IEICE TRANSACTIONS

on Fundamentals of Electronics, Communications and Computer Sciences

DOI:10.1587/transfun.2024IMP0005

Publicized:2025/01/09

A PUBLICATION OF THE ENGINEERING SCIENCES SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER Special Section on Image Media Quality

Traffic Accident Prediction Without Object Detection for Single-Vehicle Accidents

Kazuki HARADA[†a)], **Yuta MARUYAMA**[†], **Tomonori TASHIRO**[†], ***Nonmembers*, *and*** **Gosuke OHASHI**[†], ***Member***

SUMMARY Recently, the research on traffic accident prediction models via deep learning has attracted significant attention. Many recent high-accuracy accident prediction models rely on bounding boxes obtained from object detection, which cannot predict single-vehicle accidents with a high fatality rate because of their structure. This paper proposes a model that predicts single-vehicle accidents by estimating the probability of accident occurrence at the frame level. The proposed model integrates depth and segmentation information along with RGB images and optical flow information to enhance prediction accuracy. To validate the effectiveness of the proposed model in single-vehicle accident scenarios, this study constructed a CARLA Accident Dataset using a driving simulator and a dataset containing only single-vehicle accident scenes selected from the Detection of Traffic Anomaly dataset. The proposed model demonstrated high accuracy in the investigated datasets, indicating its effectiveness in predicting single-vehicle accidents.

key words: *deep learning, traffic accident prediction, single-vehicle accident, dashcam, driving simulator*

1. Introduction

Autonomous driving technology has rapidly advanced recently; however, concerns regarding traffic accidents remain. Traffic accidents are broadly categorized as "person and vehicle," "multiple vehicles," and "single vehicle" accidents. In Japan, single-vehicle accidents in Japan accounted for 4% of all accidents; however, they represented 28% of fatal accidents in 2023 [1]. Of all the accident types, single-vehicle accidents have the highest fatality rate. This trend is observed worldwide, with single-vehicle accidents comprising 56% (2021) of fatal accidents in the United States [2] and 33% (2020) in the European Union [3]. Traffic accident prediction models using deep learning are being actively researched to prevent the social and economic losses incurred during accidents and realize a safe automated driving society. Most existing high-accuracy models for accident prediction [4]–[13] estimate the probability of accidents at the object level based on object detection results. We have also proposed such a model [13]. However, such models cannot predict single-vehicle accidents because of the absence of detectable objects. Moreover, although certain accident datasets contain single-vehicle accidents, the number of single-vehicle accidents in the real world is fewer than that of person-and-vehicle and multiple-vehicle accidents. Consequently, the number of single-vehicle accident scenarios in datasets must be increased. Thus, this study constructed a dataset that includes single-vehicle accident scenes and proposed an accident prediction model based on deep learning that can address single-vehicle accidents. The study contributions are as follows:

- The proposed model estimates the probability of accident occurrence at the frame level without object detection to predict single-vehicle accidents.
- Traffic accident datasets that include single-vehicle accidents were constructed to validate the effectiveness of the proposed model.
- This study demonstrates that adding depth and segmentation information, along with RGB and optical flow information, can improve the accuracy of accident predictions.

2. Related Works

2.1 Datasets Available for Traffic Accident Prediction

Datasets containing accident scenes are essential for traffic accident prediction via deep learning. Table 1 shows the trends of available datasets for traffic accident prediction. Most existing accident datasets comprise real accident videos obtained from video-sharing platforms. These datasets include the temporal annotations of accidents. A few contain spatial annotations and other information.

The Tokyo University of Agriculture and Technology (TUAT) near-miss database [14] and Near-Miss Incident Database (NIDB) [15] focus on near-miss scenes where accidents almost occurred. The Dashcam Accident Dataset (DAD) [4] is the first available dataset for traffic accident prediction; it comprises actual accident scenes. Several subsequent datasets have been proposed [5]–[8], [16]–[20]. Car Crash Dataset (CCD) [6] and Multi-Modal Accident video Understanding (MM-AU) [16] include temporal and spatial annotations, as well as textual data explaining the reasons for the accidents. DADA-2000 [17] includes the visual attention of drivers, rendering it the most extensive dataset for predicting the attention of drivers during accidents. In addition, VIENA^2 [18] and GTACrash [19] contain synthetic accident videos rendered from video games.

Although the aforementioned datasets contain videos from vehicle-mounted dashcams, Car Accident Detection and Prediction (CADP) [7] features recordings from fixed

[†]The authors are with the Department of Electrical and Electronic Engineering, Shizuoka University, Hamamatsu-shi, 432-8561 Japan.

a) E-mail: harada.kazuki.20@shizuoka.ac.jp

Table 1 Trends of datasets for traffic accident prediction.

Dataset	Release	Videos	Annotation			Synthetic/Real	Ego vehicle		Single vehicle
			Temporal	Spatial	Others		Involved	Non involved	
TUAT near-miss database [14]	2010	207000+	✓		Text	R	✓		✓
DAD [4]	2016	1750	✓			R		✓	
NIDB [15]	2018	6200	✓			R	✓		
CADP [7]	2018	1416	✓	✓		R		✓	
VIENA^2 [18]	2018	15000	✓			S	✓		✓
DADA-2000 [17]	2019	2000	✓		Text, Att.	R	✓	✓	✓
GTACrash [19]	2019	11381	✓	✓		S	✓		
CCD [6]	2020	4500	✓		Text	R	✓	✓	
DoTA [8]	2023	4677	✓	✓		R	✓	✓	✓
ROL [5]	2023	1000	✓	✓		R	✓	✓	✓
DeepAccident [20]	2023	691	✓	✓		S	✓	✓	
MM-AU [16]	2024	11727	✓		Text	R	✓	✓	✓

Table 2 Trends of deep learning models for traffic accident prediction.

Model	Release	Component		
		Object detection	RNN	Optical flow
DSA [4]	2016	✓	✓	
Shah et al. [7]	2018	✓	✓	
Ustring [6]	2020	✓	✓	
FA [10]	2020	✓	✓	
DRIVE [26]	2021		✓	
DSTA [9]	2022	✓	✓	
CAP [27]	2022		✓	
FOL-Ensemble [8]	2023	✓	✓	✓
AM-Net [5]	2023	✓	✓	✓
THAT-Net [11]	2023	✓	✓	✓
Maruyama et al. [13]	2023	✓	✓	
DAA-GNN [12]	2024	✓	✓	
TTHF [29]	2024			

surveillance cameras. DeepAccident [20] is the first accident prediction dataset designed for Vehicle-to-Everything (V2X) applications. It contains accident scenes that are automatically collected using the Car Learning to Act (CARLA) [21] simulator.

