mathematics

MDPI

Article

Hybrid Traffic Accident Classification Models

Yihang Zhang [1] and Yunsick Sung [2,*]

[1] Department of Autonomous Things Intelligence, Dongguk University-Seoul, Seoul 04620, Republic of Korea
[2] Department of Multimedia Engineering, Dongguk University-Seoul, Seoul 04620, Republic of Korea
* Correspondence: sung@dongguk.edu; Tel.: +82-2-2260-3338

Abstract: Traffic closed-circuit television (CCTV) devices can be used to detect and track objects on roads by designing and applying artificial intelligence and deep learning models. However, extracting useful information from the detected objects and determining the occurrence of traffic accidents are usually difficult. This paper proposes a CCTV frame-based hybrid traffic accident classification model that enables the identification of whether a frame includes accidents by generating object trajectories. The proposed model utilizes a Vision Transformer (ViT) and a Convolutional Neural Network (CNN) to extract latent representations from each frame and corresponding trajectories. The fusion of frame and trajectory features was performed to improve the traffic accident classification ability of the proposed hybrid method. In the experiments, the Car Accident Detection and Prediction (CADP) dataset was used to train the hybrid model, and the accuracy of the model was approximately 97%. The experimental results indicate that the proposed hybrid method demonstrates an improved classification performance compared to traditional models.

Keywords: traffic accident classification; trajectory tracking; YOLO; Deep SORT; convolutional neural network; vision transformer

MSC: 68T99

Citation: Zhang, Y.; Sung, Y. Hybrid Traffic Accident Classification Models. *Mathematics* **2023**, *11*, 1050. https://doi.org/10.3390/math11041050

Academic Editor: Ivan Lorencin

Received: 27 December 2022
Revised: 8 February 2023
Accepted: 17 February 2023
Published: 19 February 2023

1. Introduction

The determination of accident-prone locations and accident-causing factors has received more attention owing to the increasing incidence of traffic accidents. Closed-circuit television (CCTV) devices play a vital role in recording global traffic, identifying vehicles, and analyzing the causes of accidents. Among the methods employed in this regard, the raw risk estimation approach is limited by the calculations involved in regression analysis [1]. Traditional machine learning approaches have been widely utilized for traffic accident detection and classification [2]. These approaches require rigorous data analysis and complex feature engineering, and most machine learning classification models need help to extract useful features from original input data [3]. These approaches have demonstrated limited performance in complex and dynamic traffic scenarios. The unpredictability of traffic accidents can lead to incorrect judgments, for example, mislabeling relatively safe locations as accident-prone locations. Consequently, accurate accident classification requires a thorough understanding of traffic scenes. An Artificial Neural Network (ANN) [4] processes input data by correlating the connections between multiple neurons with weights. The traffic accident classification approach based on ANNs has been demonstrated to approximate the relationship between complex nonlinear variables [5]. However, the ANN learning process is not visible and requires the analysis of a large number of parameters, which is why the results of traffic accident classification are not particularly reliable.

In the past decade, rapid developments have been made in Multi-Object Tracking (MOT), object detection, object tracking, and other approaches [6]. MOT estimates the motion of objects from video sequences and has been widely utilized in various applications, such as video stabilization [7], 3D reconstruction [8], pedestrian detection [9], and

vehicle detection [10]. MOT approaches based on CCTV frames could help improve road safety and the monitoring and evaluation of the causes of accidents. Object detection extracts relevant features by locating objects in video frames and drawing bounding boxes. Existing object detection research applies deep learning-based methods to detect objects using surveillance [11] or vehicle [12] camera footage. However, object detection cannot always handle complex traffic scenes. Object tracking is often conducted after object detection using the bounding box to track the object in consecutive frames and to obtain its trajectory [13,14]. Even in the presence of occlusion, popular object tracking approaches can maintain a high tracking accuracy. However, current object trajectory tracking-based traffic accident classification research is unable to determine the accident's location [15]. Owing to the complex motion of objects in real traffic scenes and an inability to consider the object trajectory's deep features, the accuracy of accident detection is poor. The traffic accident classification process is essential in understanding the cause of traffic accidents and deducing the connection between the object's trajectory and the accident's location.

