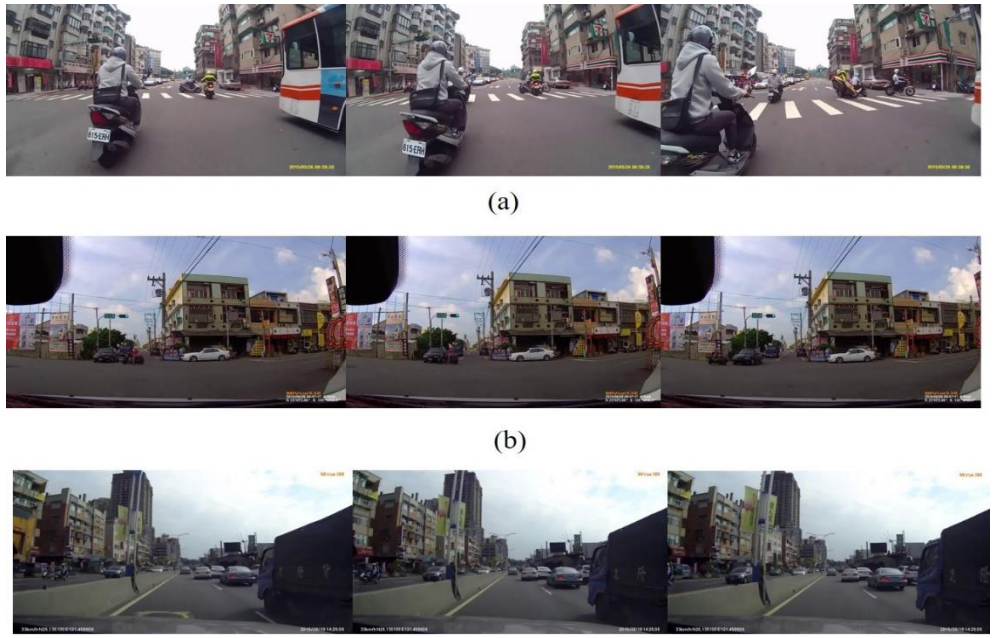
$$mAP = \int_0^1 p(r)dr$$

where $p(r)$ is precision as a function of recall r.

- **Average Time to Accident (ATTA)**
  - At every threshold β , we find the first value $\hat{t}$ in every positive video when the accident probability is above a threshold. If the accident starts at frame τ, then τ-$\hat{t}$ is Time-to-Accident (TTA).
  - Average all TTAs for all the positive videos to get a single TTA at a given threshold.
  - Average all TTAs at different thresholds to find Average Time-to-Accident (ATTA).
  - A higher ATTA value means earlier anticipation of accidents.

Street Accident Dataset

- It contains videos captured across 6 cities in Taiwan.
  - Frame rate: 20 frames per second.
  - Spatial resolution of frames: 1280 x 720.
  - Duration of a video: 5 seconds.



(a)

(b)

(c)

Accident Type	Positive examples	Negative examples	Total
Training set	455	829	1284
Testing set	165	301	466
Total	620	1130	1730

SA dataset distribution

Accident Type	Dist. (%)
Motorbike hits car	42.6
Car hits car	19.7
Motorbike hits motorbike	15.6
Others	20

SA dataset statistics

Quantitative Results

- We describe the following five variants of our Feature Aggregation block.
  - **FA-1**: Fully connected layers with parameters $\mathbf{W}_\theta$ and $\mathbf{W}_\varphi$ are removed from FA block.
  - **FA-2**: Softmax function is replaced by multiplication with $1/N$.
  - **FA-3**: Tanh activation function is replaced by ReLU.
  - **FA-4**: Instead of using dot product similarity as relation function, we use the relation network module proposed in [2] to find attention weights. It is given as,

$$e_t^{ij} = ReLU\left(\mathbf{W}\left[\mathbf{A}^i; \mathbf{B}^j\right] + \mathbf{b}\right)$$

    W and b are learnable parameters of fully connected layer that project the concatenated vector to a scalar value.
  - **FA-Final:** This is our final network with fully connected layers $\mathbf{W}_\theta$ and $\mathbf{W}_\varphi$, dot product similarity, softmax and tanh activation function.