However, the existing datasets present several challenges. First, the lack of a unified standard for temporal annotations makes it difficult to compare evaluations across datasets. Second, certain datasets contain a mix of ego-vehicle-involved and non-involved accidents. Thus, the use of such datasets may affect model performance. Finally, owing to the scarcity of single-vehicle accident scenes, the available datasets are insufficient for single-vehicle accident prediction.

2.2 Deep Learning Models for Traffic Accident Prediction

Several deep learning models have been proposed for traffic accident prediction. Table 2 summarizes the trends of accident prediction models. In particular, models using Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) [22] and Gated Recurrent Unit (GRU) [23], have been proposed to consider long-term temporal relationships [4]–[7], [9]–[13], [26], [27].

The Attention-Guided Multistream Feature Fusion Network (AM-Net) [5] and Two-Layer Hidden State Aggregation Based Two-Stream Network (THAT-Net) [11] use optical flow, which refers to the motion vector of each pixel between two consecutive frames in a video sequence. Both models incorporate optical flow estimation models, such as Recurrent All-Pairs Field Transforms (RAFT) [24] to generate optical flow images from input images. This can facilitate the extraction of motion features of objects and the estimation of the probability of accidents.

THAT-Net, UString [6], and Dynamic Attention Augmented Graph Neural Network (DAA-GNN) [12] use detected objects and their features to construct graphs and learn using GNNs [25]. In addition to RGB images, Deep ReInforced accident anticipation with Visual Explanation (DRIVE) [26], Cognitive Accident Prediction (CAP) [27], Future Object Localization (FOL) [8], [28], and Text-Driven Traffic Anomaly Detection with Temporal High-Frequency Modeling (TTHF) [29] use additional information as input. DRIVE predicts accident probability and saliency maps through reinforcement learning. However, datasets rarely provide the ground truth of gaze data, which are essential for predicting saliency maps. Maruyama et al. [13] enhanced accident prediction accuracy and proposed a method for visualizing the risk factors using the divergence between visual attention and focus of expansion (FOE). CAP uses the Transformer [30] mechanism by encoding image frames with Patch Embedding and text data describing accident situations with Bidirectional Encoder Representations from Transformers (BERT) [31], and inputs them into a Multi-Head Attention. It subsequently estimates accident probability using a Graph Convolutional Network (GCN) [32] and GRU. FOL predicts object trajectories and detects traffic accidents by identifying deviations from the vehicle's expected odometry. FOL-Ensemble [8] combines FOL-STD [28], which uses predicted bounding box similarities, with AnoPred [33]. TTHF detects traffic anomalies by leveraging contrastive learning between video clips and textual prompts via pre-trained Contrastive Language-Image Pre-training (CLIP) [34], extracting visual context and capturing dynamic scene changes without requiring sequential structures like RNNs.

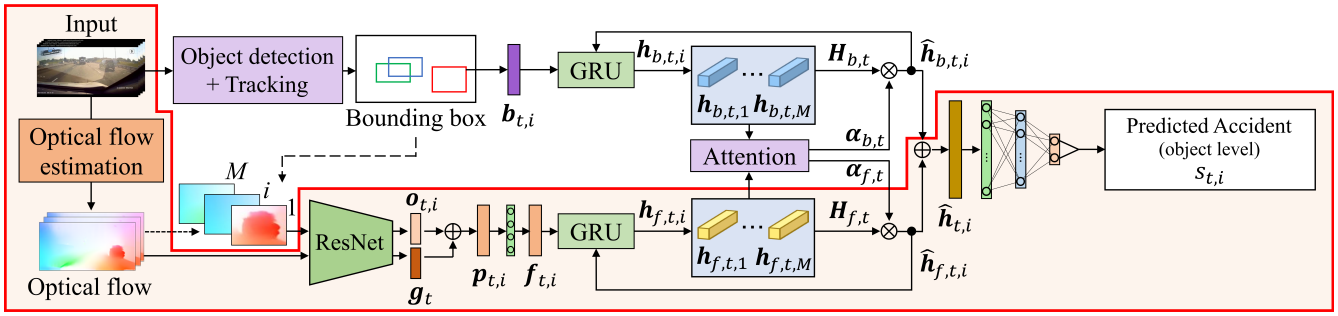


Fig. 1 Architecture of AM-Net.

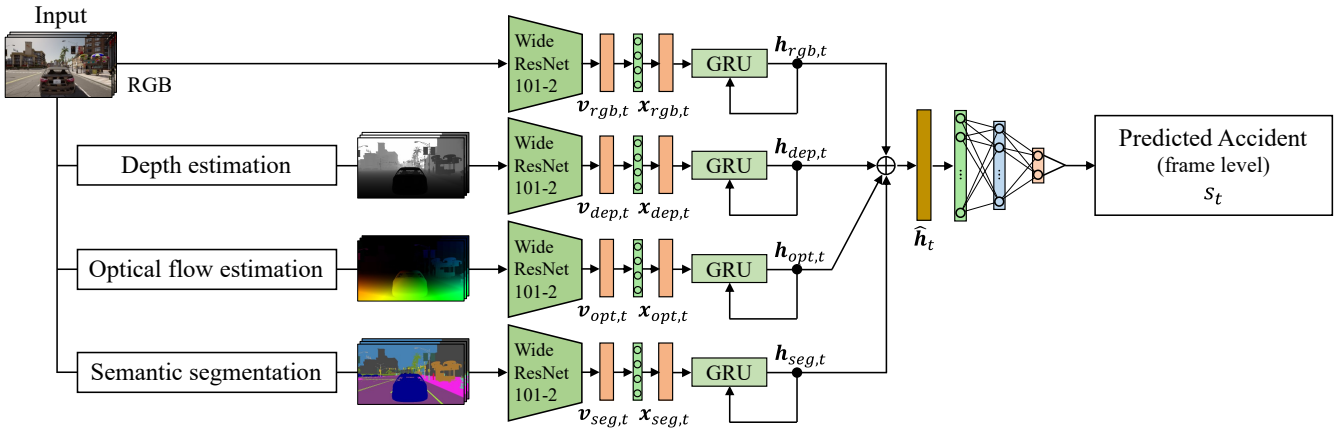


Fig. 2 Architecture of the proposed model.

Many of the aforementioned high-accuracy accident prediction models rely on object detection; thus, they encounter difficulties in estimating single-vehicle accidents. This is because colliding objects are often difficult to detect and, in certain cases, may not even exist. Furthermore, the existing models that do not employ object detection suffer the challenge of dependency on datasets, as they require the ground truth of saliency maps or textual explanations of accident situations.