Recent advances in deep learning have enabled the development of powerful and robust models for traffic accident classification. Convolutional Neural Networks (CNNs) are deep learning models suited for vision, detection, and classification tasks [16]. A CNN consists of multiple layers of convolutional filters that can extract and combine local features from the input image and generate more informative high-level representations [17,18]. Despite their impressive performance, the applications of CNNs in traffic accident classification are limited. CNNs may not effectively analyze traffic accident scenes with large or complex objects, low lighting, or other challenging conditions. Furthermore, CNNs cannot extract the spatial and temporal relationships between video frames, which could be crucial for accurately classifying traffic accidents. Many researchers used hybrid methods to overcome these disadvantages and to break through the limitations of a single model [19]. Hybrid methods combine multiple models or algorithms to create more powerful and robust systems by utilizing their respective strengths. Initially, the hybrid method was used primarily in Computer Vision (CV) [20], for instance, by combining CNNs and Support Vector Machine (SVM) [21]. It has since extended to other fields such as Natural Language Processing (NLP) [22] and recommender systems [23]. As more advanced machine learning models and algorithms have been developed, hybrid methods have been able to utilize them to achieve better performance. Additionally, the enhanced capabilities of graphical processors and computer power have made it possible to train more complex hybrid models [24], such as the combination of a CNN and transformer [25]. Overall, these developments have provided new possibilities for addressing the problem of traffic accident classification.

This paper proposes a hybrid method for traffic accident classification based on CCTV frames, which focuses on modeling fusion features to help determine the location of traffic accidents in CCTV frames. The proposed method consists of two main stages. First, a Trajectory Generator detects the objects and generates their trajectories. The Trajectory Generator consists of You Only Look Once (YOLOv5) [26] and Deep Simple Online and Real-time Tracking (Deep SORT) [27]. Next, the Traffic Accident Classifier outputs traffic accident classification results. The Traffic Accident Classifier comprises a CNN, a Vision Transformer (ViT) [28], and a Feature Fusion Network. The main contributions of this paper are summarized as follows:

- The proposed hybrid method utilizes CCTV frames as input to extract fusion features from one frame and the corresponding trajectories by applying ViT and CNN, which enhance the deduction of the relationship between frame and trajectory features to determine the area where traffic accidents occur. ViT and CNN can be combined as an end-to-end learning framework.
- This is the first attempt to use YOLOv5, Deep SORT, ViT, and CNN to classify traffic accidents. It closes the gap in the use of hybrid models in the field of traffic accident classification.

- We extracted 25 no-accident frames and 25 accident frames from each video in the Car Accident Detection and Prediction (CADP) dataset [29] to make a new CADP dataset that can be used for traffic accident classification tasks.
- The new CADP dataset was used to experimentally evaluate the effectiveness and accuracy of the proposed hybrid method, considering road and weather conditions.
- This paper mathematically defines models such as YOLOv5, CNN, and ViT, demonstrating their interpretability and providing their potential expansion.

The remainder of this paper is organized as follows. In Section 2, the algorithms used to detect traffic accidents are described. Section 3 proposes a hybrid method for detecting traffic accidents in CCTV frames using a Trajectory Generator and Traffic Accident Classifier. Section 4 details the experimental procedure and the results. Section 5 discusses the differences between the proposed hybrid method and the traditional traffic accident classification methods. Finally, Section 6 presents the conclusions of this paper.

2. Related Works

This section introduces and compares recent approaches in various industrial fields that primarily use MOT frameworks with the proposed method. Related research on traffic accident classification is then reviewed.

2.1. Multi-Object Tracking

With the wide application of deep learning, researchers have utilized deep features to establish new frameworks for MOT. SORT, introduced by Bewley et al. [30], achieved the best performance among the MOT algorithms. Subsequently, Wojke et al. [27] developed Deep SORT to integrate the appearance information of objects and to reduce the difficulty of tracking occluded objects in the SORT algorithm with a pre-trained association metric. Ricardo et al. [31] proposed a new approach for tracking and evaluating mobile robotics. Their approach defines eight new cost matrix formulas for correlating object tracking data. Deep SORT struggles to distinguish highly similar objects when applied to vehicle or pedestrian tracking and other fields. However, in the hybrid model, the motion and appearance information of the objects can be balanced by appropriate data associations. Multi-Class Deep (MCD-SORT) [32] is a granular computing approach in AlexNet [33]. The trajectory association restriction of tracked objects is placed in the same category, improving the MOT performance. However, the scenario's time information is lost when deriving trajectory features in the object feature detection stage, which hinders the accuracy of the hybrid model. To avoid the loss of feature attributes during extraction, the authors of [34] introduced an R-CNN attention mechanism to extract the scenario's global features and to make them available to the object detection stream. The attention mechanism can identify the feature attributes of objects. Similarly, Bai et al. [14] added an attention mechanism to the feature extraction network at the object detection stage, and attention mechanism channels were selectively built in the module to improve the utilization of feature attributes. However, this approach still leaves room for optimizing the determination of the object's trajectory.

