Journal of Computational Design and Engineering, 2023, **10**, 1579–1601
DOI: 10.1093/jcde/qwad063
Advance access publication date: 28 June 2023
Research Article

Split liability assessment in car accident using 3D convolutional neural network

Sungjae Lee[1] and Yong-Gu Lee[1,2,*]

[1]School of Mechanical Engineering, Gwangju institute of science and technology (GIST), 123 Cheomdangwagi-ro, Buk-gu, 61005 Gwangju, Republic of Korea
[2]Artificial Intelligence Graduate School, Gwangju institute of science and technology (GIST), 123 Cheomdangwagi-ro, Buk-gu, 61005 Gwangju, Republic of Korea
*Correspondence: lygy@gist.ac.kr

Abstract

In a car accident, negligence is evaluated through a process known as split liability assessment. This assessment involves reconstructing the accident scenario based on information gathered from sources such as dashcam footage. The final determination of negligence is made by simulating the information contained in the video. Therefore, accident cases for split liability assessment should be classified based on information affecting the negligence degree. While deep learning has recently been in the spotlight for video recognition using short video clips, no research has been conducted to extract meaningful information from long videos, which are necessary for split liability assessment. To address this issue, we propose a new task for analysing long videos by stacking the important information predicted through the 3D CNNs model. We demonstrate the feasibility of our approach by proposing a split liability assessment method using dashcam footage.

Keywords: car accident, split liability assessment, 3D convolution

1. Introduction

Every once in a while, drivers get involved in car accidents. When they are lucky, the damage to the collision will be limited to the car and not to personal injuries. The relief can be quickly replaced by worries as drivers involved in the accident must undergo settlements to determine who will be liable for the damage and to what extent. The split liability assessment in car accidents refers to revisiting the moment of the collision and finding the cause of the accidents leading to the conclusion as to who is responsible for the property loss. The responsibility is often partially accredited to each driver and often mediated by car insurance agents. In cases where both parties do not come to a satisfying agreement, legal disputes will commence, finally leading to settlements made by the court. Each defendant will submit evidence proving they are less liable for the damage than the opponent. One of the most valuable pieces of evidence in a collision accident case is recorded video footage, often considered more credible than human witness testimony. In Korea, the use of car dashcams reached more than 90%. The same trend is true for other countries. Most car insurance companies discount for the insurance fees when the cars are equipped with dashcams. In some instances, even with available videos, it may not be sufficient to understand the underlying cause of a collision. In such scenarios, advanced simulation software can digitally reconstruct the scene using a combination of accident videos. Those include PC-Crash, LS-DYNA, and MADYMO (Steffan & Moser, 1996; Steffan *et al.*, 1999; Shang *et al.*, 2021), which use 3D simulations based on object movements to understand the detailed motions of the cars during the collision. From multiple video clues and operator inputs, the velocity and position of the objects involved in the collision can be simulated through a physics engine leading to a better understanding of the collision scene. Lawyers can use these reconstructed simulations to clearly understand the cause of the accidents and defend their clients. Courts can therefore use these simulations as supporting material for their final judgment of the split liability assessment.

Split liability assessment is a scorching topic, and TV shows and Youtube channels are devoted to this subject. Many lawyers appear on these channels to give their professional opinions, and audiences also give their personal opinations. Often, there can be a gap between traffic law and common sense. Some assessments can receive large flames from audiences because, in many cases, the disputes can be controversial, and opinions can be biased.

Legal matters are an area where Artificial Intelligence (AI) has the potential to eventually surpass humans. A notable example of this occurred during the Alpha Law competition in 2019, where AI and humans competed against each other. The competition was a legal advisory challenge between a team comprising of AI and a human and another team consisting of two humans. The team that had the assistance of AI emerged as the winner. This is because legal matters are primarily logical in nature, and AI excels in domains where logical rules dominate. With this in mind, we propose a novel AI system that can accurately classify a collision accident and assign it to a category based on prior similar accidents. Our AI model is based on a convolutional neural network (CNN), which has demonstrated remarkable success in computer vision applications such as classification, detection, and feature extraction. We trained the AI model using video clips labeled by experts in split traffic liability assessments, categorized by accident types. We validate the effectiveness of our approach using real-life dashboard camera footage of traffic accidents.

Received: February 1, 2023. **Revised:** June 25, 2023. **Accepted:** June 25, 2023

2. Related Work

According to the Korea Insurance Development Institute, there were more than two million reported car accidents in a year. Among them, 80% of the cases involved full liability, where one side of the car accident was deemed fully responsible for the entire damage. However, the remaining 20% were disputed. For these disputed cases, the General Insurance Association of Korea (GIAK), jointly formed by car insurance companies that are members of the association, provides settlements that are mandatorily followed by its members. These settlements are made to avoid lengthy court disputes, which can take an average of four months to 2 years. The GIAK also operates a portal (https://accident.knia.or.kr) where well-established categories exist in the form of a hierarchy to systematically classify the types of accidents and the possible split liability assessments. Currently, there are 109 cases of car-to-car accidents. At the high level, the classification is based on the movements of the cars involved in the accident: go-straight vs go-straight, go-straight vs left-turn, go-straight vs right-turn, and left-turn vs left-turn. At the mid-level, the number of roads that meet at the junction (three-way junction, four-way junction, etc.) and the type of movement of the car (changing lane, passing, stop-and-go) are used. Lastly, at the low-level, critical factors affecting the split ratio of liability, such as the speed of the car and traffic light conditions, are used. In summary, the high-to-mid level divides the type of split liability, and the low level determines the detailed liability scores. As described, various factors affect the criteria for determining the hierarchical nature of the classification. In this study, we follow these hierarchical rules and the final score of the split liability. Detailed accident classification is given in Appendix 1.

Analysing video clips of accidents is essential to enhance the accuracy of car accident classification using artificial intelligence. CNNs are effective in analysing images. CNNs are particularly adept at extracting representational features from input data compared to other methods (Lowe, 1999; Dalal & Triggs, 2005). This is because the network can extract common features among multiple images that belong to the same class. Consequently, CNNs have shown excellent performance in various fields, such as image classification (He *et al.*, 2016; Krizhevsky *et al.*, 2017; Houssein *et al.*, 2022) and object detection (Ren *et al.*, 2015; Redmon *et al.*, 2016; Choi *et al.*, 2022). However, 2D CNNs cannot be directly applied to video clips where time is added to space leading to three dimensions.

Recurrent neural networks (RNNs) are a model for dealing with time information. RNNs can store and handle information over time. The difference between 2D CNNs and RNNs is in the recurrent character. This means that there is an iterative connection that takes previous values back as input. However, RNNs have a vanishing gradient problem in which the influence of preceding information decreases over time (Bengio *et al.*, 1994). Therefore, a model that combines the spatial information analysis ability of 2D CNNs and the temporal information analysis ability of Long Short Term Memory (LSTM) exists. And there is research that uses Gated Recurrent Units (GRU) besides using LSTM. Although GRU performs similarly to LSTM, GRU has fewer gates and model parameters (Chung *et al.*, 2014; Ballas *et al.*, 2015).

Many studies simultaneously deal with spatiotemporal information. First of all, there is a spatiotemporal field that deals with the behavior of objects in the video, such as action recognition (Le *et al.*, 2022; Jiaxin *et al.*, 2021), car accident prediction (Yao *et al.*, 2019; Bao *et al.*, 2020; Adewopo *et al.*, 2022), and video segmentation (Wan *et al.*, 2022). Another area that deals with the temporal information is video super-resolution (Xiao *et al.*, 2021; Xiao *et al.*, 2022). In this area, joint spatiotemporal enhancement of videos resulting in higher resolution and increased number of frames are achieved. Various tasks dealing with spatiotemporal information have been created with the development of deep learning, as stated above. In addition, this field acquires supplementary information from temporal data to enhance network performance (Xiao *et al.*, 2023b). This is beneficial as it enables a broader feature space. However, it also requires more data for training the model. Therefore, research has been conducted on data generation through data simulation (Dosovitskiy *et al.*, 2017; Li *et al.*, 2022). However, challenges exists in generalization due to the differences between real and simulated data. Research is available to address this problem (Xiao *et al.*, 2023a).

In this research, a study was conducted to split liability assessment in car accidents. This research extends object behavior analysis into the spatiotemporal domain with the addition of legal judgment. First, we will describe related works through the development trend of action recognition, a representative task of spatiotemporal analysis from the next part.

Action recognition is a field where video clips are analysed to understand the actions (Laptev, 2003). SIFT3D and HOG-3D that extend SIFT and HOG belong to action recognition methods (Scovanner *et al.*, 2007; Klaser *et al.*, 2008; Patel *et al.*, 2018). Recently CNNs, with a large number of datasets, have also been used for action recognition which is mainly in the field of human behavior recognition (Zhou *et al.*, 2018; Fan *et al.*, 2019; Piergiovanni & Ryoo, 2019; Gowda *et al.*, 2021; Quddus *et al.*, 2021; Xu *et al.*, 2022).

Action recognition problem requires consideration of both temporal and spatial information in videos. First, there is a way to recognize behavior using 2D CNNs. Karpathy *et al.* (2014) conducted research that uses 2D CNNs to learn spatial information and analyses temporal information through fusion. There is an advantage of using various pre-trained models when using the 2D CNNs structure. However, there are limitations because the CNNs structure only learns spatial information. Also, it is not easy to process videos of various lengths because only fixed length is allowed. Wang *et al.* (2016) segmented the video and extracted one frame from each segment. Then, they used it as an input for 2D CNNs that share weights with the corresponding frame. Through this, they secured the performance of temporal segment networks (TSN), which classifies by looking at the whole video. While TSN is good at capturing short-term temporal information, it may struggle to model longer-term temporal dependencies due to segment-level processing.

Next, there is a way to analyse spatial information using 2D CNNs and input it into an LSTM model that handles temporal information. This is called the 2D CNNs + LSTM method. There is an advantage of using the backbone that is pre-trained with a vast amount of images using 2D CNNs as a means of transfer learning. Donahue *et al.* (2016) proceeded with action recognition by using the result of the 2D convolution of the input frame as input to LSTM. Through this, end-to-end learning was possible. However, long-term recurrent convolutional network (LRCN) combines CNNs and RNNs, but it may not be as effective in capturing temporal dynamics as 3D CNNs, which can extract both spatial and temporal features from video frames simultaneously. And these characteristics of LRCN make it more computationally expensive than other models, including 3D CNNs models.

3D CNNs extend 2D convolutional networks with the added dimension of time (Ji *et al.*, 2012). 3D CNNs have shown outstanding results in the field of action recognition and categorization. Tran *et al.* (2015) have proposed the optimal kernel size for 3D convo-

Table 1: Dataset related to traffic accident.

Dataset	Number of Videos	Positive	Source	Purpose
DAD* (Chan *et al.*, 2017)	1750	620	Youtube	Accident anticipation
A3D** (Yao *et al.*, 2019)	1500	1500	Youtube	Accident anticipation
CCD*** (Bao *et al.*, 2020)	4500	1500	Youtube	Accident anticipation
Ours (GLAD)	1267	1267	Youtube, KakaoTV etc	Accident anticipation, **split liability assessment in car accident**

Each dataset can be downloaded from the links provided below:
* https://aliensunmin.github.io/project/dashcam/
** https://github.com/MoonBlvd/tad-IROS2019
*** https://github.com/Cogito2012/CarCrashDataset

lutional networks with the best accuracy. Similar to the 2D CNNs (Simonyan & Zisserman, 2014), it has been shown that each axis having three lengths of floating numbers resulting in a $3 \times 3 \times 3$ kernel performed best. Furthermore, they have investigated the effect of using only a few numbers of layers for action recognition in terms of accuracy. They have concluded that the 3D CNNs performed well on time and space data, such as video clips. Having been motivated by their work, we applied 3D CNNs to assess the split liability in car accidents. The CNN-based network is used to utilize CNN's feature extraction. In other words, the CNN backbone's feature extraction capability is important. Hara *et al.* (2018) have experimented with the classification capability of 3D CNNs. They compared the relationship between the depth of the layers and the accuracy. They concluded that the accuracy was linearly proportional to the depth. They have also tested the use of ResNet. They were able to conclude that there exists little difference between the 3D CNNs and 3D CNNs modified with ResNet. They have found that the best results can be achieved by increasing the depth of the network in the case of 3D convolutions. Therefore, we also investigated the effect of increasing the number of layers in our experiments. Furthermore, as mentioned above, we changed the pretraining dataset in addition to the number of layers and checked the result in order to check the performance of the backbone.