3. Method

3.1 Architecture Overview

The existing accident prediction models that estimate object-level accident occurrence probabilities cannot handle single-vehicle accidents. Thus, we propose a model that estimates frame-level accident occurrence probabilities independent of bounding boxes obtained from object detection. The proposed model was constructed by referencing the parts of AM-Net [5] proposed by Karim et al. that do not use object detection information. Figures 1 and 2 show the architectures of AM-Net and that of the proposed model, respectively. AM-Net employs object detection to estimate object-level accident occurrence probabilities; thus, it cannot adequately predict single-vehicle accidents. Contrarily, the proposed model estimates frame-level accident occurrence probabilities, enabling the prediction of single-vehicle accidents. In addition, the model generates depth, optical flow, and segmentation images from RGB images and extracts features from each image type.

The integration of data from RGB images with additional layers, such as optical flow and segmentation, has proven to be highly effective in visual attention research [35], [36]. This method is inspired by the hierarchical nature of human visual processing, in which the brain analyzes visual information from basic to complex stages across different cortical areas [37]. By inputting features from various sources, including texture from RGB images, depth from depth images, motion from optical flow images, and object identification from segmentation images, the proposed model enhances its ability to recognize and interpret complex scenes. As shown in Sec. 2, although models exist that use optical flow [5], [11], no models leverage depth and semantic segmentation. To our knowledge, a configuration combining RGB, optical flow, depth, and segmentation has not been previously proposed.

3.2 Additional Information from RGB Images

The existing models frequently use RGB and optical flow images as inputs, similar to the proposed model, which uniquely incorporates depth and segmentation images. Most accidents involve a form of collision, and depth images can facilitate a spatial understanding of how objects approach ego

vehicles. Segmentation images provide clearer information on object placement than RGB images. The proposed model obtained these images using Metric3DV2 [38] for depth estimation, GMFlow [39], [40] for optical flow estimation, and InternImage [41] for segmentation.

3.3 Feature Extraction and Aggregation

First, the input images are resized to 224×224 pixels, and features are extracted using Wide ResNet101-2 [42] pretrained from ImageNet, excluding its fully connected layers. This process yields a feature vector, $\boldsymbol{v}_t (\in \mathbb{R}^{2048})$, in frame t. The feature vector is transformed into a lower-dimensional feature vector, $\boldsymbol{x}_t (\in \mathbb{R}^{256})$, using a fully connected layer, which is subsequently fed into a GRU to obtain a hidden representation, $\boldsymbol{h}_t (\in \mathbb{R}^{256})$. This procedure is performed for each of the four data sources, resulting in four hidden representations, $\boldsymbol{h}_{rgb,t}$, $\boldsymbol{h}_{dep,t}$, $\boldsymbol{h}_{opt,t}$, and $\boldsymbol{h}_{seg,t}$ which are concatenated to $\hat{\boldsymbol{h}}_t (\in \mathbb{R}^{1024})$. Thereafter, $\hat{\boldsymbol{h}}_t$ is transformed into two dimensions using two fully connected layers, and the probability of accident s_t is calculated using the softmax function. The weights of the fully connected layers and GRU are shared among the four data sources.

4. Experiments

4.1 Datasets

To validate the effectiveness of the proposed model, we use two newly constructed datasets along with DoTA.

4.1.1 Dataset Construction Using CARLA Simulator

We constructed an original dataset called the "CARLA Accident Dataset (CAD)" using CARLA [21], an open-source driving simulator for autonomous driving research. CAD includes scenarios involving single-vehicle accidents and provides ground truth images for depth, optical flow, and segmentation obtained through simulations. This dataset can be employed to evaluate the effectiveness of the proposed architecture independent of the accuracy of the estimation models. Moreover, it enables a precise assessment of the performance of the model in accident scenarios, free from potential distortions caused by data estimation errors.

In CARLA, a three-dimensional simulation of real urban environments enables the reproduction of various driving conditions, and various simulated sensors are available. Inducing accidents during simulation is necessary to construct an accident dataset using CARLA. Thus, accidents are generated through autopilot and manual driving, and accident scenes are collected. Table 3 lists the conditions for dataset construction. On autopilot, the ego vehicle is programmed to engage in hazardous driving behaviors, allowing the simulator to automatically generate accidents. Conversely, during manual driving, a Logicool G29 steering wheel is used for the hands-on manipulation of steering, acceleration, and braking, capturing data that more closely simulates actual driving scenarios. Furthermore, this approach intentionally produces accident scenes that are difficult to generate on autopilot, ensuring a substantial collection of single-vehicle accident scenes. The simulations use three different maps and operate under diverse weather and light conditions. Outputs from the RGB, depth, optical flow, and semantic segmentation sensors are recorded for each frame. A dataset scene comprises 100 frames, ending with the frame where a collision occurs, resulting in 400 images per scene, including the outputs from the four sensors across 100 frames.

(a) Scenes of autopilot

(b) Scenes of manual driving

Fig. 3 Examples of CAD accident scenes.

In deep learning, it is important to split the data into training and evaluation sets to prevent overfitting. Initially, the model is trained using training data, and the accuracy of the resulting model is evaluated using evaluation data. The collected 600 scenes were divided into 480 for training and 120 for evaluation. In addition, simply distributing the entire dataset randomly can lead to the under-representation of rare scenes in the evaluation set. Thus, after categorizing all the videos by accident type, they were randomly distributed to approximately maintain the 4:1 training-to-evaluation ratio.

Many of the existing accident datasets comprise videos collected from video-sharing platforms, frequently including edited, processed content, such as that with subtitles and zooming in on vehicles at the moment of accidents. When constructing datasets using such videos, caution is necessary because of the potential for copyright infringement with unauthorized postings and the inclusion of privacy-sensitive information related to individuals involved in accidents. The use of driving simulators for dataset construction can circumvent these challenges, ensuring ethical compliance and

Table 3 Simulation conditions for dataset construction.

	Autopilot	Manual driving
CPU	Intel® Core i7-12700F CPU @2.10 GHz	Intel® Core i7-9700 CPU @2.10 GHz
CPU memory	16.0 GB	
GPU	NVIDIA GeForce RTX 3050 (8.0 GB)	NVIDIA GeForce RTX 3090 (24.0 GB)
Simulator	CARLA 0.9.14	
Map	Town07, Town10, Town12	
Image acquisition	100 frames/scene (fps = 10)	
Image size	640 × 360 pixel	1280 × 720 pixel
Autopilot setting	Ignoring red lights, other vehicles, pedestrians, and traffic signs: 100% Safe distance between vehicles: 0 m, Speed limit increase: 20%–100%	-
Steering wheel	-	Logicool G29

Table 4 Categories in the DoTA dataset.

Label	Category
ST	Collision with another vehicle that starts, stops, or is stationary
AH	Collision with another vehicle moving ahead or waiting
LA	Collision with another vehicle moving laterally in the same direction
OC	Collision with another oncoming vehicle
TC	Collision with another vehicle that turns into or crosses a road
VP	Collision between vehicle and pedestrian
VO	Collision with an obstacle in the roadway
OO	Out-of-control and leaving the roadway to the left or right
UK	Unknown

protecting privacy.