To enhance the model generalization ability and to address the limitations of object trajectory determination, Shivani et al. [35] introduced a framework for autonomous tracking by unmanned aerial vehicles. The framework combines objects' appearances and motion features into the depth correlation matrix; this improves the detection of objects and the accuracy of the generated trajectory from the perspective of unmanned aerial vehicles. Additionally, some researchers have focused on establishing a trajectory tracking framework with consistent spatiotemporal relationships [36]. Fang et al. introduced a new fusion model based on a first-person view [37], called SSC-TAD, to fuse vehicle appearance, motion, and context consistency. Here, first person indicates that the camera was placed directly in front of the person or vehicle. Motion features are enhanced by embedding optical flow images to help with trajectory determination and object position prediction. However, the object size and illumination easily affect the SSC-TAD model. Meanwhile, owing to the limited field of view of the first-person vehicle camera, it is impossible to detect the

perspectives other than the front perspective, and the detection object is susceptible to occlusion, which increases the difficulty of trajectory determination and the issue of objects being ignored during the detection process. Huang et al. [38] introduced a two-stream convolutional network architecture that integrates spatial and temporal streams. The two-stream convolutional network architecture is a frame-based spatial pixel segmentation approach used during object detection to obtain accurate bounding boxes for multiple objects. The tracking algorithm improved the metric learning approach, strengthened the trajectory determination, and made the two-stream convolutional network more robust with regard to its tracking performance. The dataset used in this paper was based on drone video and fisheye cameras [39] with a top-down view. The fisheye camera has a nearly 180° field of view and can record at a high frame rate. However, this increases the data complexity and raises the requirement for GPU support. Thus, the training efficiency and accuracy of the network still need to be improved.

2.2. Traffic Accident Classification

Traffic accident classification calculates the probability that a traffic accident has occurred in each video frame. The scenarios can be classified as an accident or no-accident scenario by defining a threshold probability. Taccari et al. [40] described a novel approach for the crash and near-crash accident classification in videos. Their algorithm directly extracts accident features from an input video based on machine learning. In [41], collision and non-collision samples were preprocessed using traffic accident data. The long short-term memory (LSTM)-based LSTMDT model was introduced to detect the evolution of traffic conditions before a collision, representing the traffic trends across different time intervals. The performance of the LSTMDT model is better than ordinary machine-learning-based traffic accident classifiers. Kang et al. [42] proposed the Vision Transformer-Traffic Accident (ViT-TA) classifier to analyze traffic accidents based on first-person video data to improve autonomous vehicle safety. Ideally, the ViT-TA would accurately classify key situations around traffic accidents and automatically point out possible causes based on attention maps. Singh et al. [43] introduced a new automatic traffic accident classification framework that uses a denoising autoencoder without applying traditional deep feature representations from raw pixels. The accident probability was determined based on depth representation. Inspired by this, Vishnu et al. [44] performed hybrid stop filtering on traffic accident videos to remove noise and vehicle tracking using SVM [21] to detect accidents from traffic density and vehicle statistics data. However, this increases computational resources significantly. These models only focus on deep features that improve classification while ignoring the connection of accident-related objects or other features that constitute a traffic accident.

Therefore, the hybrid method proposed in this paper strengthens the connection between features through the Trajectory Generator and Traffic Accident Classifier stages. In particular, the Traffic Accident Classifier uses trajectory and frame features. Traffic Accident Classifiers have two primary advantages. First, not only is the trajectory position learned, but the fusion features in the frame emphasize objects with a strong correlation to the occurrence of traffic accidents. Compared to traditional approaches, most of it is trained on carefully selected datasets, resulting in a model with poor generalization ability. The Traffic Accident Classifier considers the relationship between trajectory and frame features, fuses the two features, and uses a large amount of data for learning to improve the accuracy of the classification results. Table 1 shows the differences between the previous traffic accident classification frameworks and our method by comparing the dataset, neural network, and model type.

Table 1. Differences between recent accident classification models and the proposed method.

Recent Related Research	LSTMDTR [41]	ViT-TA [42]	Stacked Autoencoder [43]	The Proposed Method
Dataset	Simulator	First-Person Video	CCTV	CCTV
Neural Networks	LSTM	Vision Transformer	Autoencoder	CNN, Vision Transformer
Model Types	Single Model	Single Model	Single Model	Hybrid Model

3. Traffic Accident Classification Model

In this section, we describe the proposed hybrid method for traffic accident classification utilizing a Trajectory Generator and Traffic Accident Classifier analyzing CCTV frames. The method extracts object trajectories and determines whether accidents have occurred.

3.1. Overview of Traffic Accident Classification Processes

The proposed hybrid method has two modules: a Trajectory Generator and Traffic Accident Classifier. The proposed hybrid method classifies traffic accidents by analyzing the extracted features of the frames and trajectories. The Trajectory Generator consists of two parts: *Bounding Box Detector* and *Trajectory Tracker.* The *Bounding Box Detector* draws 2D object bounding boxes in CCTV frames using the multi-object detection algorithm You Only Look Once (YOLO) [26]. The *Trajectory Tracker* takes 2D object bounding boxes as input and applies the Deep SORT algorithm [27] to calculate their trajectories.