As described above, action recognition is a research field that analyses video by adding time information to 2D image processing. Typical examples of action recognition datasets are Sports-1M and HMDB-51 (Kuehne *et al.*, 2011; Karpathy *et al.*, 2014). Deep learning-based action recognition research has been conducted based on the dataset. The data class of Sports-1M consists of clips about behavior such as track cycling, running and kayaking. HMDB-51 is a human motion recognition dataset consisting of jump, kick, and kiss. The characteristic of the two datasets is that clips are made up of short and repetitive simple actions. This is true of several datasets used in the study of action recognition (Soomro *et al.*, 2012; Caba et al., 2015; Abu-El-Haija *et al.*, 2016; Goyal *et al.*, 2017). Therefore, the clips used in the study serve the purpose of classifying repeated images into one class.

However, the goal of this study, the split liability assessment, differs from previous studies. The information must be accumulated over time in the long video. The split liability assessment in car accidents extracts information from images played over time and collects the information to perform the final classification. For example, in the case of a lane change accident, the information about lane change and the information about collision are needed. For this reason, this study proposed a voting method. The final classification was selected from the information from the long video by stacking the real-time information predicted through the 3D CNNs model.

There are also datasets related to car accidents. Research has been conducted to predict and prevent accidents. Accident prediction is a study that derives the probability of accidents happening while driving a car. The objective of this study is to prevent accidents from ever happening. Table 1 shows these related datasets. The 'Positive' column indicates the number of videos that contain accident scenes among all the videos. Dashcam Accident Dataset (DAD) aims to find the car accident, with 1750 videos and 620 positive samples. AnAn Accident Detection (A3D) is a dataset of car accidents recorded by dashcams, with 1500 videos and 1500 positive samples. The Car Crash Dataset (CCD) is used to predict when car accidents might occur by assessing anomalies in the movements of on-road participants, with 4500 videos and 1500 positive samples. While this dataset has the most data, it cannot be used for split liability assessment because it aims to anticipate accidents. Our dataset comprises 1267 videos, all of which contain accidents and are labeled as an accident type. We propose a new task that does not exist in the field, which sets our dataset apart from the others.

In summary, as previously described, this study proposed a new task to perform classification by collecting information over time from video rather than the simple classification of existing action recognitions. For the split liability assessment, the final classification must be performed by collecting important elements from the video. We propose a method of stacking the result values of 3D CNNs to find features relevant to assessing the split liability. The network's performance depends on the depth of the network layers. Therefore, we experimented with checking the performance of the feature extraction of the network by changing the number of layers and the pretraining dataset.

3. Experimental Configuration

3.1 Dataset

3.1.1 *Data collection*

We have collected the car accident video clips by web-crawling on Youtube, KakaoTV, and Bobaedream. Our objective was to collect worldwide car accident cases, and 1267 cases were gathered. Figure 1 illustrates featured data statistics that include resolutions, install location of the video camera (front/back), filmed time of day (daytime/night), FPS, and running time.

Many of the outsourced video clips are not used. Although generality could be achieved by collecting all possible video clips, in many cases, the number of samples having specific special conditions was extremely rare such that these particular conditions could interfere with the generality of the available cases. Therefore, we have excluded these types of video clips. We explain these samples in Section 3.1.2.

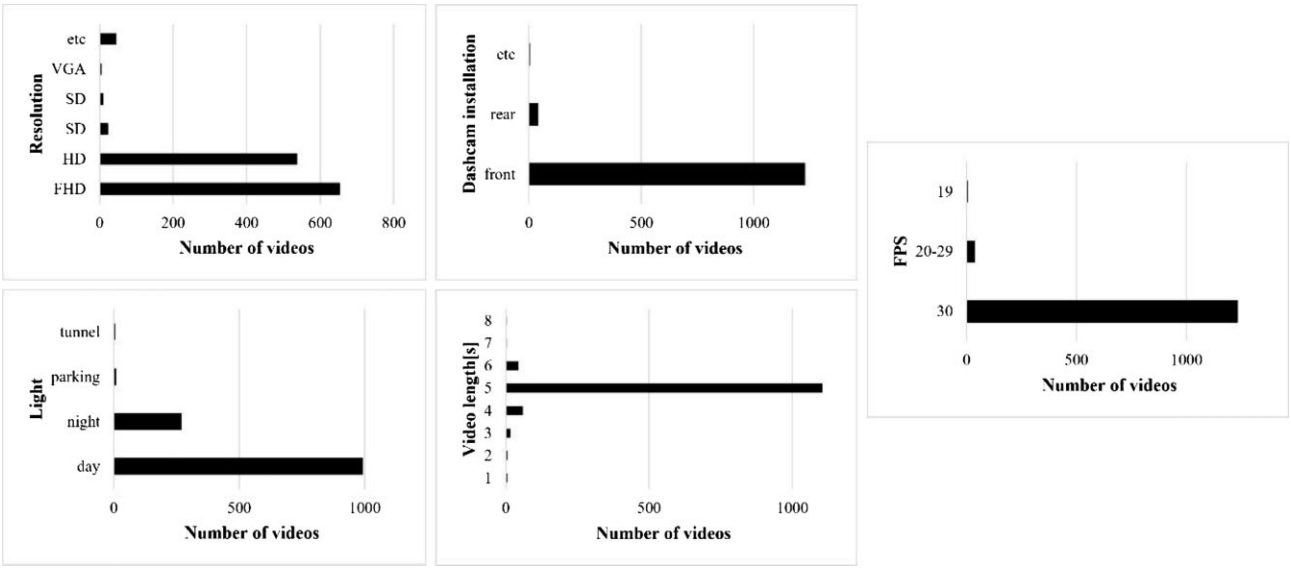


Figure 1: Features of collected videos.

The traffic laws were different because the collected video clips originated from various countries. For simplicity, we have excluded split liability assessment that requires the consideration of the traffic laws. The fact that some countries enforce keeping to the left and others enforcing the opposite was of no concern because this did not affect the split liability assessment.

3.1.2 *Pre-processing*

As explained above, we have reduced the data dimensions to assess the split liability in car accidents. We have preprocessed the data based on five areas. These include the length of the video clip, objects involved in the accident, the location of the dashcam, the time of the accident and the quality of the videos. First, the video clip's length was trimmed to 3 seconds just before the crash. The trim duration was experimentally chosen by comparing it with a five-second trim, which showed lower accuracy. We found that in most accidents, the first 2 seconds of the video (which accounts for 40% of the total 5-second running time) typically showed normal driving patterns without any actions such as lane changes. This reduced the possibility of excluding the fact that longer video clips would result in a certain bias of accident classification. Thus only 3 seconds of uniformly timed video clips were used. There were some special cases where video clips that were longer than 3 seconds possessed information that would affect the split liability assessment, but they were not used for simplicity. This simplification allowed us to achieve better network performances as to the reduction of dimensions of the features.

We only considered accidents involving car vs car. Accidents involving car vs motorcycle and car vs pedestrian collisions were not included due to the scarcity of available samples and to reduce the feature dimensions. We also excluded video clips obtained from rear-facing dash cams for similar reasons. Only video clips taken from the cars involved in the accident were used. In other words, we did not use video clips from a third-person perspective. We also excluded video clips with lower resolutions and images of low quality where objects were not clearly identifiable. Lastly, video clips that were taken by recording the playback and scenes that showed the aftereffect of the accident were also not used. Following Fig. 2 shows sample screenshots of video clips that were not used.

3.1.3 *Post-processing*

In this section, we explain the data labeling. To label the data, we provided the labelers with the accident type classification criteria specified by GIAK. We used written instructions and direct feedback to assist them in the labeling process. We selected good labelers by analysing the labeling results of several accident type cases in the screening test. These selected labelers then proceeded with the annotation process. This method is widely used in the development of deep learning datasets to ensure objectivity. Ouyang *et al.* (2022) secured the objectivity of the language model dataset using a similar approach.

The data labeling was based on 109 categories of car accidents which belonged to one of 11 steps (0–100%) of split liability proportions of the proponent. Figure 3 illustrates the statistics of 11 split liability ratios. The x-axis represents the split liability ratio of the proponent, and the y-axis represents the number of videos in that ratio. Those that have 20% or fewer proportions of the proponent's liability occupy more than 80% of total video clips. Those that pertain to 0% of the proponent's liability occupy more than 60%. We have found that the publicly collected video clips were quite biased toward the proponent. This is understandable because no one would post video clips that showed his or her fault. In other words, it is human nature to post video clips that are more favorable to themselves. It is also concerned that there is not much similarity among particular liability proportions. For example, for 0% of the proponent's liability cases, there existed categories such as wrong-way driving, collisions while parked, and rear collisions. We found these categories belonging to the same portion of the proponent's liability possessed no visual semantic similarity.

Thus, we have chosen to train the classifier based on the categories of the accident. We used 109 classes of car vs car accidents which is a subset of the full classes as described by GIAK. The clear benefit is that, in this way, a clear visual semantic similarity was achieved within the same class. The detail of the classes is described in the appendix 1. We evaluated the class for three items. It is 'Whether split liability assessment is checkable with my dash cam video', 'Necessity of opponent's dash cam video', and 'Necessity of rear installed dash cam video'. It can be seen that accident analysis is possible only with my dash cam video for 64 of 109 classes. And it can be seen that accident analysis is

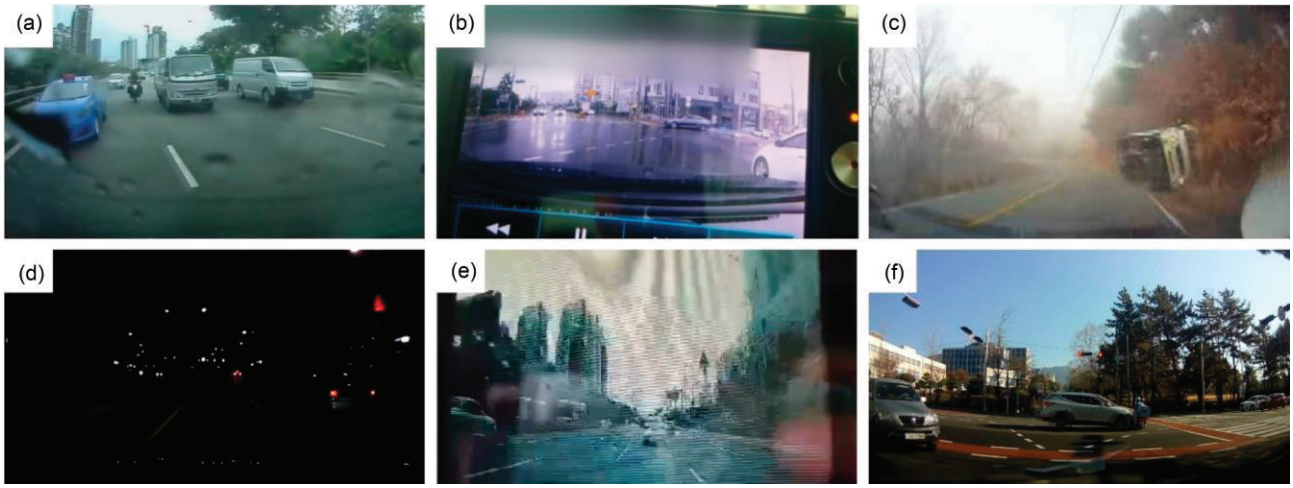


Figure 2: Excluded video clips: (a) rear-faced dash cams, (b) recorded on the playback, (c) aftereffect of the accident, (d) low light, (e) poor image quality and (f)third point of view.