4.1.2 Reconstruction of DoTA Dataset

The Detection of Traffic Anomaly (DoTA) dataset [8] produced by Yao et al. is the first traffic anomaly video dataset that provides spatio-temporal annotations of risky objects in driving scenarios. It comprises 4677 dashcam video clips, each with a resolution of 1280 × 720 pixels, captured under various weather and light conditions.

Each DoTA video is annotated with the start and end times of the traffic anomaly and categorized into one of the nine types listed in Table 4. Among them, VO and OO are considered to be related to single-vehicle accidents. However, the classifications are based on the actions of the involved parties (ego vehicle or another vehicle) rather than on the type of colliding object, which is important for accident prediction. Thus, we selected only scenes involving single-vehicle accidents related to the ego vehicle from the 4677 dashcam video clips and reclassified them into three types: "collision with structures," "road departure," and "turnover." By clipping the footage prior to the end time of traffic anomaly in the DoTA videos, we standardized the scene composition between the dataset and CAD. We denoted the constructed dataset as "selected DoTA."

4.2 Experimental Conditions

Table 5 presents the experimental overview. Although CAD and selected DoTA are constructed by category and condition, the models were evaluated on overall performance

Table 5 Overview of the experiments.

Dataset		Ground truth	Additional images from RGB
Synthetic	CAD	Experiment 1	Experiment 2
Real	Selected DoTA	-	Experiment 3
	Full DoTA	-	Experiment 4

across diverse conditions, following common practice in traffic accident prediction research. Experiment 1 evaluates the architecture using ground truth images from CARLA for depth, optical flow, and segmentation images. Experiments 2, 3, and 4 use additional information from RGB images. Experiment 2 uses CAD, Experiment 3 uses selected DoTA, and Experiment 4 uses full DoTA for evaluation. Table 6 lists the experimental conditions.

The proposed model was implemented using PyTorch [43] on a system equipped with an Intel® Core i7-12700F CPU @2.10 GHz, 16.0 GB of memory, and an NVIDIA GeForce RTX 3050 GPU (8.0 GB). Wide ResNet101-2 functioned as the feature extractor, Metric3DV2 was used for depth estimation, GMFlow was used for optical flow estimation, and InternImage was used for segmentation. The training parameters were a batch size of 1, 20 epochs, and an initial learning rate of 0.001, with Adam as the optimizer and cross-entropy as the loss function. The evaluation metrics included the area under the receiver operating characteristic curve (AUC), precision, recall, F1 score, and time-to-accident (TTA). AUC represents the model's capacity to distinguish traffic accident occurrence, while F1 score, the harmonic mean of precision and recall, evaluates the overall model balance. The TTA indicates how early an accident can be predicted, with larger values indicating earlier predictions. This metric is defined as the largest time difference between τ and t, where t is the time when the accident probability, s_t, first exceeds the threshold, $\bar{s}$. The definition is given by the following formula:

$$\mathrm{TTA} = \max\{\tau - t \mid s_t \geq \bar{s}, 0 \leq t \leq \tau\} \tag{1}$$

The definition of τ varies by dataset. In CAD, τ denotes the collision time, specifically when the simulator's Collision Detector registers a collision. On the other hand, in DoTA, τ denotes the moment a traffic anomaly becomes unavoidable, specifically determined as the average time marked by three annotators with different driving experiences. This study adopted a commonly used threshold ($\bar{s} = 0.5$) to evaluate

Table 6 Experimental conditions.

	Experiment 1	Experiment 2	Experiment 3	Experiment 4
CPU	Intel® Core i7-12700F CPU @2.10 GHz			
CPU memory	16.0 GB			
GPU	NVIDIA GeForce RTX 3050 (8.0 GB)			
Dataset	CAD		Selected DoTA	DoTA
Train videos	480		192	3275
Test videos	120		48	1402
Image size	640 × 360 pixel			1280 × 720 pixel
Batch size	1			
Feature extraction	Wide ResNet101-2			
Depth estimation	-	Metric3DV2		
Optical flow estimation	-	GMFlow		
Segmentation	-	InternImage		
Epochs	20			
Initial learning rate	1.0×10^{-3}			
Optimizer	Adam			
Loss function	Cross Entropy			
Metrics	AUC, Precision, Recall, F1 score, TTA			

Table 7 Number of scenes of CAD.

Category		Autopilot	Manual driving
Multiple vehicles	Angle	21	14
	Sideswipe	3	17
	Rear-end	393	32
	Head-on	5	18
Single vehicle	Structures	4	84
	Road departure	0	3
	Turnover	0	0
Person and vehicle		1	5
Map	Town07	47	12
	Town10	292	114
	Town12	88	47
Light condition	Daytime	155	109
	Dawn/Sunset	13	8
	Nighttime	259	56
Weather condition	Sunny/Cloudy	299	126
	Rainy	128	47
Total		427	173

Table 8 Number of scenes of selected DoTA dataset.

Category		Number of scenes
Accident type	Structures	146
	Road departure	93
	Turnover	1
Light condition	Daytime	205
	Nighttime	35
Weather condition	Sunny/Cloudy	103
	Rainy	41
	Snowy	96
Slip	Yes	135
	No	105
Total		240

the accident prediction model.

4.3 Results

4.3.1 Dataset Construction

Table 7 presents the CAD construction results, with accident types used in Japan. Figure 3 shows examples of the collected accident scenes. On autopilot, 92% of the scenes captured were rear-end collisions, resulting in a biased dataset. The four scenes of collision with structures that occurred on autopilot were caused by malfunctions in the autonomous driving system of CARLA, specifically when the vehicle failed to correctly change lanes during lane reductions. Oppositely, manual driving enabled the intentional creation of various accident scenes beyond rear-end collisions. The "turnover" category refers to accidents where a vehicle turns over without colliding with an object. Although we captured scenes of vehicles turning over after colliding with an object, no scenes of "turnover" were collected because of the difficulty in simulating incidents where a vehicle turns over without colliding with an object.

Table 8 presents the construction results of the selected DoTA dataset, detailing 240 selected single-vehicle accident videos. The videos were categorized into training and evaluation sets at a ratio of 4:1, resulting in the construction of the first dataset focused on ego-vehicle-involved single-vehicle accidents. Notably, 40% of the scenes involved adverse weather conditions and slippery surfaces, indicating a bias in the accident scenarios represented.

4.3.2 Quantitative Evaluation

Tables 9, 10, 11, and 12 present the results of Experiments 1, 2, 3, and 4, respectively.