[2] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, “A simple neural network module for relational reasoning,” in Advances in neural information processing systems, 2017, pp. 4967–4976.

Quantitative Results

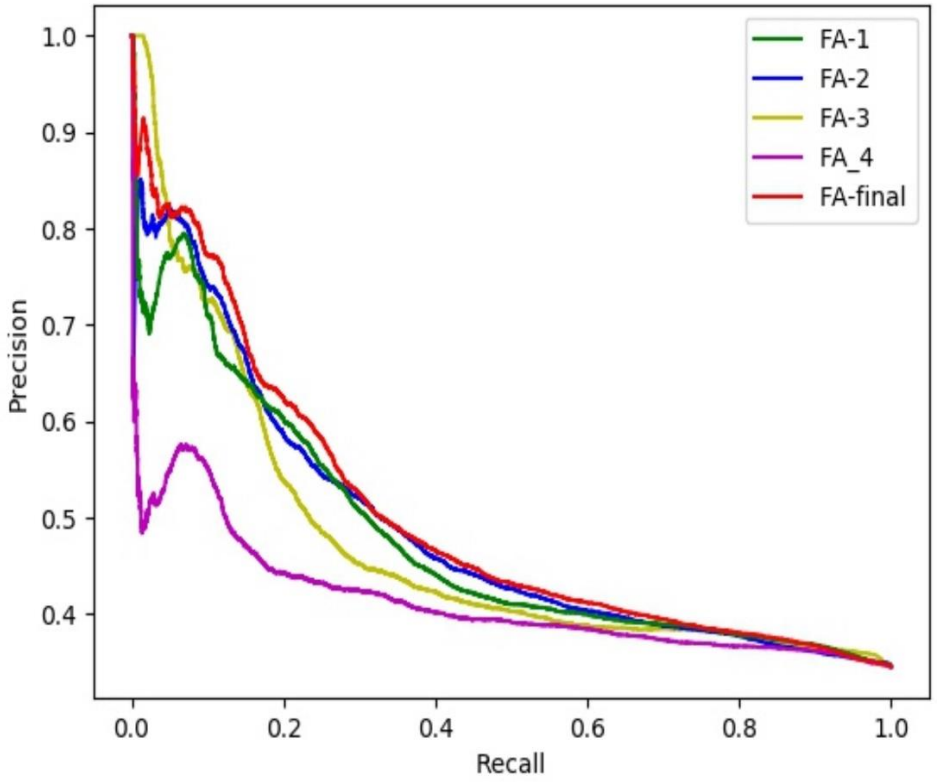
Method	mAP (%)	ATTA(s)
DSA	48.1	1.34
SP	47.3	1.66
L-R*CNN	37.4	3.13
L-RA	49.1	3.04
L-RAI	51.4	3.01
AdaLEA	53.2	3.44
VGG + full frame	37.3	3.21