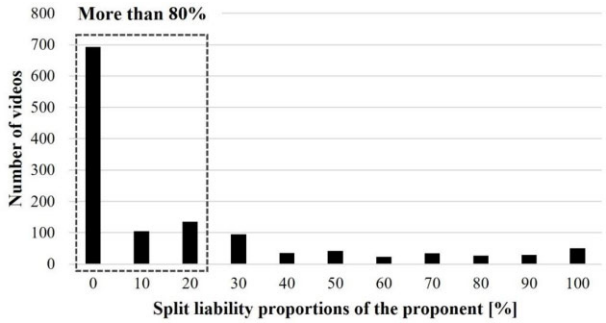


Figure 3: The count of video clips as a function of the proportion of the liability.

possible only with my video in some situations for 32 cases. Thus, out of the 109 classes in dash cam videos, 96 classes can be used for our purpose. Figure 4 shows the count of video clips available for each class. We can see that classes 252, 249, and 253 clearly dominate the number of video clips.

There is a difference in the number of accident types between the ranking of GIAK (https://accident.knia.or.kr/ranking) and our constructed dataset. Class 249 is the second most frequent class in our dataset, but it does not appear in GIAK's ranking. This phenomenon arises from the regional bias in the dataset. GIAK has regional dependence in Korea, whereas our constructed dataset was built through web crawling, therefore, being more international. Such cultural bias has been reported in various deep learning applications (Gebru *et al.*, 2021; Buolamwini & Gebru, 2018).

For simplicity, we have chosen to only concentrate on the top three classes (249, 252, 253) to verify the performance of our CNN-based spatiotemporal network. If we examine the count of the available video clips among these three classes, we can see that class 252 almost doubles in the number of counts as compared to the remaining two classes (249, 253). The bias on the available count in each class can result in unsatisfactory accuracy. In Section 3.2.2, we explain data augmentation methods to overcome this bias problem. Through pre-processing, we obtained a total of 192 video clips for three classes. We have split the dataset into 171 training sets and 21 test sets.

***Three classes of the dataset**

The classification criteria for the aforementioned three classes of car accidents are illustrated in the following Fig. 5. The blue car is the proponent's car. The red car is the opponent's car. Class 249 is wrong-way driving. The opponent's car seriously crosses the center line and hits the proponent's car. Class 252 is a lane change accident. While the proponent and opponent are driving in the same direction, the proponent changes the lane and collides with the proponent's car. Class 253 is a rear collision accident. While both are driving, the proponent's car collides with the rear of the opponent's car. The proportions of the liability, as seen by the proponent, are 0%, 30%, and 100%. Each class of accident can be classified based on the movements of the proponent's car and the opponent's car. Furthermore, because there exists little resemblance in the movements of the cars involved in the accident, it is suitable for the classifier used in this study.

3.2 Model

3.2.1 3D CNNs

The split liability assessment is determined by the trajectories of the cars and the environmental factors before the collision. Here, it is important that the information leading up to the collision is considered as a criterion for classifying the accident case. That is, the trajectories of the car are accumulated as a principle, and the accident cases are classified based on the accumulated trajectories. Therefore, to classify the accident case, it is necessary to analyse not only spatial information but also temporal information. The spatial and temporal development of the cars is of primary interest. To analyse these subjects of interest, we used 3D CNNs, which are capable of processing time and space.

3D CNNs are based on 2D convolution on 2D images in the directions of height and width and extended in the third direction of time. In other words, the filter (kernel) is a cuboid. Therefore, the time-stacked images are convolved in three dimensions. By utilizing these features, we can assess the time-dependent motion of the cars involved in the accident. We discussed the feature map, including the motion of the car in 4.4, with the class activation map (CAM).

In this study, the feature extraction capability of the backbone of the 3D CNNs plays an important role in accident case classification. Therefore, using C3D as the basic backbone, the performance of the backbone was measured by changing the number of layers and the pre-trained dataset. C3D is an AI model used in the field of action recognition, and because temporal and spatial data can be simultaneously processed, it was found to be most appropriate for

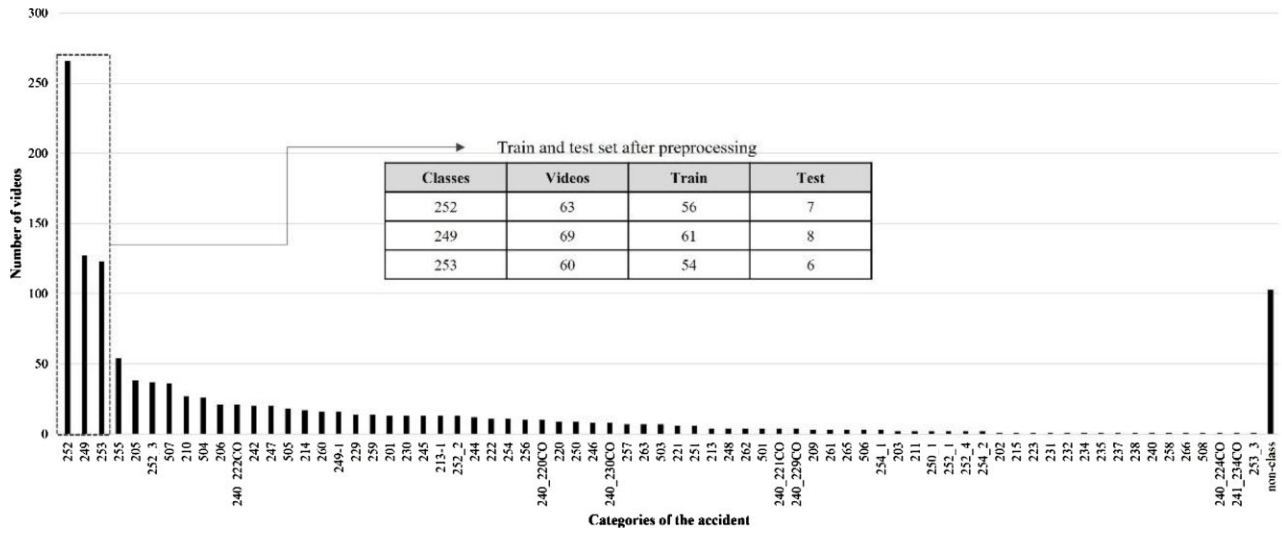


Figure 4: The count of video clips as a function of the car accident categories.

our problem. To see the effect of the depth of the backbone layer, we have compared by changing the layers to 11, 13, 16, and 19 layers. Table 2 shows the configuration of each layer-specific model. Each layer is denoted as conv3D-<number of channels > . The activation function is the RELU. The size of the filter is 3 × 3 × 3, which performs best in a 3D convolutional network (Tran *et al.*, 2015). Stride and padding are set to one.

Each layer is a VGG network, and the purpose of using this layer was to take the benefit of 2D CNNs. The head of the network is responsible for the classification and is made of soft-max functions. The network was used to classify three car-accident classes. The loss function used in the learning is cross-entropy losses.

We also experimented with the effect of transfer learning. As the depth of the layer is deepened, a larger dataset is required to train the network. This is a common problem in the training because a small dataset can result in over-fitting and will not generalize well. Thus, we have pre-trained our network using a huge Sport-1M dataset and HMDB51. Afterward, the pre-trained backbone was fixed, and only the classification part was fine-tuned. We compared the transfer learning effect with those done by training from random parameters.

The hyperparameters are as follows. The learning rate was set to 0.00001, the clip length was set to 16, and the batch size of 10 was used. Each parameter was optimized experimentally.

Figure 6 shows the overall network diagram. As mentioned previously, when the input images (clip) are input, the data is pipelined to the backbone and the head, and the resulting classification is reported.

3.2.2 *Data augmentation*

As noted previously, we have added more video clips through data augmentation. We chose data augmentation methods that do not interfere with the position and movement of the cars. First, we applied a horizontal flip. Figure 7 (top) illustrates the horizontal flip. Because we are dealing with video clips, entire sets of images are identically flipped. When the images are horizontally flipped, notice the directions of the movements of the cars, and the lane marks also change. However, these changes do not interfere with classifying classes 249, 252, and 253 because they are invariant to horizontal flips. For example, when a car changes the driving lane from left to right, it will change from right to left when horizontal flipping is applied. The fact that there was a lane change (class 252) does not change. Similarly, the horizontal flipping is invariant to classes 249 and 253.

Next, consider random cropping. When large cropping is used, it can lead to the loss of the cars involved in the accident. Therefore, we limited the cropping to be modest and only used cropping of images less than 20% of the original width and height of the images. Notice that cropping resulted in images with a reduced number of resolutions. Random cropping is illustrated in Fig. 7 (bottom). In addition, basic augmentation was performed to change the brightness and saturation of the image.

Here is another important method of augmentation, shuffle. In this study, the input video clip was shuffled. Thus, the video was divided into several clips and shuffled. In other words, the entire video was not used as input consecutively. One clip in our model consists of 16 images. Through this, we were able to solve data shortage problems and achieve results in performance improvements. In addition, the evaluation of the results described in the next chapter is constructed based on these shuffle augmentation results.

3.2.3 *Results evaluation*

A dash cam video from car accidents is split into multiple clips and input to the model with shuffle augmentation in training sessions. Moreover, each clip passes the backbone classifier, where the classified result is output in the test session. This means that one dash cam video can trigger many classified results. These results work as votes, and the majority of the voting result is used to determine the class of the dashcam video.

The dash cam video is composed of multiple clips denoted as c_i. The total number of the clip is defined as n. c_i is input to the network N(x). The output of the network is shown as $R(c_i)$. This classifier gives three probabilities. The classification result is shown as p_m. m denotes the class number: 249, 252, or 253. m∋{249, 252, 253}. Therefore, p_m denotes the probability that the clip belongs to a particular class, m. For any one clip c_i, $\sum p_m$ sums to 1.

For any given dash cam video V, the result of V can be represented as

$$V = \{R(c_0), R(c_1), R(c_2), R(c_3), \ldots R(c_n)\}\, R(c_n) \sim p_m.$$

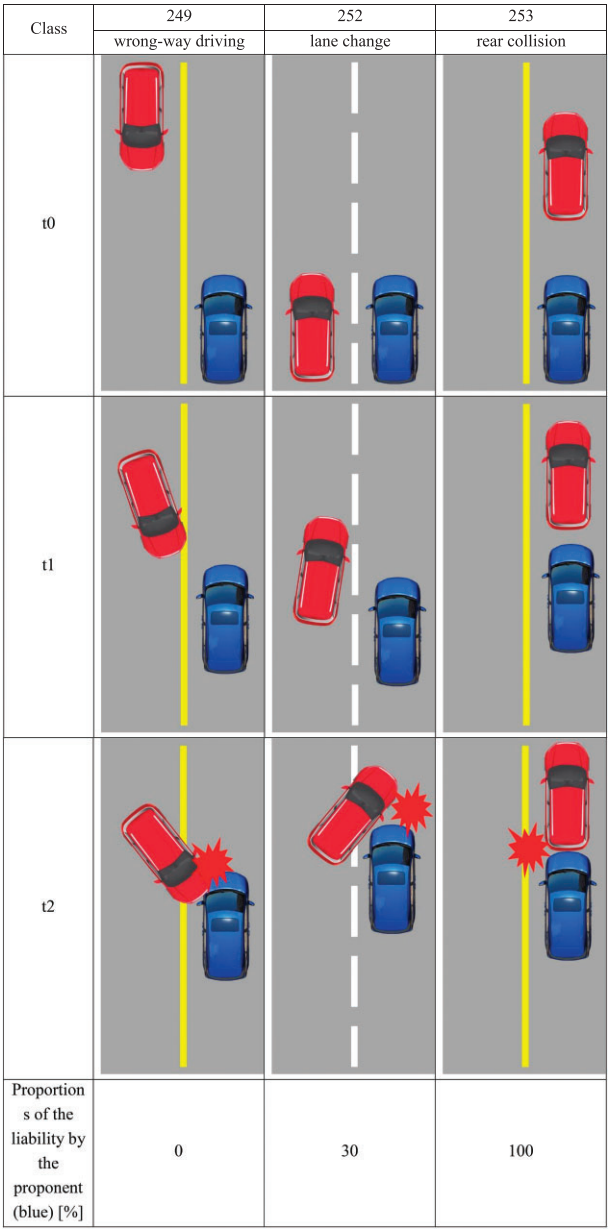


Figure 5: The time sequential movements and the proportional liability (as seen by the proponent (blue)) for the classes 249, 252, and 253.