For Experiment 1, the evaluation results for each accident type indicated that the overall discriminatory ability of the model, as measured by the AUC, was high across all categories. It achieved an impressive score of 0.923, even for single-vehicle accidents, which are frequently predicted inadequately by the existing models via object-level evaluation. However, the AUC for "head-on collisions" was the lowest at 0.772, suggesting room for improvement in the discriminatory power of the model. For precision, "rear-end" recorded the highest value at 0.845, demonstrating that the prediction of the model was accurate. However, the precision was low for many categories, particularly "head-on,"

Table 9 Evaluation results of Experiment 1.

Accident type		AUC	Precision	Recall	F1 score	TTA [s]
Multiple vehicles	Angle	0.860	0.658	0.979	0.787	9.314
	Sideswipe	0.856	0.763	0.913	0.832	7.840
	Rear-end	0.848	0.845	0.943	0.891	8.661
	Head-on	0.772	0.423	1.000	0.595	8.140
Single vehicle	Structures	0.926	0.493	0.938	0.646	5.994
	Road departure	0.967	0.395	1.000	0.566	7.800
Person and vehicle		1.000	0.470	1.000	0.640	10.000
Non-single-vehicle accidents		0.836	0.805	0.946	0.870	8.653
Only single-vehicle accidents		0.923	0.485	0.942	0.640	6.094
Overall		0.863	0.772	0.946	0.850	8.269

Table 10 Evaluation results of Experiment 2.

Accident type		AUC	Precision	Recall	F1 score	TTA [s]
Multiple vehicles	Angle	0.735	0.681	0.727	0.703	8.086
	Sideswipe	0.864	0.859	0.820	0.839	6.560
	Rear-end	0.867	0.864	0.902	0.883	8.173
	Head-on	0.780	0.495	0.981	0.658	5.324
Single vehicle	Structures	0.899	0.556	0.917	0.692	5.324
	Road departure	0.998	0.833	1.000	0.909	3.700
Person and vehicle		1.000	0.505	1.000	0.671	9.700
Non-single-vehicle accidents		0.849	0.834	0.890	0.861	8.064
Only single-vehicle accidents		0.907	0.568	0.922	0.703	5.233
Overall		0.863	0.809	0.892	0.849	7.639

Table 11 Evaluation results of Experiment 3.

Accident type		AUC	Precision	Recall	F1 score	TTA [s]
Single vehicle	Structures	0.966	0.812	0.985	0.890	3.925
	Road departure	0.965	0.793	0.963	0.870	5.288
Overall		0.966	0.803	0.975	0.881	4.314

Table 12 Evaluation results of Experiment 4.

Method	Features	AUC
ConvAE [44]	Flow	0.663
ConvLSTMAE [45]	Flow	0.625
FOL-Ensemble [8]	RGB + Bbox + Flow + Ego	0.730
AM-Net [5]	Bbox + Flow	0.793
TTHF [29]	RGB	0.847
Ours	RGB + Depth + Flow + Seg	0.845

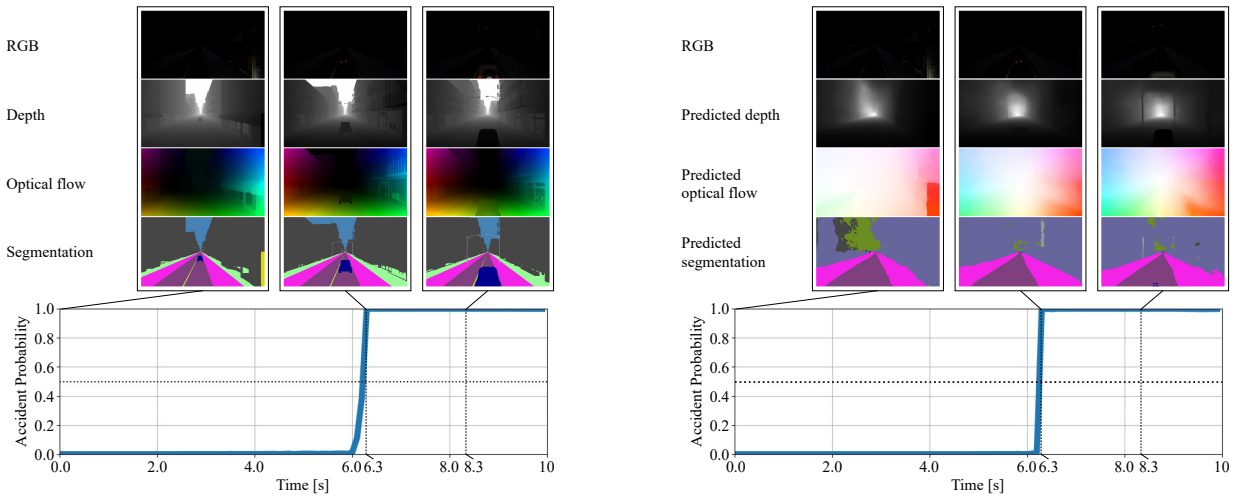
where it decreased to 0.423. This low precision led to many false positives, resulting in over-detection. An increase in over-detection can generate many unnecessary warnings in practical applications, potentially diminishing the usability of the model. For recall, "head-on," "road departure," and "person and vehicle" achieved perfect scores of 1.000, indicating that the model did not miss any accidents of these types. However, even with a high recall, the challenge of over-detection remains if the precision is low; thus, the balance between these metrics is critical. The F1 score, a metric that balances precision and recall, was the highest for "rear-end" at 0.891. However, "head-on" (0.595) and "road departure" (0.566) had low scores because of the very low precision. Although using a threshold of $\overline{s} = 0.5$ in this experiment may improve precision by maximizing the F1 score, it could potentially decrease the recall. In real-world accident prediction, where avoiding accidents and minimizing the impact on human lives is paramount, prioritizing the accuracy of the recall even at the expense of low precision might be more critical. Thus, it may not be necessary to use a threshold that maximizes the F1 score.

In Experiment 2, despite a reduction in the accuracy of the estimated depth, optical flow, and segmentation images of the model, the accident prediction accuracy was comparable to that achieved in Experiment 1, in which ground truth images were used.

In Experiment 3, the proposed model achieved high discriminatory accuracy across all the accident types.