The Traffic Accident Classifier consists of three parts: *Trajectory Analyzer*, *Frame Analyzer*, and *Feature Fusion Network*. The *Trajectory Analyzer* extracts 2D object trajectory features with a CNN. The *Frame Analyzer* extracts frame features using the ViT [28]. Next, the trajectory and frame features are fused, and then, the *Feature Fusion Network* detects accidents by superimposing the fused features with attention weights. Figure 1 shows the processes of using the Trajectory Generator and Traffic Accident Classifier to detect traffic accidents.

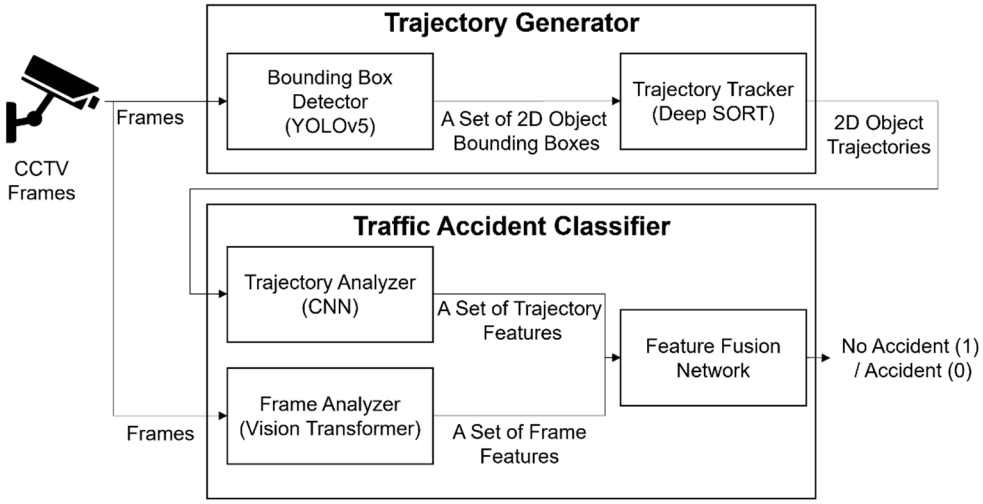


Figure 1. Processes of traffic accident classification with Trajectory Generator and Traffic Accident Classifier.

3.2. Mathematical Definition

The sequence of consecutive CCTV frames used as input is denoted as S. YOLOv5 maps S to a set of 2D object bounding boxes. B_i denotes a 2D object bounding box as a tuple containing five values $\left(x_i, y_i, w_i, h_i, v_i \times \frac{\alpha_i \cap \beta_i}{\alpha_i \cup \beta_i}\right)$, where i is the i-th 2D object bounding box

in the current CCTV frame and is used to estimate the location of a specific object detected in the current CCTV frame as expressed in Equation (1):

$$B_i = \left(x_i, y_i, w_i, h_i,\ v_i \times \frac{\alpha_i \cap \beta_i}{\alpha_i \cup \beta_i} \right) \tag{1}$$

where x_i and y_i denote the coordinates of the center of the 2D object bounding box, and w_i and h_i denote the width and height of the 2D object bounding box, respectively. $v_i \times \frac{\alpha_i \cap \beta_i}{\alpha_i \cup \beta_i}$ denotes the confidence score, where the 2D object bounding box of the ground truth is denoted as α_i, the predicted 2D object bounding box is denoted as β_i, and v_i denotes the variable indicating whether an object is detected or not in the i-th 2D object bounding box B_i. Specifically, v_i is 1 if an object is detected and 0 otherwise. The accuracy of the position of the predicted 2D object bounding box β_i is proportional to the Intersection Over Union (IOU) score between α_i and β_i, as a higher IOU score implies a closer match between the predicted 2D object bounding box β_i and 2D object bounding box α_i of the ground truth.