The final class of V is determined by the dominant class of V. In other words, the most frequent $R(c_n)$ in V represents the final class of V.

4. Results and Discussion

4.1 Results with layers and parameters

We tested the effect of changing the depth of layers on the accuracy of 3D CNNs. We used the most popular 2D CNNs backbones of varying depths VGG11, VGG13, VGG16, and VGG19. And additionally, we conducted experiments on the SlowFast model (Feichtenhofer *et al.*, 2019), which is commonly cited in the field of action recognition. SlowFast has a ResNet-50 backbone, and we followed the criteria presented in the study for the training parameters. We conducted experiments on the scratch model and the pretraining model. The pretraining model was fine-tuned after being trained on the Kinetics-400 dataset.

Table 2: 3D Convolutional networks configurations.

11 layers	13 layers	16 layers	19 layers
input (224 × 224 × 224 RGB video)			
Conv3D-64	Conv3D-64 **Conv3D-64**	Conv3D-64 Conv3D-64	Conv3D-64 Conv3D-64
maxpool			
Conv3D-128	Conv3D-128 **Conv3D-128**	Conv3D-128 Conv3D-128	Conv3D-128 Conv3D-128
maxpool			
Conv3D-256 Conv3D-256	Conv3D-256 Conv3D-256	Conv3D-256 Conv3D-256 **Conv3D-256**	Conv3D-256 Conv3D-256 Conv3D-256 **Conv3D-256**
maxpool			
Conv3D-512 Conv3D-512	Conv3D-512 Conv3D-512	Conv3D-512 Conv3D-512 **Conv3D-512**	Conv3D-512 Conv3D-512 Conv3D-512 **Conv3D-512**
maxpool			
Conv3D-512 Conv3D-512	Conv3D-512 Conv3D-512	Conv3D-512 Conv3D-512 **Conv3D-512**	Conv3D-512 Conv3D-512 Conv3D-512 **Conv3D-512**
maxpool			
FC-4096			
FC-4096			
FC-3			
softmax			

Table 3 shows the result obtained by varying the depth of the backbone layers. Accuracy calculated whether the final prediction result obtained through stacking is correct with ground truth. The first column is the number of layers. The second to last columns are accuracy. The second column is the average accuracy for the entire video. The third to fifth columns show the accuracy of each case.

In average accuracy, the model with 11 layers proposed by C3D shows 71.4% accuracy. And the models with 13 layers and 16 layers show the same accuracy at 66.7%. Lastly, the layer 19 model shows the lowest accuracy of 61.9%. The complexity of the model increases with the number of layers. In the classification of the three accident cases used in this study, it can be seen that the performance of the small model is high. These results can also be observed in comparison with the SlowFast model. The SlowFast model achieved accuracies of 38.1% and 57.1% in the scratch and pretraining models, respectively.

The accuracy of 253 case of the layer 11 model is the lowest at 33.3%. Thus, the average accuracy is the highest due to the high accuracy of 249 and 252 cases. On the other side, it can be seen that the layer 16 model is evenly distributed for each case. That is, it can be seen that the performance of the layer 16 model is good in terms of generalization. However, as mentioned in Section 3.1.3, when 249 case is not the majority, one could also consider choosing 11 or 13 layer models that achieve high accuracy on other cases. This approach, which emphasizes the importance

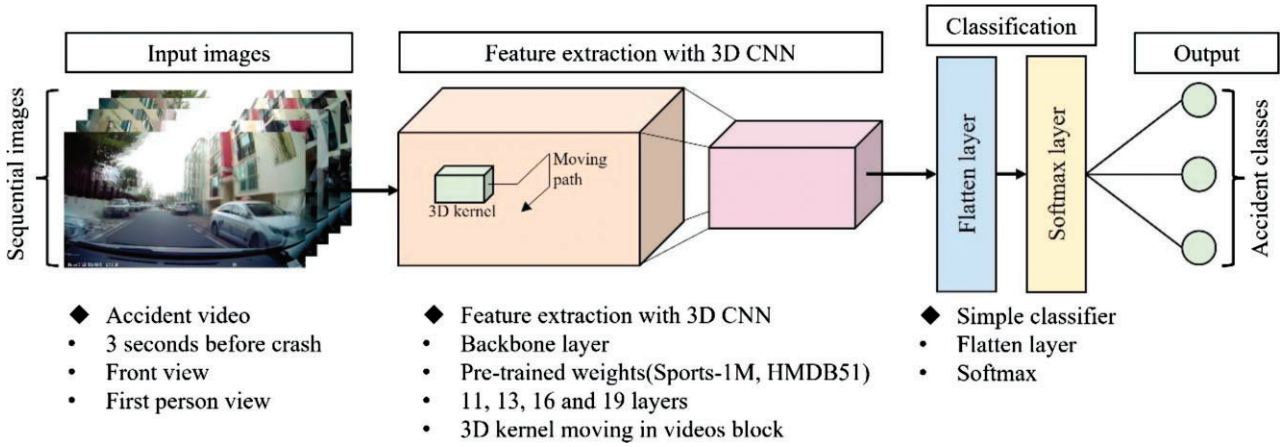


Figure 6: The network diagram of the split liability assessment based on 3D CNN.

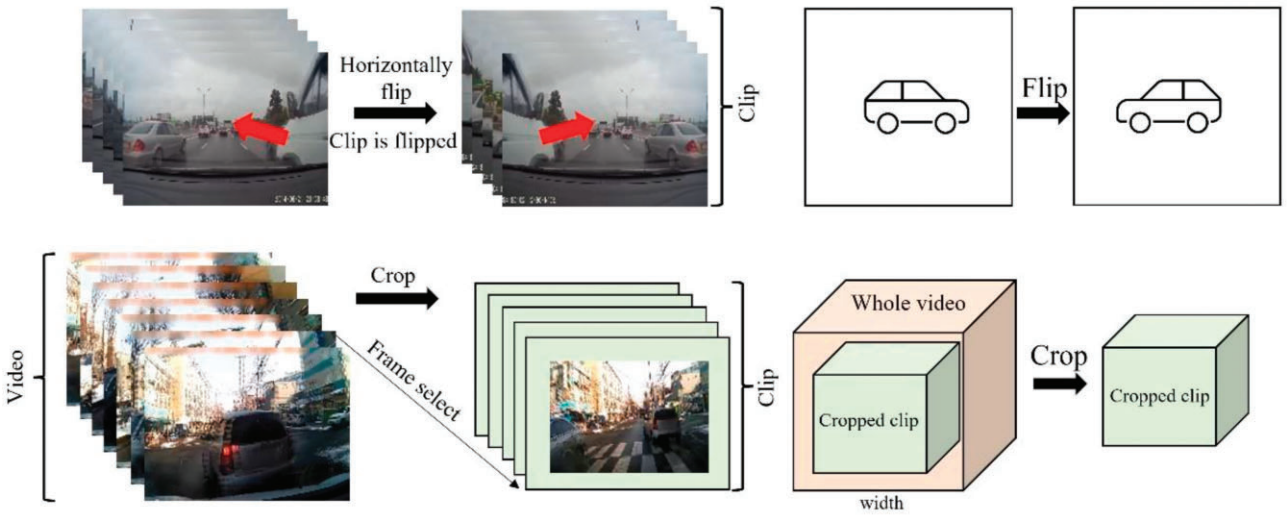


Figure 7: Data augmentation: (top) horizontal flipping, (bottom) random cropping.

Table 3: Accuracy obtained by varying the network depth.

Number of layers	Accuracy [%]	Accuracy_249	Accuracy_252	Accuracy_253
11	**71.4**	**71.4**	**100**	33.3
13	66.7	42.9	75.0	**83.3**
16	66.7	**71.4**	62.5	66.7
19	61.9	**71.4**	75.0	33.3
50*	38.1	**42.9**	25.0	50.0
50**	57.1	**85.7**	62.5	16.6

* slowfast, ResNet-50 backbone, scratch
** slowfast, ResNet-50 backbone, kinetic-400 pre-trained

of accurate classification for specific classes, has been used in the field of medicine (Fotouhi *et al.*, 2019).

4.2 Results with pre-trained dataset

As described above, the model performance according to the pretraining dataset is confirmed as a method for evaluating the performance of the backbone. Table 4 is a brief description of the pretraining dataset. Pretraining is carried out using Sports-1M and HMDB51. Sports-1M consists of more than 1 million videos. It has a total of 487 classes and 1000–3000 clips per class. It contains labels such as skiing, judo and yoga because it is made for sports videos. HMDB51 is a dataset with 51 action classes. It consists of 6849 video clips. It has classes such as walking and shaking hands because it deals with the action of a person's motion.

Table 5 represents the accuracy according to the pretraining dataset. Other parameters other than the pretraining dataset are the same for the model used in this section. In other words, it has the same parameters as the layer 11 model of Section 4.1. This network learns 200 times (epochs) for the pretraining dataset. After that, the head part is fine-tuned using the accident video dataset.

Table 4: Pre-trained dataset.

Pre-trained dataset	Size	Class	Clips per class	Objects example
Sports-1M	Over million	487	1000–3000	Skiing, judo etc
HMDB51	6849	51	101	Walking, shaking hands etc

Table 5: Accuracy obtained by varying the dataset.

Pre-trained dataset	Accuracy [%]	Accuracy_249	Accuracy_252	Accuracy_253
Sports-1M	71.4	71.4	100	33.3
HMDB51	**90.5**	**71.4**	**100**	**100**

Experiment results show that the model that learned the HMDB51 shows higher performance than the Sports-1M model. It can be seen that the average accuracy is 90.5%, showing higher performance even in the effect of the change of the layer tested earlier. In addition, it can be seen that the accuracy for each case is also higher. It can be seen that the performance of the backbone by the pretraining dataset is more important than the effect of the number of layers through this result.

4.3 Results with GUI

We have implemented a simple GUI to show the intermediate and final outcome of the split liability assessment. The left pane shows the loaded dashcam video. The bottom area, it shows the location of the loaded video. The right pane shows the analysis result. The analysis includes the intermediate (on the fly) classifications and also the final classification. In this picture, the first, second, and third lines show the top three classification results that have been processed by the dashcam videos played up until that moment. The first line shows two colored bars. The upper bar (sky color) shows the predicted likelihood, and the bottom (green color) shows the number of votes accredited to the particular class. For example, the moment of the dash cam image is predicted to be 'lane change (252)' as shown in Fig. 8. It has 67% likelihood. And this video has been predicted 32 times with the lane change case so far. The two bars for the lines second and third also show similar bars. Notice the sum of the three upper bar (sky color) percentile values is 100%. The final conclusion shown at the bottom of the right panel is determined by a majority vote. The classification number, accident types, and split liability assessment follow immediately after. In this section, we discuss the result as shown in these dashcam videos for several examples.