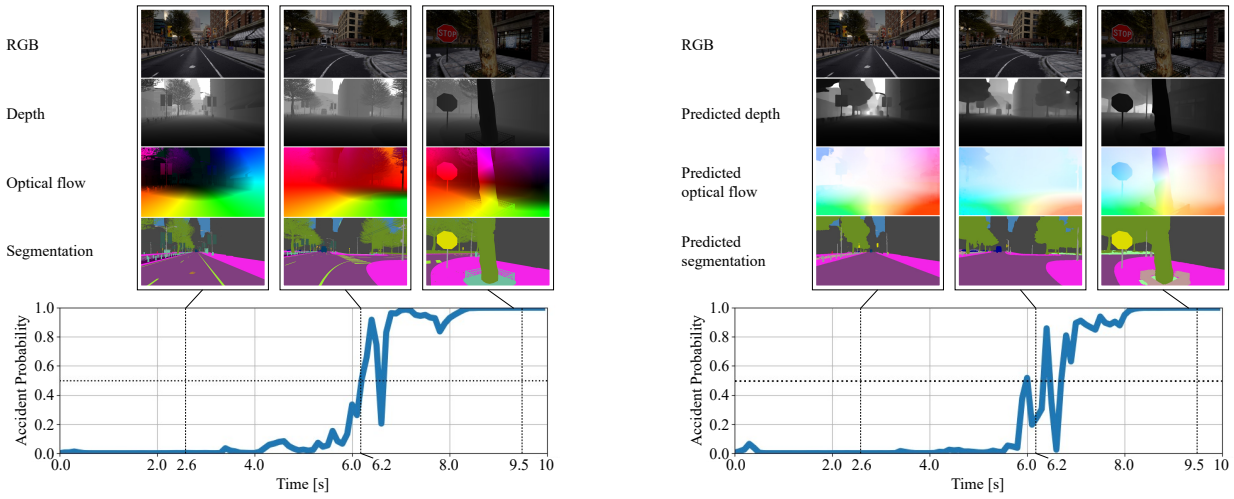
In Experiment 4, the proposed model achieved AUC comparable to the state-of-the-art model, TTHF. While TTHF requires textual prompts of accident scenes during training, our model uses RGB images alone, making it a simpler and more practical approach.

In Experiment 1, under the "person and vehicle" and "angle" conditions, as well as in Experiment 2 under "person and vehicle," very high TTA values were obtained. However, the precision was low, leading to over-detection. This high TTA probably resulted from the model predicting most frames in many scenes as having an accident probability ($s_t \geq \overline{s} = 0.5$). Thus, over-detection can lead to early accident predictions, thereby increasing the TTA, which may



(a) Result of Experiment 1 **(b)** Result of Experiment 2

Fig. 4 Example of multiple-vehicle accident prediction by the proposed model.



(a) Result of Experiment 1 **(b)** Result of Experiment 2

Fig. 5 Example of single-vehicle accident prediction by the proposed model.

render the TTA an inappropriate metric.

4.3.3 Qualitative Evaluation

The accident prediction results of the proposed model in Experiments 1 and 2 were demonstrated using specific scenes. First, Fig. 4 shows the prediction results for a multiple-vehicle accident scene. In this scenario, the ego vehicle rear-ends the vehicle ahead on a poorly visible single-lane road. The amount of information extracted in RGB images is limited; thus, it is challenging to identify hazards. However, using depth and segmentation images can provide additional information regarding the scene. The features extracted from these images enabled the model to successfully predict the occurrence of an accident 3.6 s in advance in Experiments 1 and 2. These results suggest that the model can extract features and predict accidents even in scenes that are not readily discernible to the human eye.

Figure 5 shows the prediction results for a single-vehicle accident scene. In this scene, around frame 60, the ego vehicle suddenly turns right. In frame 99, it collides with a tree on the sidewalk, which would be difficult for an object detection model to accurately detect. In Experiment 1, the proposed model detected a traffic anomaly in frame 62 and predicted the accident 3.7 s before it occurred. In Experiment 2, the model predicted the accident 3.9 s beforehand. In these scenes, the optical flow images capture the movement of the ego vehicle, and the segmentation images detect lane departure, significantly contributing to the accident prediction.

4.3.4 Ablation Study

In Experiment 1, ablation studies of the input sources for the proposed model were conducted to demonstrate how additional information enhances accuracy in accident prediction.

Table 13 Evaluation results of ablation study.

Model	RGB	Depth	Optical flow	Segmentation	AUC	Precision	Recall	F1 score
1	✓				0.833	0.812	0.879	0.844
2		✓			0.840	<u>0.854</u>	0.753	0.801
3			✓		0.811	0.669	<u>0.988</u>	0.797
4				✓	0.838	**0.874**	0.749	0.807
5	✓	✓			0.851	0.852	0.801	0.826
6	✓		✓		0.854	0.781	0.914	0.843
7	✓			✓	0.831	0.817	0.878	<u>0.846</u>
8		✓	✓		0.856	0.753	0.935	0.834
9		✓		✓	**0.864**	0.853	0.828	0.840
10			✓	✓	0.857	0.779	0.904	0.837
11	✓	✓	✓		0.857	0.670	**0.997**	0.801
12	✓	✓		✓	0.850	0.795	0.897	0.843
13	✓		✓	✓	0.855	0.800	0.894	0.844
14		✓	✓	✓	0.856	0.703	0.965	0.813
15	✓	✓	✓	✓	<u>0.863</u>	0.772	0.946	**0.850**

Table 13 presents the results; the most accurate metrics are highlighted in bold, and the second most accurate metrics are underlined. The experimental conditions were the same as those described in Table 6. Using ground truth images obtained from CARLA as ideal inputs enabled the evaluation of the contribution of each data source to predictive accuracy.

Initially, the performance of models 1–4, which each use a single data source, was dependent on the characteristics of the specific source. Model 1, which uses only RGB images, exhibited a relatively good balance between precision and recall, resulting in a remarkable F1 score. However, Model 3, which solely relies on optical flow images, exhibited low precision. Models 2 and 4, using only depth and segmentation images, respectively, exhibited a low recall.

Models 5–15, which integrate multiple data sources, generally achieved a higher AUC and recall than the single-source models. This suggested that the complementary information provided by multiple data sources contributed to an overall better performance. However, the precision of these models declined, particularly in those including optical flow images, where a trend toward a low precision was observed. For instance, Model 11 achieved a high recall, but its precision was relatively low at 0.670, indicating occurrences of over-detection. This suggested that although using multiple data sources increased the available information for accident prediction, it increased the potential risk of over-detection.

Comparing Models 1 to 15 from the perspectives of AUC and F1 score, although Model 9, using depth and segmentation sources, achieved the highest AUC, it ranked seventh in F1 score, indicating an unbalanced performance. Conversely, Model 15, using all four data sources, achieved the highest F1 and the second-highest AUC, thus representing the most balanced and high-performing model.

Considering the commonly used combination of RGB and optical flow images in existing accident prediction models (Model 6), the addition of depth images (Model 11), segmentation images (Model 13), or both (Model 15) improved the AUC. Thus, the ablation studies showed that by integrating multiple data sources, the proposed accident prediction model can capture features of accidents that are undetectable by a single data source, thereby enhancing predictive accuracy.