Deep SORT maps the set of 2D object bounding boxes to a set of 2D object trajectories. The equation for defining a 2D object trajectory by Deep SORT is $t_{m,\,n} = \lambda d^1(m,n) + (1-\lambda) d^2(m,n)$ [27], where λ denotes the hyperparameter that controls the association between $d^1(m,n)$ and $d^2(m,n)$ considering the n-th bounding box detection and the m-th track distribution, $d^1(m,n)$ denotes the Mahalanobis distance, and $d^2(m,n)$ denotes the smallest cosine distance. $d^1(m,n) = (d_n - y_m)^T S_m^{-1}(d_n - y_m)$ calculates the Mahalanobis distance between the n-th bounding box detection and the mean of the track distribution to associate the motion features of the corresponding object where d_n denotes the n-th bounding box detection, y_m denotes the mean of the track distribution, S_m denotes the covariance matrix, and $(y_m,\ S_m)$ denotes the m-th track distribution after making y_m and S_m be in the same measurement. Motion features are temporal information. $d^2(m,n) = min\{1 - r_n^T r_k^m \mid r_k^m \in R_m\}$, where the appearance features of each object within the bounding box detection d_n is denoted as r_n, r_n^T denotes the transposed r_n, index k denotes the index of the tracking object, r_k^m denotes the appearance feature of the k-th tracking object, and R_m denotes the appearance features of all the bounding box detections. Appearance features are spatial information. Therefore, the 2D object trajectory $t_{m,n}$ is obtained by linear weighting the λ of the Mahalanobis distance and the smallest cosine distance.

The CNN maps the set of 2D object trajectories to a set of one trajectory feature maps. The one trajectory feature map F by the CNN is F_j of the maximum of j, where F_j denotes the trajectory feature map of the j-th convolutional layer as expressed in Equation (2):

$$F_j = W_j \otimes F_{j-1} + b_j \tag{2}$$

where W_j denotes the weight matrix of the j-th convolutional kernel applied to the $(j-1)$-th feature map, b_j denotes the bias of the j-th convolutional kernel, and $\otimes$ denotes the convolution operation.

The input of the ViT maps the sequence S of consecutive CCTV frames to a set of a frame feature map. The frame feature map P by ViT is as shown in Equation (3):

$$P = MLP(LN(MHA(LN(S)))) \tag{3}$$

where MLP denotes a multi-layer perceptron, LN denotes layer normalization, and MHA denotes multi-head attention.

The *Feature Fusion Network* takes as input both a set of trajectory feature maps and a set of frame feature maps. The output O of the traffic accident classification by the *Feature Fusion Network* is as expressed in Equation (4):

$$O = \sigma(FC(FC(F_j)^T \times FC(P))) \tag{4}$$

where $\sigma(\cdot)$ denotes a sigmoid function, which maps the result to a value between 0 and 1, and *FC* denotes a fully connected layer. After passing through one fully connected layer, the transposed trajectory feature map denoted by $FC(F_j)^T$ and multiplied by $FC(P)$ is passed through another fully connected layer to obtain the fusion feature $(FC(F_j)^T \times FC(P))$. The fusion feature is processed by the sigmoid function $\sigma(\cdot)$ to obtain of the traffic accident classification result.

3.3. Traffic Accident Classification Models

The proposed hybrid method identifies whether accidents are included by extracting trajectories. In the Trajectory Generator, the size of the CCTV frame input to the YOLOv5 network is $224 \times 224 \times 3$. A set of CCTV frame features is extracted by the YOLOv5 network, in which all dynamic objects, such as vehicles and pedestrians, are detected on the road. YOLOv5 is pre-trained on the CADP dataset [29], a dataset for traffic accident analysis. The dataset comprises 1416 video segments collected from YouTube, including full spatial and temporal annotations. In this paper, the YOLOv5 network outputs a set of 2D object bounding boxes using three scales, $20 \times 20 \times 255$, $40 \times 40 \times 255$, and $80 \times 80 \times 255$, which support the different sizes of objects owing to depth disparity in CCTV frames. A trajectory is deduced for each object via the Deep SORT algorithm, starting from the second CCTV frame with 2D object bounding boxes.

The Traffic Accident Classifier uses a CNN to extract a set of trajectory feature, which are compressed down to 1000 dimensions by the subsequent linear layer. The CNN is modified based on the VGG16 architecture. In this paper, we modified the VGG16 structure by removing the softmax layer. The CNN receives a set of 2D object trajectories of size 224×224 as input. It consists of five collections of convolutional layers, where filters with small receptive fields are utilized 3×3. The convolution stride is fixed at 1 pixel, and the spatial resolution is preserved after convolution. Simultaneously, pooling is carried out by five max-pooling layers, where each convolutional layer collection is followed by one max-pooling layer performed over a 2×2 kernel with a stride of 2 pixels. Three fully connected layers follow the five collections of convolutional and max-pooling layers. The first two fully connected layers have 4096 channels, and the first fully connected layer accepts the output size of the last max-pooling layer as the input. The last fully connected layer has 1000 channels, which compress the feature vector size for processing in the *Feature Fusion Network*. In addition, ViT is vanilla ViT and helps determine the location of traffic accidents along with the Traffic Accident Classifier. CCTV frames with a size of 224×224 pixels are flattened into patches and then embedded as input, utilizing ViT to extract a set of frame features. The multi-head attention mechanism in ViT assigns higher weights to the location of traffic accidents to aggregate the attention score. The frame features are compressed into 1000 dimensions by the subsequent linear layer for the *Feature Fusion Network*.