4.3.1 *253-class*

Figure 9 shows the time lapsed (t0∼t3) screenshots of the dashcam video and the accompanying assessment results. We used a single Tesla-V100 GPU for inference. The test performance of the model (11 layers, HMDB51 pre-trained) is 40 FPS. Real-time first denotes the most probable class, and real-time second shows the second most probable class. At the time t0, we can see that there are two cars, one in the front and the other in the second lane. Notice that the model predicts that the top two predictions are first 252 and second 253 with similar probabilities. And most noticeably, the car in the second lane is closer than the car in the front. At the time t1, notice the two cars are closer. Notice the prediction of the model is now first 253 and second 252. Since the influence of the car in the front is stronger, the first value now changes to 253. At the time t2, the first 253 and second 252 remain unchanged. And the probability of third 249 is lower than at the time t0 and t1. In other words, two cars are in the front view, and the likelihood of the imminent accident is distributed among the top two predictions. Lastly, at the time t3, first 253 accounts for 74% probability, and class 253 seems to be more dominant. Thus, the probability of first 253 has the highest value. From these results, we can see that the network can distinguish between the two classes, 252 and 253. This is due to the fact that the second lane car is totally out of sight, and only the front car strongly affects the predicted accident class.

As described above, we can understand the ability of the proposed network to classify the type of accident by analysing dashcam videos. In addition, the network is capable of predicting an accident that may occur in real-time while watching the video. This demonstrates that the model is able to analyse the movement of the car. Furthermore, the model is more accurate in predicting ground truth data in the latter part of the video, closer to the moment of collision. Lastly, the final assessment of the model's performance shows that it is equal to the ground truth label of 253.

4.3.2 *249-class*

Figure 10 shows the time lapsed (t0∼t3) assessments of a case about class 249. The assessment results for specific moments are shown in Fig. 10. As previously shown, real-time first denotes the most probable class, and real-time second shows the second most probable class.

At the time t0, we see that the car (ego) is driving in the second lane. Notice that the model predicts that the top two predictions are first 249 and second 252. At the time t1, we suddenly see a car in the first lane. The assessment classes are first 252 and second 249. Because a car appeared in the first lane, it is most likely that an imminent lane-change class accident may occur due to the appearance of this car. At the time t2, we can see that the model continues to pay close attention to lane-change class as identical to time t1. Lastly, at time t3, we see that a car traveling in the opposite direction is headed toward the car with the dashcam. The assessment now reads first 249 and second 253. The reason why the second prediction showed an accident while driving is presumed to be because the last moment of the video shows the opponent's rear side of the car.

The screenshots show that the assessment focuses on wrong-way driving. This can be presumed to be because the scene just prior to the collision depicted scenes often seen from wrong-way driving in training sessions. Additionally, we can see that the assessment changes over time. This is due to the effect of the processed clips. We can understand each clip possesses enough information to predict the class of the accident in the near future. In other words, the model can be used to perceive upcoming accidents.

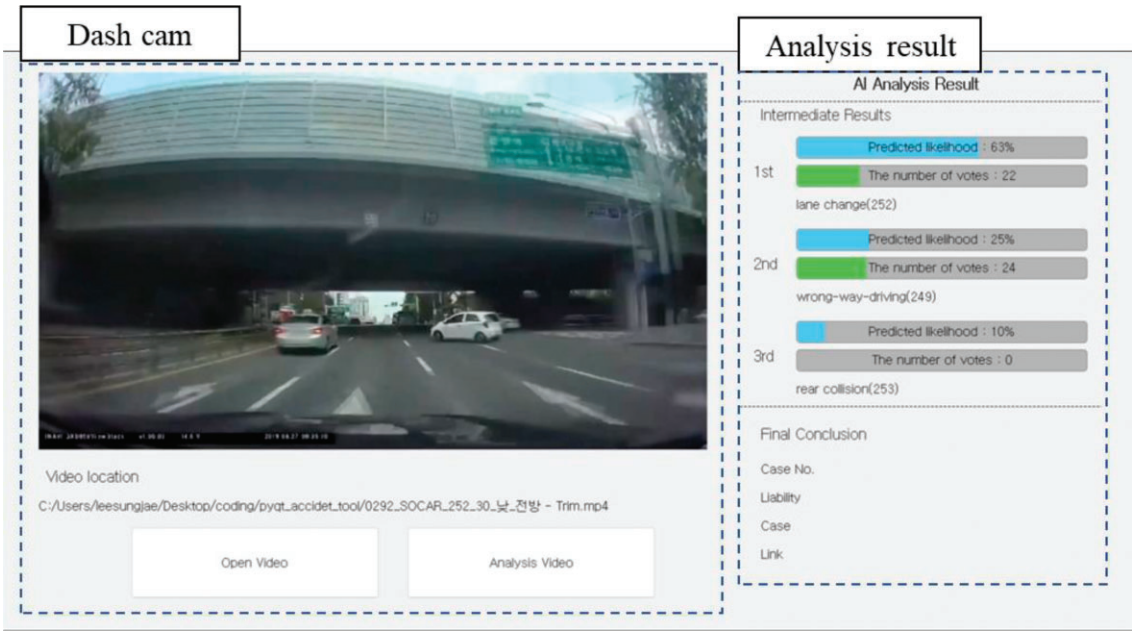


Figure 8: GUI software for the split liability assessment.

4.3.3 *252-class*

Lastly, we discuss the 252 class. Figure 11 shows the time lapsed (t0∼t3) assessments of a case about class 252. At the time t0, the model assesses first 249 and second 253. At the time t1, first is 252, and second is 249. And at the time t2, first is 252, and second is 249. At times t0 and t1, because cars in sight are very far away, the model assumes accidents that might occur during such situations. At the time t1, the car on the left lane is assessed to increase the probability of causing the lane-change accident. In other words, at the time t2, it appears that the effect of accumulated voting counts influenced the assessment, not the actual car on the right-hand side that would eventually collide. We can conclude that when the likelihood of the same class of accidents is high, this tendency is preserved through accounted voting counts. Lastly, at the time t3, first is 252, and the second is 253. Now at the moment when the collision occurred, we can see that the model correctly and quickly assessed the accident to be of class 252. It is interesting to see the second assessment, which is 253. In summary, the model concluded the case to be of class 252, which is the lane-change accident.

By examining Sections 4.3.1 through 4.3.3, we can make the following conclusions:

1) The model assessment with quantized clips proved to show accurate results.
2) When two conflicting classes of accidents exist, the model lists both with more votes for the true accident class.
3) For clips that are far from the moment of collisions, the model predicts the most likelihood of the accident considering the current situation.
4) The voting scheme can be used to conclude the final assessment. This is the reason why we have limited the clip to be of 3 seconds.

In the next section, we look into how the model is making the assessment through the use of the CAM.

4.4 Results with CAM

There exist many approaches to understanding how CNNs work (Zeiler & Fergus, 2014). For example, the filters are visualized to see that the edges are sought in the shallow part of the layers, and more high-level features are detected in the deeper layers. In classification problems, it is of primary interest to understand which parts of the network can be attributed to deciding on a particular class. Therefore, we used CAM to understand which parts were responsible for concluding certain classes (Zhou *et al.*, 2016).

CAM can be obtained by replacing the fully connected network with Global Average Pooling in the output layer of a convolutional network. A CNN layer with the number of channels equal to the number of classes is attached at the end of the backbone, where feature maps are extracted. Each channel of the corresponding layer represents one class. By adding a FC layer, a classification is achieved with softmax. CAM can be obtained by multiplying the weights at the Fully Connected (FC) layer with the feature map.

Figure 12 illustrates a snap show about the accident class 252 (lane change). We obtained the CAM by assuming three classes at the last layer. Time-lapsed results are shown from t0 to t3. The left image shows the raw image, and the right image shows the CAM result. Areas of high attention are colorized with red. At time t0, it appears fewer areas are showing red colors. The lower left corner pulls some attention. This is due to the fact for lane-change classes, lower-left and lower-right corners are the areas where lane changes are happening.

For the t1 image, we still see much attention is given to the lower left lane marks. We also see some attention given to the center of the image as well as to the right. The attention given at the center is due to the fact that there exists a car in the front. It is interesting to see that the parked car on the right is pulling attention even if it is not in a drivable space. We can understand that all cars, whether it is in drivable space or not, are attracting much attention. In fact, we were able to see that

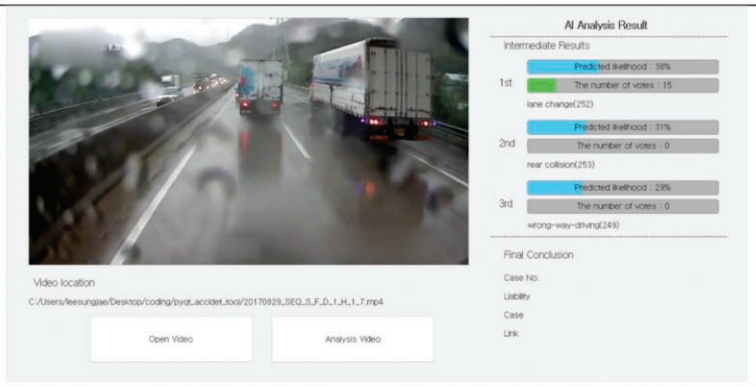
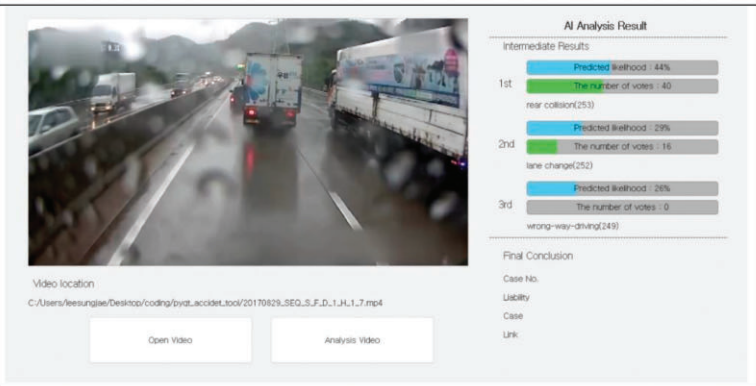
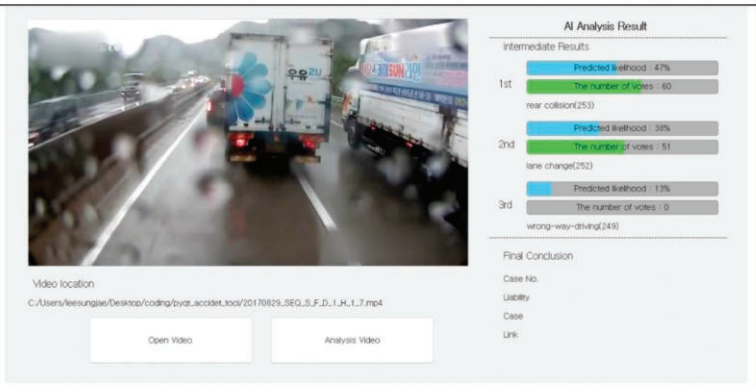
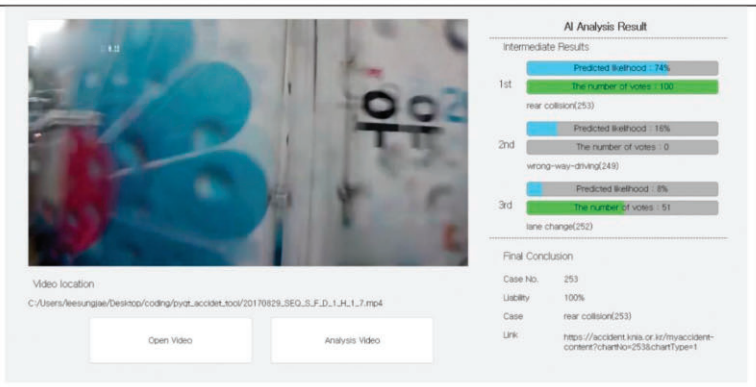
Time	Images	Results
t0		Real-time 1st : 252 Real-time 2nd : 253
t1		Real-time 1st : 253 Real-time 2nd : 252
t2		Real-time 1st : 253 Real-time 2nd : 252
t3		Real-time 1st : 253 Real-time 2nd : 252

Figure 9: Time lapsed predictions of the case of accident class 253.

much attention was given to all cars in the image for the trained dataset.