Experimental Results with SOTA approaches

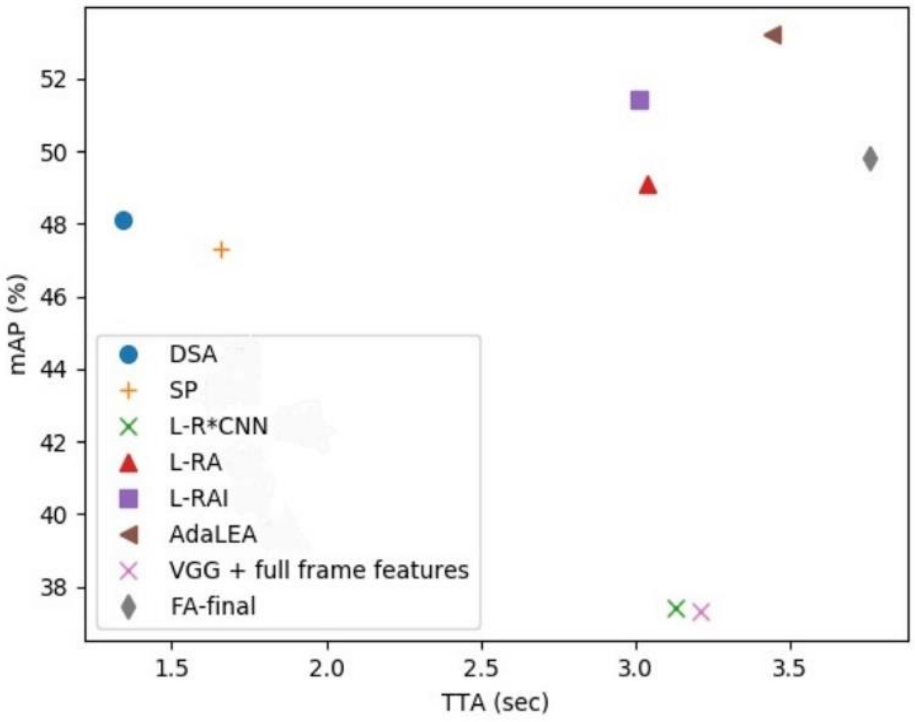
Figure Variants	mAP (%)	ATTA(s)
FA-1	47.7	3.29
FA-2	48.6	3.21
FA-3	47.2	3.23
FA-4	41.3	3.56
FA-Final	**49.8**	**3.76**

Experimental Results with different variants of FA block.