Next, the *Feature Fusion Network* extracts a set of trajectory features by CNN and a set of frame features by ViT. The network combines two types of features to improve the classification accuracy of the proposed method and to enhance the model's ability to understand feature maps. The fusion features are obtained by the matrix multiplying the transposed frame and trajectory features. Finally, the fusion features are compressed to 1 dimension by another fully connected layer, and a traffic accident classification value is obtained using the sigmoid function. The result of the proposed method is expressed as an Accident (0) or No Accident (1) case based on the classification value. When the traffic accident classification value is more than 0.5, considered as 1, the classification result is not included in the current frame. In contrast, when the traffic accident classification value is less than or equal to 0.5, considered as zero, a traffic accident has been detected and included in the current frame. Figure 2 shows the architecture of the proposed hybrid traffic accident classification method.

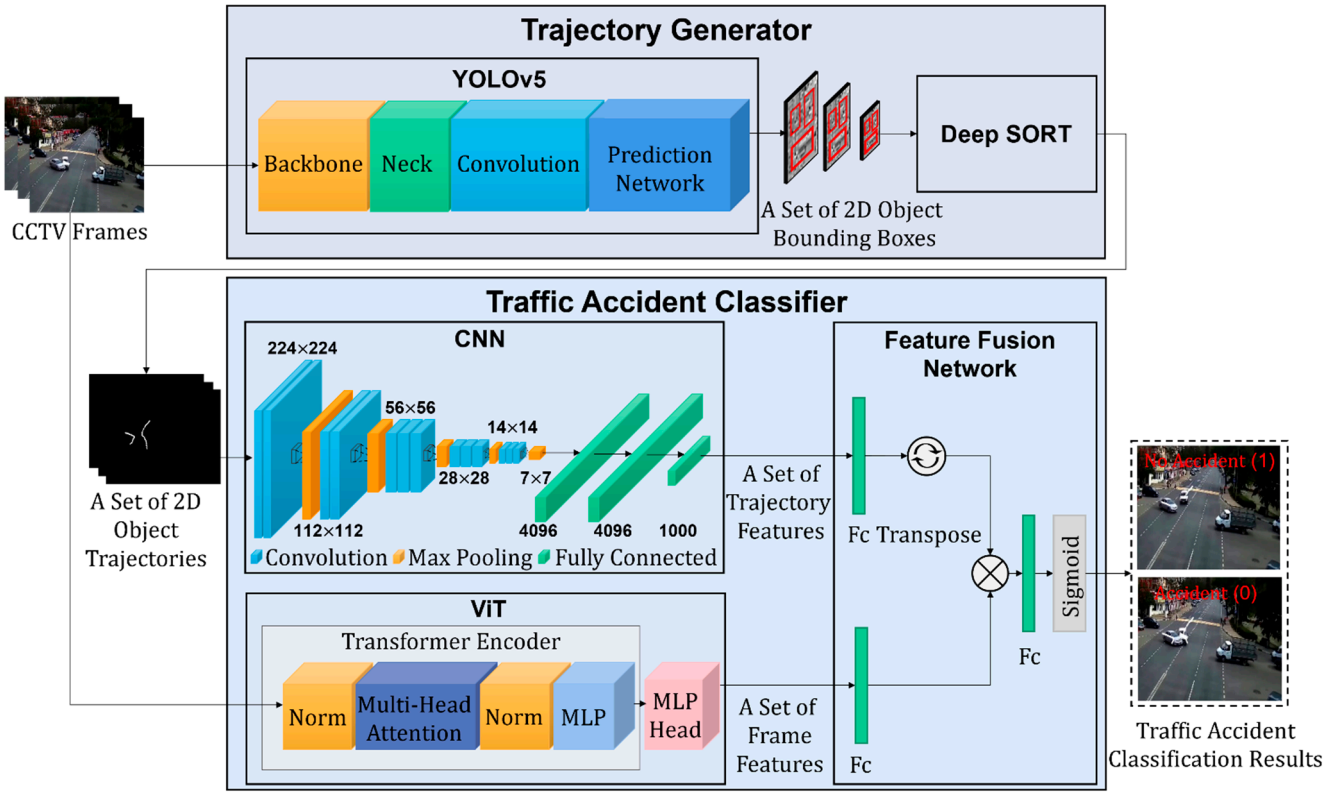


Figure 2. Model of the proposed hybrid traffic accident classification method.

4. Experiment

In this section, we describe the experimental objectives and provide detailed parameters for the hybrid model training and traffic accident classification. This section includes the details of an ablation experiment, and we compare the results obtained with the classification results. Finally, we show the visualization of the attention regions for CCTV frames extracted using ViT.