At the time t2, we can see the left car is receiving attention. At the same time, the rest of the areas (lower left corner at the time t0 and t1) is receiving less attention. This is due to the fact that just prior to the collision car that is most likely to collide is given more attention.

Lastly, the CAM image at the time t2 shows that after the moment of collision, the attention is increased for the collided car. By examining the CAM results, we can see that the model pays special attention to cars. And when cars are not present (t0), more attention is paid to the generalized circumstances of the relevant class. Throughout this study, we were able to conclude that our model is able to categorize car accidents by the trained classes.

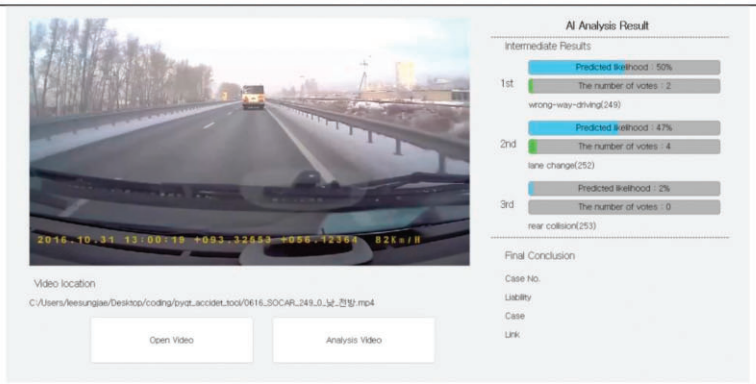
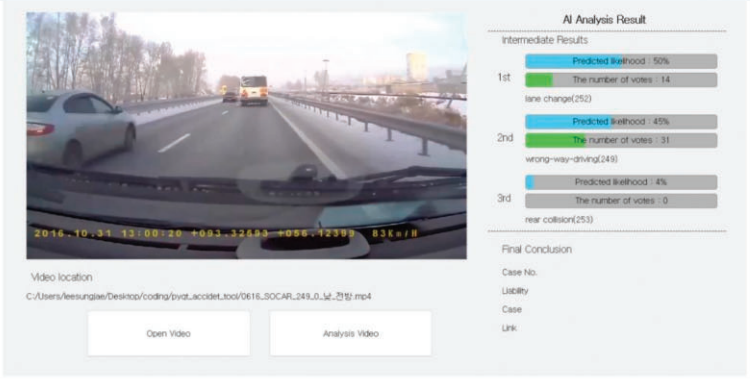
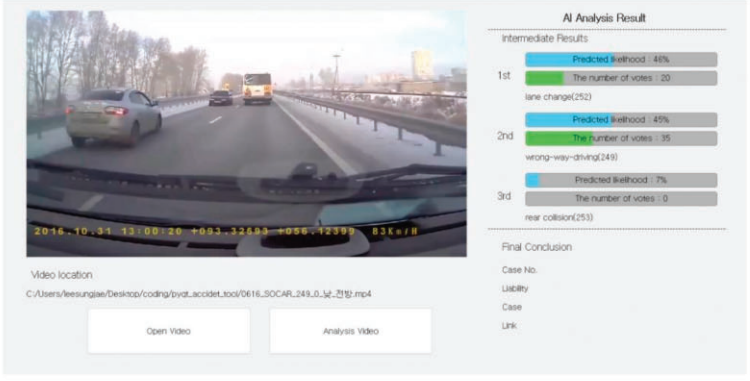
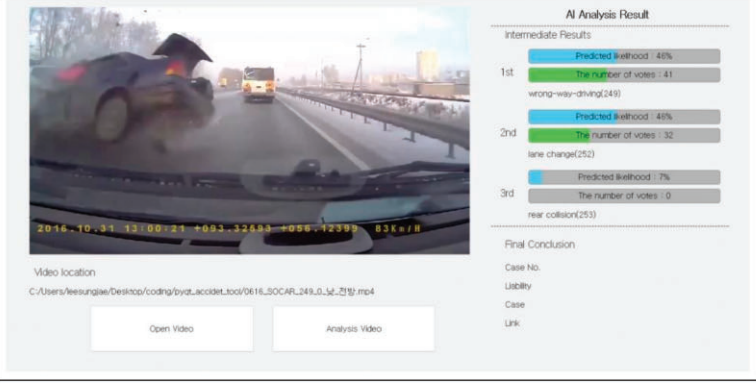
Time	Images	Results
t0		Real-time 1st : 249 Real-time 2nd : 252
t1		Real-time 1st : 252 Real-time 2nd : 249
t2		Real-time 1st : 252 Real-time 2nd : 249
t3		Real-time 1st : 249 Real-time 2nd : 252

Figure 10: Time lapsed predictions of the case of accident class 249.

5. Conclusions

Usually, the field of video recognition is known as the field of behavior recognition using a short video clip such as sports-1M. However, video analysis often analyses information on long videos in the real world. Accordingly, as a representative example, a task for the split liability assessment in a car accident is proposed in this study. This task involves generating a final classification result by combining factors that influence a liability assessment

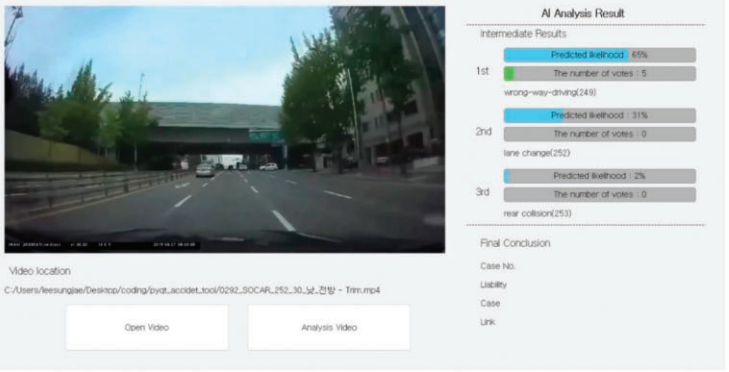
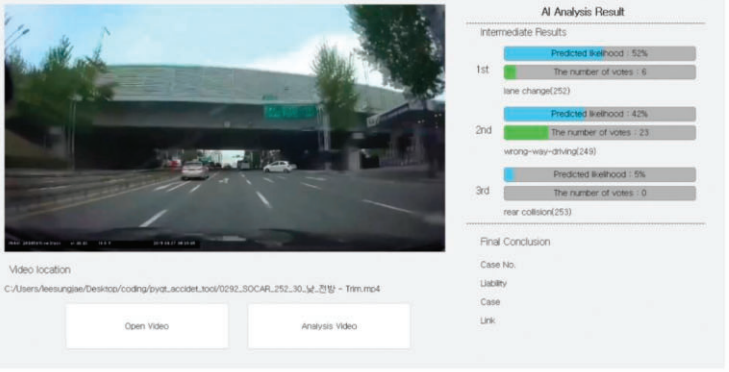
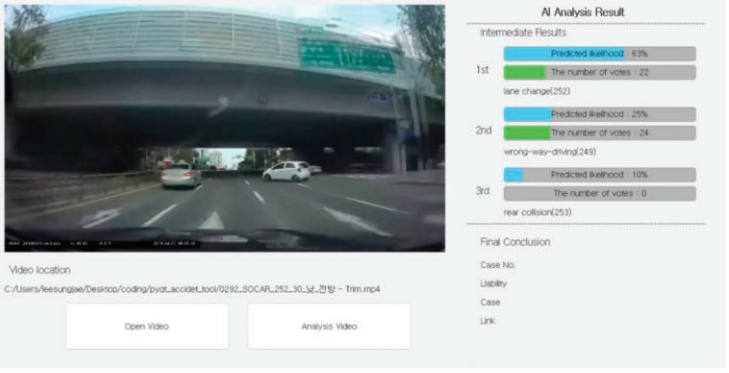
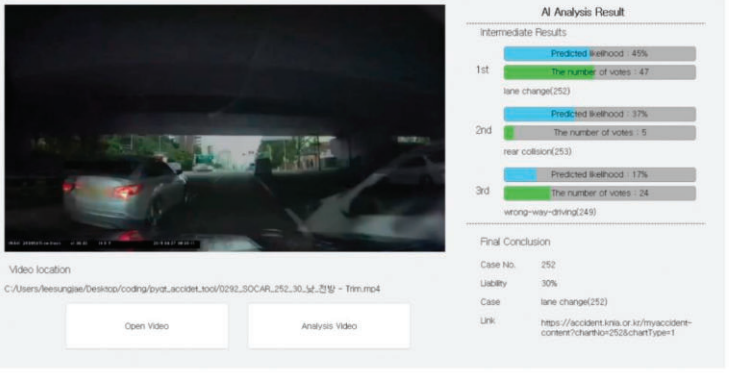
Time	Images	Results
t0		Real-time 1st : 249 Real-time 2nd : 252
t1		Real-time 1st : 252 Real-time 2nd : 249
t2		Real-time 1st : 252 Real-time 2nd : 249
t3		Real-time 1st : 252 Real-time 2nd : 253

Figure 11: Time lapsed predictions of the case of accident class 252.

over time in a long video. Therefore, we propose a new task for analysing long videos.

In this study, video analysis is carried out by using 3D CNNs. It is a network that analyses short videos from the past. Therefore, long video classification was performed through a method of stacking each clip result. The results of acquiring important information from long videos were shown through this method, unlike previous studies. The CAM results are also shown in this study to see where the network concentrates. The network focuses on and tracks the movement of the vehicle over time.