5. Conclusion

This study focused on the limitations of the existing high-accuracy models for predicting single-vehicle accidents. We propose a model that predicts such accidents without object detection. By incorporating depth and segmentation information into the model, we improved its recognition capability for accident scenes. To verify the effectiveness of the proposed model in predicting single-vehicle accidents, we constructed a CAD using a driving simulator and a dataset containing only single-vehicle accident scenes selected from the DoTA dataset. When applied to the datasets, the proposed model achieved high accuracy. The study findings show that accident prediction models incorporating depth and segmentation information, along with RGB and optical flow information, can significantly improve accuracy.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP23K11117.

References

[1] National Public Safety Commission and National Police Agency, "Annual report (2023)," Statistics about Road Traffic, https://www.e-stat.go.jp/en/stat-search/files?tclass=000001020602&cycle=7&year=20230, accessed June 7. 2024.
[2] National Highway Traffic Safety Administration, "Traffic safety facts 2021 a compilation of motor vehicle traffic crash data," Traffic Safety Facts Publications, https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813527, accessed June 7. 2024.
[3] European Road Safety Observatory, "Facts and figures single vehicle crashes 2023," Facts and Figures, https://road-safety.transport.ec.europa.eu/system/files/2023-03/ff_single_vehicle_crashes_20230221.pdf, accessed June 7. 2024.
[4] F.H. Chan, Y.T. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," Proc. 13th Asian Conf. Computer Vision, Taipei, Taiwan, vol.10114, pp.136–153, Nov. 2016. DOI:10.1007/978-3-319-54190-7_9

[5] M.M. Karim, Z. Yin, and R. Qin, "An attention-guided multistream feature fusion network for early localization of risky traffic agents in driving videos," IEEE Trans. Intell. Veh., vol.9, no.1, pp.1792–1803, 2024. DOI:10.1109/TIV.2023.3275543

[6] W. Bao, Q. Yu, and Y. Kong, "Uncertainty-based traffic accident anticipation with spatio-temporal relational learning," Proc. 28th ACM Int. Conf. Multimedia, Virtual Event / Seattle, USA, pp.2682–2690, Oct. 2020. DOI:10.1145/3394171.3413827

[7] A.P. Shah, J.B. Lamare, T.N. Anh, and A.G. Hauptmann, "CADP: a novel dataset for cctv traffic camera based accident analysis," Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill., Auckland, New Zealand, pp.1–9, Nov. 2018. DOI:10.1109/AVSS.2018.8639160

[8] Y. Yao, X. Wang, M. Xu, Z. Pu, Y. Wang, E.M. Atkins, and D.J. Crandall, "DoTA: unsupervised detection of traffic anomaly in driving videos," IEEE Trans. Pattern Anal. Mach. Intell., vol.45, no.1, pp.444–459, 2023. DOI:10.1109/TPAMI.2022.3150763

[9] M.M. Karim, Y. Li, R. Qin, and Z. Yin, "A dynamic spatial-temporal attention network for early anticipation of traffic accidents," IEEE Trans. Intell. Transp. Syst., vol.23, no.7, pp.9590–9600, 2022. DOI:10.1109/TITS.2022.3155613

[10] M. Fatima, M.U.K. Khan, and C.M. Kyung, "Global feature aggregation for accident anticipation," Proc. 25th Int. Conf. Pattern Recognition, Virtual Event / Milan, Italy, pp.2809–2816, Jan. 2020. DOI:10.1109/ICPR48806.2021.9412338

[11] W. Liu, T. Zhang, Y. Lu, J. Chen, and L. Wei, "THAT-Net: two-layer hidden state aggregation based two-stream network for traffic accident prediction," Inf. Sci., vol.634, pp.744–760, 2023. DOI:10.1016/J.INS.2023.03.075

[12] W. Song, S. Li, T. Chang, K. Xie, A. Hao, and H. Qin, "Dynamic attention augmented graph network for video accident anticipation," Pattern Recognit.," vol.147, p.110071, 2024. DOI:10.1016/J.PATCOG.2023.110071

[13] Y. Maruyama and G. Ohashi, "Accident prediction model using divergence between visual attention and focus of expansion in vehicle-mounted camera images," IEEE Access, vol.11, pp.140116–140125, 2023. DOI:10.1109/ACCESS.2023.3339855

[14] Smart Mobility Research Center, "Smart mobility research center," Tokyo University of Agriculture and Technology, https://web.tuat.ac.jp/ smrc/, accessed June 7. 2024.

[15] H. Kataoka, T. Suzuki, S. Oikawa, Y. Matsui, and Y. Satoh, "Drive video analysis for the detection of traffic near-miss incidents," Proc. 2018 IEEE Int. Conf. Robotics and Automation, Brisbane, Australia, pp.3421–3428, 2018. DOI:10.1109/ICRA.2018.8460812

[16] J. Fang, L.L. Li, J. Zhou, J. Xiao, H. Yu, C. Lv, J. Xue, and T.S. Chua, "Abductive ego-view accident video understanding for safe driving perception," arXiv:2403.00436, 2024. DOI:10.48550/ARXIV.2403.00436

[17] J. Fang, D. Yan, J. Qiao, J. Xue, H. Wang, and S. Li, "DADA-2000: can driving accident be predicted by driver attention? analyzed by a benchmark," Proc. 22nd IEEE Intell. Transp. Syst. Conf., Auckland, New Zealand, pp.4303–4309, Oct. 2019. DOI:10.1109/ITSC.2019.8917218

[18] M.S. Aliakbarian, F.S. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "VIENA^2: a driving anticipation dataset," Proc. 14th Asian Conf. Computer Vision, Perth, Australia, vol.11361, pp.449–466, Dec. 2018. DOI:10.1007/978-3-030-20887-5_28

[19] H. Kim, K. Lee, G. Hwang, and C. Suh, "Crash to not crash: learn to identify dangerous vehicles using a simulator," Proc. 33rd AAAI Conf. Artif. Intell., Honolulu, USA, pp.978–985, Feb. 2019, DOI:10.1609/AAAI.V33I01.3301978

[20] T. Wang, S. Kim, W. Ji, E. Xie, C. Ge, J. Chen, Z. Li, and P. Luo, "DeepAccident: a motion and accident prediction benchmark for v2x autonomous driving," Proc. 38th AAAI Conf. Artif. Intell., Vancouver, Canada, pp.5599–5606, Feb. 2024. DOI:10.1609/AAAI.V38I6.28370

[21] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: an open urban driving simulator," Proc. 1st Annu. Conf. Robot Learning, Mountain View, USA, vol.78, pp.1–16, Nov. 2017.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol.9, no.8, pp.1735–1780, 1997. DOI:10.1162/NECO.1997.9.8.1735