4.1. Experimental Objectives and Environment

Experiments were conducted to ascertain the accuracy of traffic accident classification utilizing the Trajectory Generator and the Traffic Accident Classifier. An ablation experiment was conducted to verify whether the proposed hybrid method improves the classification performance compared to the CNN-based Traffic Accident Classifier.

In the Trajectory Generator, YOLOv5 was chosen as the backbone network. YOLOv5 was the fastest and most accurate detector in the world, beating all SOTA benchmarks at the time. YOLOv5 was also trained on the COCO dataset of annotated images and achieved 48.2% average precision at a speed of 13.7 ms. This ensured that it could provide accurate object detection in the CADP dataset, thus making it well-suited for the Trajectory Generator in the proposed hybrid model.

We tested the proposed hybrid method using CCTV frames as input. The CCTV frames were passed to a Trajectory Generator to generate 2D object trajectories, which were uniformly cropped to 224 × 224 pixels and then passed to the CNN in a Traffic Accident Classifier. At the same time, the ViT in the Traffic Accident Classifier received CCTV frames, which were also uniformly cropped to 224 × 224 pixels. The training parameters for the proposed hybrid method are listed in Table 2. During the training process, the parameters were adjusted according to the accuracy and loss values obtained, until they resulted in the parameters of the proposed hybrid model. The batch size was set to 40, evenly dividing the total training data of 4000 and ensuring that each training batch contained the same number of samples. It is a common practice to choose the batch size and learning rate in

direct proportion. Therefore, based on the smaller batch size of 40, we set a small learning rate of 1×10^{-6}, which allowed the model to make small updates to the weights in the early stages of training. The learning rate decayed according to the cosine method [45], with a decay rate of 1×10^{-4}, as represented by Equation (5):

$$a_t = \frac{1}{2} a_0 (1 + \cos(t_\pi / T)) \quad (5)$$

where a_t is the decayed learning rate, a_0 is the initial learning rate, t is the current epoch, and T is the total number of training epochs. The learning rate decay can help the model avoid overshooting the optimal solution during the training process. When the training process exceeded 500 epochs, the proposed hybrid model suffered from overfitting. Therefore, the total number of training epochs was set to 500 with 100 steps per epoch for the proposed hybrid model to perform enough updates while preventing overfitting. The Adam optimizer was used to smooth out the gradients. The sigmoid function was chosen as the objective function.

Table 2. Parameters for training the proposed hybrid method.

Hyperparameter	Value
Input size of CCTV frames	224×224
Input size of 2D object trajectories	224×224
Batch size	40
Learning rate	1×10^{-6}
Decay learning rate	1×10^{-4}
Total epochs	500
Steps per epoch	100
Optimizer	Adam
Objective function	sigmoid function

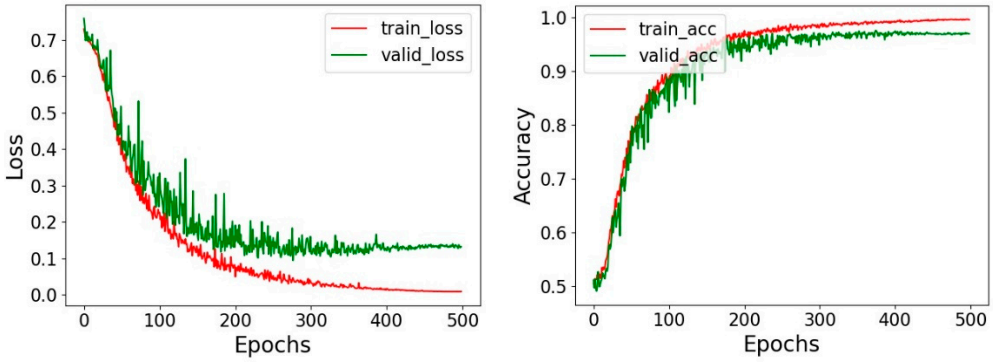
The experiments were conducted on a Windows 10 machine with an Intel i7-6850K processor and four Nvidia Titan RTX GPUs with 48 GB of memory. CNN, ViT, and the CNN-based classifier were implemented in Python 3.6, using the PyTorch deep learning library version 11.3 to exploit the GPU's computing capabilities.