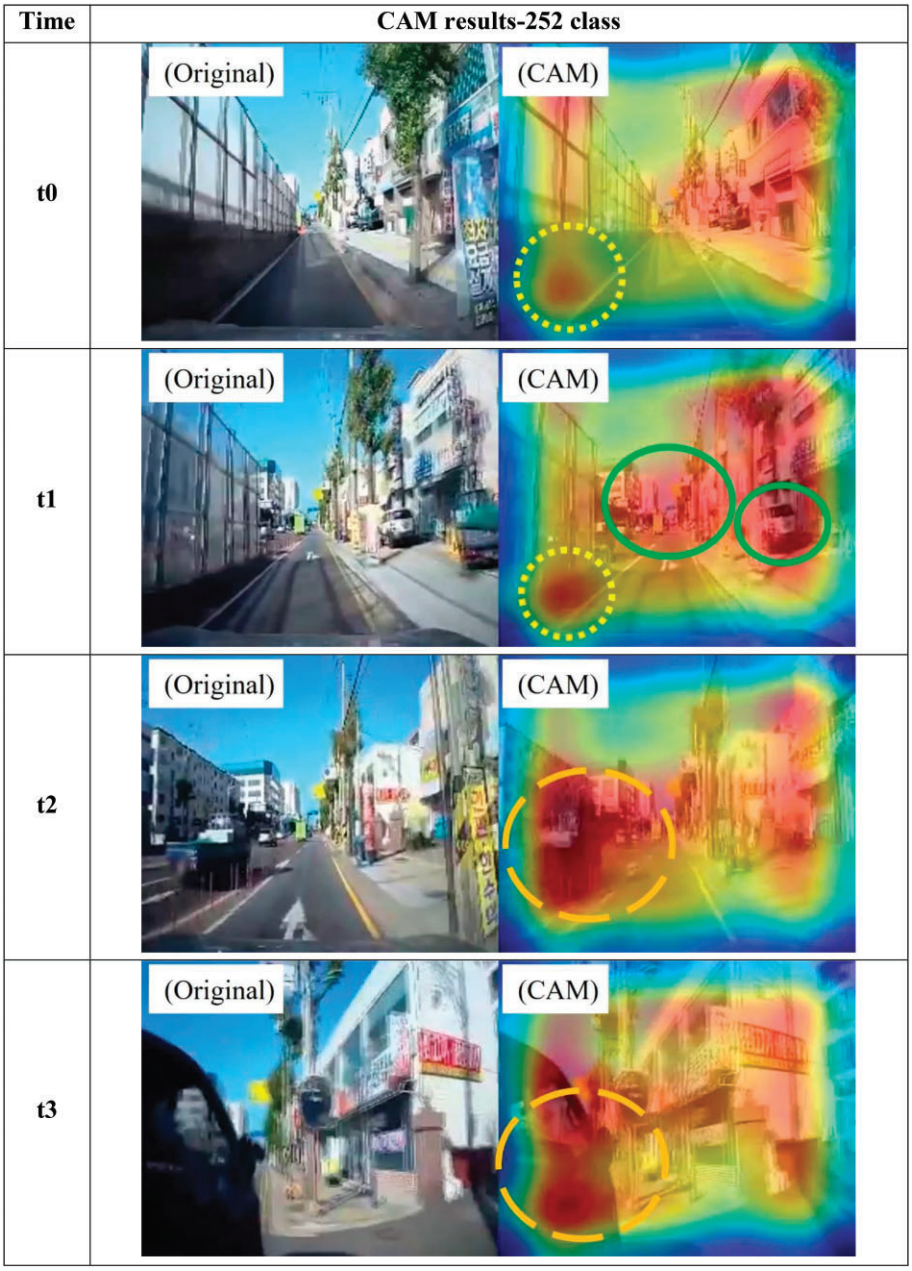


Figure 12: CAM results for the class 252: t = time, original = raw input video, CAM = 252 class CAM results.

From the perspective of using constructed data, we deleted ill-conditioned data and used the remaining data, as shown in Fig. 2. However, in real accidents, the videos taken from the vehicles often contain a number of peculiar footage. Therefore, to enable the wide use of the developed network in actual accident situations, these videos must be dealt with. The camcorder footage obtained by filming the dashcam screens with a cell phone cannot be used because the AI model confuses the dashboard background and the relevant traffic actors in the dashcam playback. Perhaps it may be possible to extract and use the video by using specialized filtering techniques to prune the background dashboard or remove unwanted motion from the screen. Low-definition videos could also be pre-processed using video super-resolution techniques before applying them. We plan to extend our work to these ill-conditioned videos in the future.

The proposals and results of this study show a new video task in artificial intelligence. Therefore, it will lead to applying existing video recognition field research to the real world. It will also be the foundation for the AI field that can be applied to the legal field through split liability assessment research.

In our study, we analysed spatiotemporal information using 3D CNNs. The 3D CNN has the advantage of being able to simultaneously process spatiotemporal information. The operation of 3D CNN involves the convolution of kernels along the temporal axis with limited window. Consequently, as the window progresses over time, information outside the window is lost. While it is capable of handling full temporal information in short videos, it exhibits limitations when processing longer videos. Therefore, our future work is to develop a transformer-based network that can capture important spatiotemporal information in long videos.

Acknowledgments

This work was partly supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) [P0020535, The Competency Development Program for Industry Specialist], Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [No. 2019–0-01842, Artificial Intelligence Graduate School Program (GIST)] and GIST Research Project grant funded by the GIST in 2023.

Conflict of interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., & Vijayanarasimhan, S. (2016). Youtube-8m: A large-scale video classification benchmark. *arXiv preprint*. https://doi.org/10.48550/arXiv.1609.08675.

Adewopo, V., Elsayed, N., ElSayed, Z., Ozer, M., Abdelgawad, A., & Bayoumi, M. (2022). Review on action recognition for accident detection in smart city transportation systems. *arXiv preprint*. https://doi.org/10.48550/arXiv.2208.09588.

Ballas, N., Yao, L., Pal, C., & Courville, A. (2015). Delving deeper into convolutional networks for learning video representations. *arXiv preprint*. https://doi.org/10.48550/arXiv.1511.06432.

Bao, W., Yu, Q., & Kong, Y. (2020). Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2682–2690. https://doi.org/10.1145/3394171.3413827

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, **5**, 157–166. https://doi.org/10.1109/72.279181

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.

Caba Heilbron, F., Escorcia, V., Ghanem, B., & Carlos Niebles, J. (2015). Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 961–970. https://doi.org/10.1109/CVPR.2015.7298698

Chan, F. H., Chen, Y. T., Xiang, Y., & Sun, M. (2017). Anticipating accidents in dashcam videos. *In Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part IV*. **13**, 136–153. https://doi.org/10.1007/978-3-319-54190-7_9

Choi, M., Kim, C., & Oh, H. (2022). A video-based SlowFastMTB model for detection of small amounts of smoke from incipient forest fires. *Journal of Computational Design and Engineering*, **9**, 793–804. https://doi.org/10.1093/jcde/qwac027

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint*. https://doi.org/10.48550/arXiv.1412.3555

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, **1**, 886–893. https://doi.org/10.1109/CVPR.2005.177

Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., & Darrell, T. (2016). Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 677–691. https://doi.org/10.1109/TPAMI.2016.2599174

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). CARLA: An open urban driving simulator. In *Conference on Robot Learning*. **78**, 1–16. https://doi.org/10.48550/arXiv.1711.03938

Fan, Q., Chen, C. F. R., Kuehne, H., Pistoia, M., & Cox, D. (2019). More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation. *Advances in Neural Information Processing Systems*, **32**. https://doi.org/10.48550/arXiv.1912.00869

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6201–6210. https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00630

Fotouhi, S., Asadi, S., & Kattan, M. W. (2019). A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of Biomedical Informatics*, **90**, 103089. https://doi.org/10.1016/j.jbi.2018.12.003

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, **64**, 86–92. https://doi.org/10.1145/3458723

Gowda, S. N., Rohrbach, M., & Sevilla-Lara, L. (2021). Smart frame selection for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 1451–1459. https://doi.org/10.1609/aaai.v35i2.16235.

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., & Memisevic, R. (2017). The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*. 5842–5850. https://doi.org/10.1109/ICCV.2017.622

Hara, K., Kataoka, H., & Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6546–6555. https://doi.org/10.48550/arXiv.1711.09577

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://doi.org/10.48550/arXiv.1512.03385

Houssein, E. H., Abohashima, Z., Elhoseny, M., & Mohamed, W. M. (2022). Hybrid quantum-classical convolutional neural network model for COVID-19 prediction using chest X-ray images. *Journal of Computational Design and Engineering*, **9**, 343–363. https://doi.org/10.1093/jcde/qwac003.

Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 221–231. https://doi.org/10.1109/TPAMI.2012.59

Jiaxin, Y., Fang, W., & Jieru, Y. (2021). A review of action recognition based on convolutional neural network. In *Journal of Physics: Conference Series. 1827*, 012138. https://doi.org/10.1088/1742-6596/1827/1/012138

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1725–1732. https://doi.org/10.1109/CVPR.2014.223

Klaser, A., Marszałek, M., & Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, 99.1-99.10. https://doi.org/10.5244/C.22.99.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, **60**, 84–90. https://doi.org/10.1145/3065386

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*, 2556–2563. https://doi.org/10.1109/ICCV.2011.6126543

Laptev, I. (2003). Lindeberg, "Space-time interest points". In *Proceedings of the 9th IEEE Inter. Conf. Computer Vision (ICCV)*. 13–16. https://doi.org/10.1109/ICCV.2003.1238378

Le, V.-T., Tran-Trung, K., & Hoang, V. T. (2022). A comprehensive review of recent deep learning techniques for human activity recognition. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2022/8323962

Li, Y., Ma, D., An, Z., Wang, Z., Zhong, Y., Chen, S., & Feng, C. (2022). V2X-Sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, **7**, 10914–10921. https://doi.org/10.1109/LRA.2022.3192802

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, **2**, 1150–1157. https://doi.org/10.1109/ICCV.1999.790410

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., & Ray, A. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*. **35**, 27730–27744. https://doi.org/10.48550/arXiv.2203.02155

Patel, C. I., Garg, S., Zaveri, T., Banerjee, A., & Patel, R. (2018). Human action recognition using fusion of features for unconstrained video sequences. *Computers & Electrical Engineering*, **70**, 284–301. https://doi.org/10.1016/j.compeleceng.2016.06.004

Piergiovanni, A. J., & Ryoo, M. S. (2019). Representation flow for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9945–9953. https://doi.org/10.1109/CVPR.2019.01018

Quddus, A., Zandi, A. S., Prest, L., & Comeau, F. J. (2021). Using long short term memory and convolutional neural networks for driver drowsiness detection. *Accident Analysis & Prevention*, **156**, 106107. https://doi.org/10.1016/j.aap.2021.106107

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788. https://doi.org/10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, **28**. https://doi.org/10.48550/arXiv.1506.01497

Scovanner, P., Ali, S., & Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM International Conference on Multimedia*, 357–360. https://doi.org/10.1145/1291233.1291311

Shang, S., Masson, C., Llari, M., Py, M., Ferrand, Q., Arnoux, P.-J., & Simms, C. (2021). The predictive capacity of the MADYMO ellipsoid pedestrian model for pedestrian ground contact kinematics and injury evaluation. *Accident Analysis & Prevention*, **149**, 105803. https://doi.org/10.1016/j.aap.2020.105803

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, https://doi.org/10.48550/arXiv.1409.1556

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint*. https://doi.org/10.48550/arXiv.1212.0402.

Steffan, H., & Moser, A. (1996). The collision and trajectory models of PC-CRASH. *SAE Technical Paper*. (960886). https://doi.org/10.4271/960886

Steffan, H., Geigl, B., & Moser, A. (1999). A new approach to occupant simulation through the coupling of PC-Crash and MADYMO. *SAE Transactions*, 785–793. https://doi.org/10.4271/1999-01-0444

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 4489–4497. https://doi.org/10.1109/ICCV.2015.510

Wan, S., Ding, S., & Chen, C. (2022). Edge computing enabled video segmentation for real-time traffic monitoring in internet of vehicles. *Pattern Recognition*, **121**, 108146. https://doi.org/10.1016/j.patcog.2021.108146

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, 20–36. https://doi.org/10.48550/arXiv.1608.00859

Xiao, Y., Su, X., Yuan, Q., Liu, D., Shen, H., & Zhang, L. (2021). Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection. *IEEE Transactions on Geoscience and Remote Sensing*, **60**, 1–19. http://dx.doi.org/10.1109/TGRS.2021.3107352

Xiao, Y., Yuan, Q., He, J., Zhang, Q., Sun, J., Su, X., & Zhang, L. (2022). Space-time super-resolution for satellite video: A joint framework based on multi-scale spatial-temporal transformer. *International Journal of Applied Earth Observation and Geoinformation*, **108**, 102731. https://doi.org/10.1016/j.jag.2022.102731

Xiao, Y., Yuan, Q., Jiang, K., He, J., Wang, Y., & Zhang, L. (2023a). From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution. *Information Fusion*, **96**, 297–311. https://doi.org/10.1016/j.inffus.2023.03.021

Xiao, Y., Yuan, Q., Jiang, K., Jin, X., He, J., Zhang, L., & Lin, C.-w. (2023b). Local-Global Temporal Difference Learning for Satellite Video Super-Resolution. *arXiv preprint*. https://doi.org/10.48550/arXiv.2304.04421

Xu, W., Wang, J., Fu, T., Gong, H., & Sobhani, A. (2022). Aggressive driving behavior prediction considering driver's intention based on multivariate-temporal feature data. *Accident Analysis & Prevention*, **164**, 106477. https://doi.org/10.1016/j.aap.2021.106477

Yao, Y., Xu, M., Wang, Y., Crandall, D. J., & Atkins, E. M. (2019). Unsupervised traffic accident detection in first-person videos. In *International Conference on Intelligent Robots and Systems*, 273–280. https://doi.org/10.1109/IROS40897.2019.8967556

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proceeedings of the European Conference on Computer Vision*, 818–833. https://doi.org/10.48550/arXiv.1311.2901

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2921–2929. https://doi.org/10.1109/CVPR.2016.319

Zhou, B., Andonian, A., Oliva, A., & Torralba, A. (2018). Temporal Relational Reasoning in Videos. In *Computer Vision–ECCV2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, 831–846. https://doi.org/10.1007/978-3-030-01246-5_49.