[23] K. Cho, B.V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," Proc. 2014 Conf. Emp. Meth. Nat. Lang. Process., Doha, Qatar, pp.1724–1734, Oct. 2014. DOI:10.3115/V1/D14-1179

[24] Z. Teed and J. Deng, "RAFT: recurrent all-pairs field transforms for optical flow," Proc. 16th Eur. Conf. Comput. Vis., Glasgow, UK, vol.12347, pp.402–419, Aug. 2020. DOI:10.1007/978-3-030-58536-5_24

[25] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," IEEE Trans. Neural Networks, vol.20, no.1, pp.61–80, 2009. DOI:10.1109/TNN.2008.2005605

[26] W. Bao, Q. Yu, and Y. Kong, "DRIVE: deep reinforced accident anticipation with visual explanation," Proc. IEEE/CVF Int. Conf. Comput. Vis., Montreal, Canada, pp.7599–7608, Oct. 2021. DOI:10.1109/ICCV48922.2021.00752

[27] 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:2212.09381, 2022. DOI:10.48550/ARXIV.2212.09381

[28] Y. Yao, M. Xu, Y. Wang, D.J. Crandall, and E.M. Atkins, "Unsupervised traffic accident detection in first-person videos," Proc. 2019 IEEE/RSJ Int. Conf. Intell. Robot. Syst., Macau, China, pp.273–280, Nov. 2019. DOI:10.1109/IROS40897.2019.8967556

[29] R. Liang, Y. Li, J. Zhou, and X. Li, "Text-driven traffic anomaly detection with temporal high-frequency modeling in driving videos," IEEE Trans. Circuits Syst. Video Technol., vol.34, no. 9, pp.8684–8697, 2024. DOI:10.1109/TCSVT.2024.3390173

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Adv. Neural Inf. Process. Syst. 30, Long Beach, USA, pp.5998–6008, Dec. 2017.

[31] J. Devlin, M.W. Chang, K. Lee, and K.N. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," Proc. 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol., Minneapolis, USA, pp.4171–4186, June 2018. DOI:10.18653/V1/N19-1423

[32] T.N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv:1609.02907, 2016. DOI:10.48550/arXiv.1609.02907

[33] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection – a mew baseline," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Salt Lake City, USA, pp.6536–6545, June 2018. DOI:10.1109/CVPR.2018.00684

[34] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, "Learning transferable visual models from natural language supervision," Proc. 38th Int. Conf. Mach. Learn., vol.139, pp.8748–8763, July 2021.

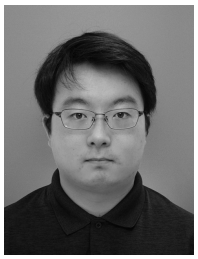
[35] Z. Hu, Y. Zhang, Q. Li, and C. Lv, "A novel heterogeneous network for modeling driver attention with multi-level visual content," IEEE Trans. Intell. Transp. Syst., vol.23, no.12, pp.24343–24354, 2022. DOI:10.1109/TITS.2022.3208004

[36] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "DADA: driver attention prediction in driving accident scenarios," IEEE Trans. Intell. Transp. Syst., vol.23, no.6, pp.4959–4971, 2022. DOI:10.1109/TITS.2020.3044678

[37] H. Yamamoto and Y. Ohtani, "Functional brain imaging and visual psychophysics," Japanese Journal of Optics, vol.33, no.2, pp.80–88, 2004.

[38] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen, "Metric3D v2: a versa-

tile monocular geometric foundation model for zero-shot metric depth and surface normal estimation," arXiv:2404.15506, 2024. DOI:10.48550/ARXIV.2404.15506
[39] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao, "GMFlow: learning optical flow via global matching," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., New Orleans, USA, pp.8121–8130, June 2022. DOI:10.1109/CVPR52688.2022.00795
[40] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, F. Yu, D. Tao, and A. Geiger, "Unifying flow, stereo and depth estimation," IEEE Trans. Pattern Anal. Mach. Intell., vol.45, no.11, pp.13941–13958, 2023. DOI:10.1109/TPAMI.2023.3298645
[41] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, X. Wang, and Y. Qiao, "InternImage: exploring large-scale vision foundation models with deformable convolutions," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Vancouver, Canada, pp.14408–14419, June 2023. DOI:10.1109/CVPR52729.2023.01385
[42] S. Zagoruyko and N. Komodakis, "Wide residual networks," Proc. Brit. Mach. Vis. Conf., York, UK, Sept. 2016.
[43] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, Pytorch: an imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32, Vancouver, Canada, no.721, pp.8026–8037, Dec. 2019.
[44] M. Hasan, J. Choi, J. Neumann, A.K. Roy-Chowdhury, and L.S. Davis, "Learning temporal regularity in video sequences," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Las Vegas, USA, pp.733–742, June 2016. DOI:10.1109/CVPR.2016.86
[45] Y.S. Chong and Y.H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," Proc. 14th Int. Symp. Neural Networks, Sapporo, Hakodate, and Muroran, Japan, vol.10262, pp.189–196, June 2017. DOI:10.1007/978-3-319-59081-3_23



Kazuki Harada received the B.E. degree from the Department of Electrical and Electronic Engineering, Faculty of Engineering, Shizuoka University, Hamamatsu, Japan, in 2024. He is currently with the Department of Electrical and Electronic Engineering, Graduate School of Integrated Science and Technology, Shizuoka University. His research interests include deep learning, artificial intelligence, and computer vision.

Yuta Maruyama received the B.E. and M.E. degrees from Shizuoka University, Hamamatsu, Japan, in 2022 and 2024, respectively. He is currently with Panasonic Connect Corporation. His research interests include deep learning, artificial intelligence, and computer vision.

Tomonori Tashiro received the B.E., M.E., and D.E. degrees from Utsunomiya University, Tochigi, Japan, in 2009, 2011, and 2014, respectively, and the Ph.D. degree from the University of Eastern Finland, Joensuu, Finland, in 2016. He was a Postdoctoral Researcher with Utsunomiya University, from 2014 to 2016, and Yamagata University, from 2016 to 2021. He was a Senior Researcher with the Industrial Research Institute of Shizuoka Prefecture, from 2021 to 2023. He is currently an Assistant Professor with the Department of Electrical and Electronic Engineering, Shizuoka University. His research interests include cognitive, vision, and color science.

Gosuke Ohashi received the B.E., M.E., and D.E. degrees from Keio University, Yokohama, Japan, in 1992, 1994, and 1997, respectively. He has been an Assistant Professor, since 1997. He was a Visiting Researcher with the University of California, Santa Barbara, from 2003 to 2004. He is currently a Professor with the Department of Electrical and Electronic Engineering, Shizuoka University. His research interests include image processing, computational vision, and visual perception.