Experimental Results

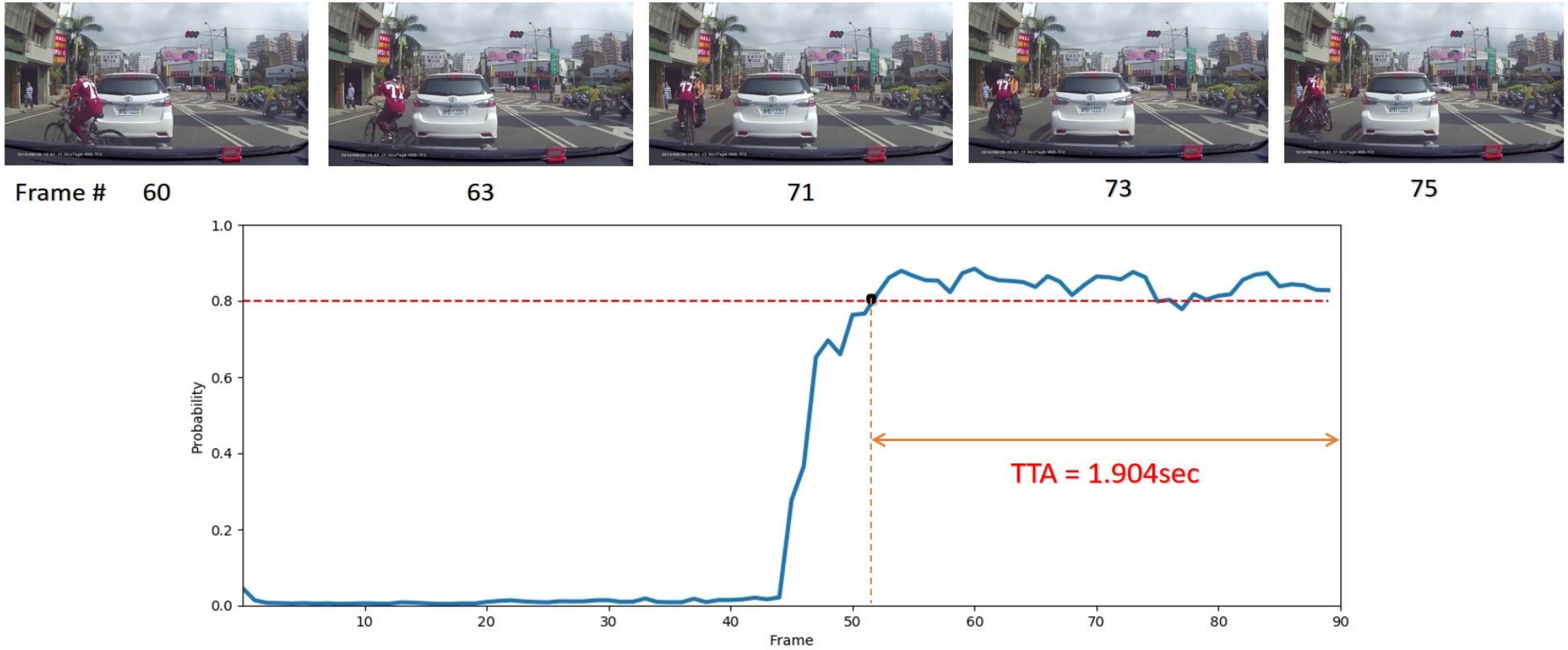


Precision vs. Recall curves for different variants of FA block.