Appendix 1. Accident cases for the split liability assessment

Accident case	Large category	Medium category	Small category	Category number	Whether split liability assessment is checkable with my video data	Necessity of opponent's video data	Front dash cam	Rear dash cam
Car-to-Car accident	Straight-to-straight accident	Crossroads (traffic lights are on the both sides)	Traffic violence-related accident by a car from one side	201	O	X	O	X
			Accident by failing to escape the crossroads	202	O	X	O	X
			Traffic violence-related accident by cars from both sides (yellow-light vs red-light)	203	O	X	O	X
			Traffic violence-related accident by cars from both sides (red-light vs red-light)	204	△	O	O	X
		Crossroads (based on the width of road)	Entering from right lane versus from left lane (same road width)	205	O	X	O	X
			Entering from right lane versus from left lane (different road width)	206	O	X	O	X
			Stop sign vioation-related accident	207	△	O	O	O
			One-way sign vioation-related accident	208	△	O	O	X
	Straight-to-left-turn accident (head)	Crossroads (traffic lights are opposite one anthoer)	Signal violation-related accident (going straight on red-light)	209	O	X	O	X
			Signal violation-related accident (turning left on green-light for straight)	210	O	X	O	X
			Signal violation-related accident (going straight on yellow-light)	211	O	X	O	X
			Signal violation-related accident by cars from both sides (same traffic light)	212	O	X	O	X
			Accident by failing to escape the crossroads	213	O	X	O	X
		Crossroads (unprotected left turn)	Unprotected left turn-related accident	213–1	O	X	O	X
		Crossroads (no traffic light)	Crossroads-related accident (no traffic light)	214	O	X	O	X
	Stright-to-left-turn accident (Side)	Crossroads (traffic lights are across the crossroads)	Signal violation-related accident (going straight on red-light)	215	O	X	O	X

Continued

Accident case	Large category	Medium category	Small category	Category number	Whether split liability assessment is checkable with my video data	Necessity of opponent's video data	Front dash cam	Rear dash cam
			Signal violation-related accident by cars from both sides (red-light)	216	O	X	O	X
			Signal violation-related accident by cars from both sides (going straight on yellow-light vs turning left on red-light)	217	O	X	O	X
			Signal violation-related accident by cars from both sides (going straight on red-light vs turning left on yellow-light)	218	O	X	O	X
			Accident by failing to escape the crossroads	219	O	X	O	X
		Crossroads (based on the width of road)	Straight from the left road vs left-turn from the right road accident	220	O	X	O	X
			Straight from the right road vs left-turn from the left road accident	221	O	X	O	X
			Going straight on the wide road vs turning left on the narrow road	222	O	X	O	X
			Going straight on the narrow road from the left vs turning left from the right wide road	223	O	X	O	X
			Going straight on the narrow road from the right road vs turning left from the left wide road	224	O	X	O	X
		Crossroads (the sign is only on one side)	Stop sign violation-related accident (turning left)	225	△	O	O	X
			Stop sign violation-related accident (going straight from left road)	226	△	O	O	X
			Stop sign violation-related accident (going straight from right road)	227	△	O	O	X
	Traffic light is only on one side	Traffic light is only on the direction where a car is going straight	Traffic light is only on the direction where a car is going straight	228	△	O	O	X
	Straight vs right-turn accident	Crossroads (based on the width of road)	Right-turn vs straight-related accident (same road width)	229	O	X	O	X
			Right-turn on the narrow road vs straight on the wide road	230	O	X	O	X
			Right-turn on the wide road vs straight on the narrow road	231	O	X	O	X

Continued

Accident case	Large category	Medium category	Small category	Category number	Whether split liability assessment is checkable with my video data	Necessity of opponent's video data	Front dash cam	Rear dash cam
		Crossroads (the stop sign is only on the one side)	Stop sign violation-related accident (turning right)	232	△	O	O	X
			Stop sign violation-related accident (going straight)	233	△	O	O	X
		Among others	Accident by changing a lane in the crossroads	233–1	△	O	O	X
	Left-turn vs left-turn accident	Crossroads (based on the width of road)	Left-turn from the right road vs left-turn from the left road (same road width)	234	O	X	O	X
			Left-turn on the narrow road vs left-turn on the wide road	235	O	X	O	X
		Crossroads (the stop sign is only on the one side)	Stop sign violation-related accident (turning left)	236	△	O	O	X
	Among other crossroads-related accidents	Passing accident on the crossroads		237	O	X	O	O
		The accident at the road where two cars can drive next to each other		238	O	X	O	O
		Crossroads where the angle is less than 90 degree for (left) right-turn		239	X	X	X	X
	T-junction road-related accidents	Straight vs (left) right-turn accident	Straight from the left road vs left-turn from the right road accident	240–220CO	O	X	O	X
			Straight from the right road vs left-turn from the left road accident	240–221CO	O	X	O	X
			Going straight on the wide road vs turning left on the narrow road	240–222CO	O	X	O	X
			Going straight on the narrow road from the left vs turning left from the right wide road	240–223CO	O	X	O	X
			Going stright on the narrow road from the right road vs turning left from the left wide road	240–224CO	O	X	O	X

Continued

Accident case	Large category	Medium category	Small category	Category number	Whether split liability assessment is checkable with my video data	Necessity of opponent's video data	Front dash cam	Rear dash cam
			Stop sign violation-related accident (turning left)	240–225CO	△	O	O	X
			Stop sign violation-related accident (going straight from left road)	240–226CO	△	O	O	X
			Stop sign violation-related accident (going straight from right road)	240–227CO	△	O	O	X
			Straight vs right-turn (same road width)	240–229CO	O	X	O	X
			Going straight on the wide road vs turning left on the narrow road	240–230CO	O	X	O	X
			Going straight on the narrow road from the left vs turning on the wide road	240–231CO	O	X	O	X
			Stop sign violation-related accident (turning right)	240–232CO	△	O	O	X
			Stop sign violation-related accident (going straight)	240–233CO	△	O	O	X
		Left-turn vs left-turn accident	Left-turn from the right road vs left-turn from the left road (same road width)	241–234CO	O	X	O	X
			Left-turn from the right narrow road vs left-turn from the left wide road	241–235CO1	O	X	O	X
			Left-turn from the left on the narrow road vs left-turn on the wide road	241–235CO2	O	X	O	X
			Stop sign violation-related accident (turning left)	241–236CO	△	O	O	X
	Car accidents of among other road types	Entering to the driveway from where is not the driveway		242	O	X	O	X
		Entering to where is not the driveway from the driveway		243	O	X	O	X
		Accident on the parking lot		244	X	X	X	X
		Second car accident		245	△	O	O	X
		Merging accident where the number of lanes decrease		246	O	X	O	X

Continued

Accident case	Large category	Medium category	Small category	Category number	Whether split liability assessment is checkable with my video data	Necessity of opponent's video data	Front dash cam	Rear dash cam
		Right (left)-turning accident happens cotemporaneously from two different lanes		247	O	X	O	X
		Opened door-related contact accident		248	X	X	X	X
		Wrong-way driving-related accident		249	O	X	O	X
		Intersection accident on the narrow road		249–1	O	X	O	X
		Passing accident happens where passing manner is prohibited		250	O	X	O	X
		Two drivers' overtaking behavior-related accident in where overtaking is prohibited		250–1	△	O	O	O
		Overtaking accident (center lane is the dotted)		251	O	X	O	X
		Changing lane-related accident		252	O	X	O	X
		Lane change-related accident (from solid lane to overtaking lane)		252–1	O	X	O	O
		Lane change happens simultaneously		252–2	△	O	O	X
		Lane change all of sudden during traffic congestion		252–3	△	O	O	O
		Lane change-related accident (passing safty zone)		252–4	△	O	O	O
		Collision betweem two cars while driving		253	△	O	O	O
		U-turn-related accident (straight vs U-turn)		254	△	O	O	X
		U-turn-related accident (right-turn vs U-turn)		254–1	△	O	O	X

Continued

Accident case	Large category	Medium category	Small category	Category number	Whether split liability assessment is checkable with my video data	Necessity of opponent's video data	Front dash cam	Rear dash cam
		U-turn-related accident (both cars are U turning)		254–2	△	O	O	X
		Car collision while pulling		255	O	X	O	O
		Left-turn vs right-turn		256	O	X	O	X
		Car departure after pulling		257	△	O	O	X
	Pavement marking violation accidents	Going straight on left-turn road mark		258	O	X	O	X
		Left-turn on straight road mark		259	O	X	O	X
		Right-turn on straight road mark		260	O	X	O	X
		Going straight on right-turn road mark		261	O	X	O	X
	Accidents on round-about	Roundabouts (one lane type)	Entering to roundabouts vs driving on the roundabouts	262	O	X	O	X
		Roundabouts (two lane type)	Driving second lane on the roundabouts vs changing first lane to the second lane	263	O	X	O	X
			Simultaneously entering to roundabouts	264	O	X	O	X
			Entering to roundabouts vs driving on the roundabouts	265	O	X	O	X
			Exiting from first roundabout lane vs entering to first roundabout	266	O	X	O	X
	Emergency car accident	Emergency car goes straight (signal violation)		267	X	X	X	X
		Emergency car's wrong-way driving on the left lane (1)		268	X	X	X	X
		Emergency car's wrong-way driving on the left lane (2)		269	X	X	X	X

Continued

Accident case	Large category	Medium category	Small category	Category number	Whether split liability assessment is checkable with my video data	Necessity of opponent's video data	Front dash cam	Rear dash cam
		Emergency car enters to the road (no signal)		270	X	X	X	X
		Emergency car overtakes others		271	X	X	X	X
		Emergency car changes the lane		272	X	X	X	X
		Emergency car vs a car that changes the lane		273	X	X	X	X
Highway	Merging road-related accident	Merging road-related accident		501	△	O	O	X
	Decrease in number of lane-related accident	Decrease in number of lane-related accident		502	△	O	O	X
	Lane change-related accident	Lane change to the fast lane		503	△	O	O	X
		Lane change while driving		504	△	O	O	X
	Rear-end car collision	Car collision while pulling		505	O	X	O	O
		Car collision while pulling over		506	O	X	O	O
		Car collision while driving		507	O	X	O	O
	Falling object-related accident	Falling object-related accident		508	X	X	X	X
	Pedestrian related-accident	Pedestrian related-accident (unreasonable walking behavior)		509	X	X	X	X
		Pedestrian related-accident (reasonable walking behavior)		510	X	X	X	X
	Lane change to the road shoulder	Lane change to the road shoulder		511	△	O	O	X

Received: February 1, 2023. **Revised:** June 25, 2023. **Accepted:** June 25, 2023