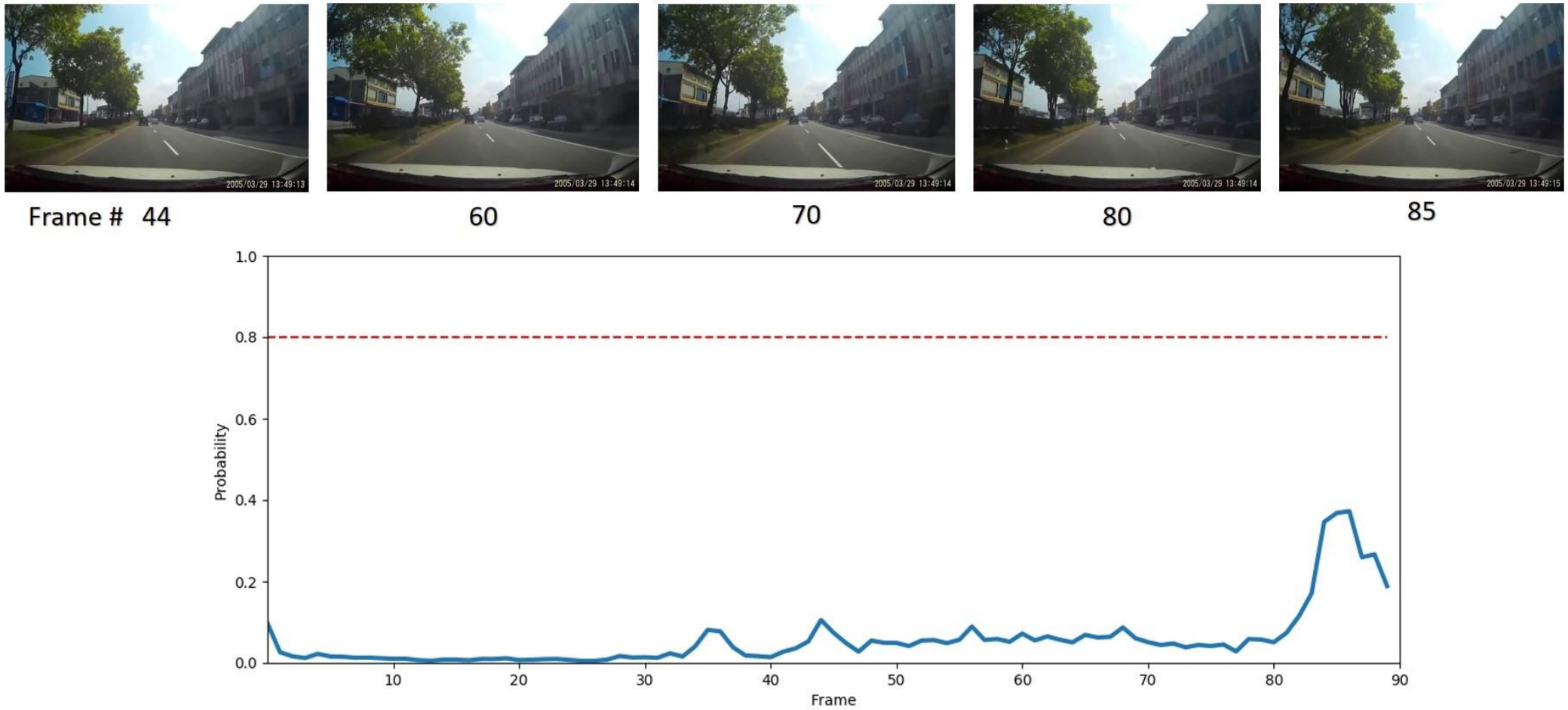
Anticipation and accuracy comparison of different methods.

Qualitative Results



A positive example. Threshold for triggering accident anticipation is kept at 0.8.

Qualitative Results



A negative example. The probability value never exceeds the threshold.

Qualitative Results

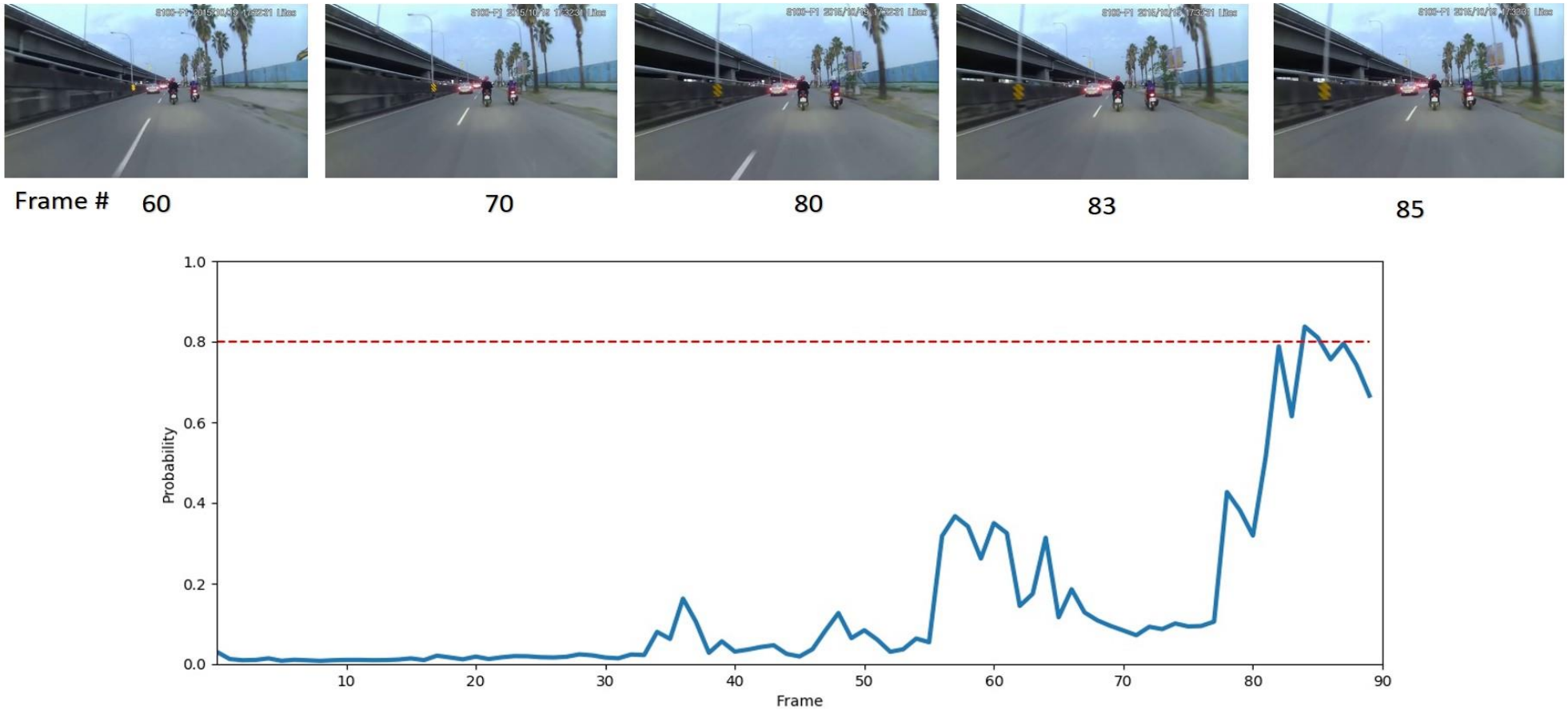
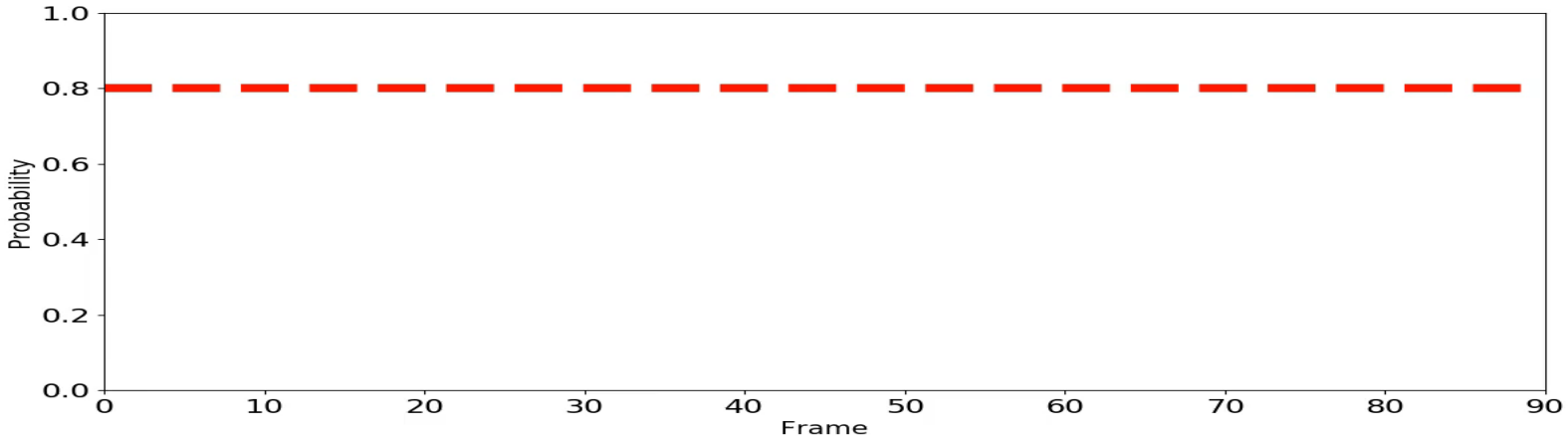


Fig. 13. A false positive. The failure can be attributed to the fact that vehicles are too close to one another.


源遠路
Yuanyuan Rd
多時相號誌
MIVue538
2015/06/03 09:28:35 0 Kmh
N 25°06'16.5" E 121°44'6.10"


1.0
0.8
0.6
0.4
0.2
0.0
Probability
0
10
20
30
40
50
60
70
80
90
Frame

Conclusion

Conclusion

- A novel Feature Aggregation block is proposed that is used for anticipation of road accidents.
- The FA block refines each object's features by using the appearance relation between different objects in a given frame.
- Using FA block along with an LSTM provides us with complementary information related to both spatial and temporal domain of a video sequence.
- The quantitative and qualitative results on the SA dataset show superior performance compared to other state-of-the-art approaches.

Questions and Feedback