4.2. Experimental Data

The CADP dataset [29] comprised video data related to car accidents. It was designed for use in machine learning, deep learning, and data mining research, with a focus on solving the problem of data labeling in public traffic accident data for detecting and predicting traffic accidents. The CADP dataset contains 1416 accident detection sample videos, including information about the car accidents, road conditions, weather conditions, and traffic flow. The average length of the video data in the CADP dataset was 366 frames, and the longest comprised 554 frames. We took 100 videos from the CADP dataset, and from each video, we extracted 25 frames before and 25 frames after the traffic accident. In total, 5000 frames were obtained as inputs for the Trajectory Generator and ViT. Furthermore, the Trajectory Generator utilized 5000 frames to generate 5000 corresponding 2D object trajectories as the CNN input. To ensure the reliability and robustness of the proposed hybrid model and to avoid including validation in the training dataset, the dataset was split using cross-validation. More specifically, 90% of the input data was used for training, and the remaining 10% was used as validation datasets. This allowed for the evaluation of the performance of the proposed hybrid method on unknown data and avoided overfitting.

4.3. Experimental Results

The accuracy and loss convergence plots are shown in Figure 3 to illustrate the training and validation results of the proposed hybrid method. The proposed hybrid training method required 500 epochs. As shown in Figure 3a, the initial training loss of the proposed hybrid method was approximately 0.73, while the initial validation loss was approximately 0.76. After 500 epochs of training, the training loss gradually converged to 0.007, whereas the validation loss converged to 0.13. Figure 3b shows the accuracy convergence for the training and validation datasets over the training. During the first epoch, the training accuracy was approximately 0.51. After 500 training epochs, the accuracy rate increased to approximately 0.99. The validation accuracy during the first epoch was similar to the training accuracy at approximately 0.50. After 500 training epochs, the validation accuracy reached approximately 0.97. The results show that the accuracy of the proposed hybrid method improves with more training epochs. Figure 3 indicates that the model could learn and generalize well on the training and validation datasets.



(**a**) (**b**)

Figure 3. Training and validation results of the proposed hybrid method. (**a**) Loss of the proposed hybrid method. (**b**) Accuracy of the proposed hybrid method.

Figure 4 presents a confusion matrix that illustrates the proposed hybrid method's ability to distinguish "Accident" from "No Accident" in 1000 test CCTV frames. The confusion matrix was divided into four quadrants, each representing a combination of the predicted and true labels. In the matrix, we used a blue gradient to indicate the number of frames that fall into each category. A darker color indicates more frames in that category. The proposed hybrid model demonstrates strong classification performance on the test CCTV frames. The majority of the test CCTV frames were accurately classified into the "Accident" or "No-Accident" category. However, the performance of the proposed hybrid method was not perfect, and there were few misclassifications. In particular, the proposed hybrid method was more likely to misclassify a "No-Accident" CCTV frame as an "Accident." This suggests that the proposed hybrid method was slightly more conservative in its decision making and was more likely to predict an accident when there was uncertainty.

Table 3 shows the traffic accident classification evaluation indicator results for the proposed hybrid model. Table 3 presents four metrics: accuracy, precision, recall, and f1-score for evaluating the performance of the proposed hybrid model. Accuracy represents the proportion of correct predictions made by the proposed hybrid method, while precision represents the proportion of true positives among all positive predictions. The recall represents the proportion of true positives among all actual positives. The f1-score is a combination of precision and recall. These metrics were calculated based on the confusion matrix. The values of accuracy, precision, and f1-score were all greater than 0.95, and recall reached 0.943, indicating that the proposed hybrid model was able to accurately classify the majority of the test CCTV frames. Additionally, four metrics for the proposed hybrid

method all scored higher than 0.94 on test data, indicating that the classification results of the hybrid method were reliable and accurate.

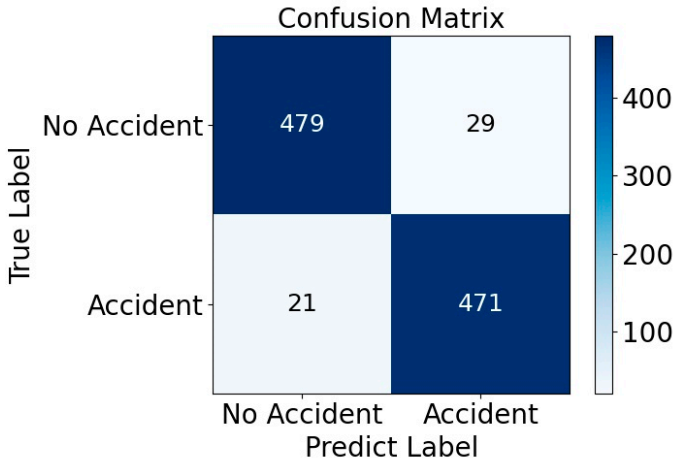


Figure 4. Confusion matrix for the proposed hybrid method.

Table 3. Evaluation indicator results for traffic accident classification.

Evaluation Indicators	Results
Accuracy	0.950
Precision	0.958
Recall	0.943
F1-score	0.95

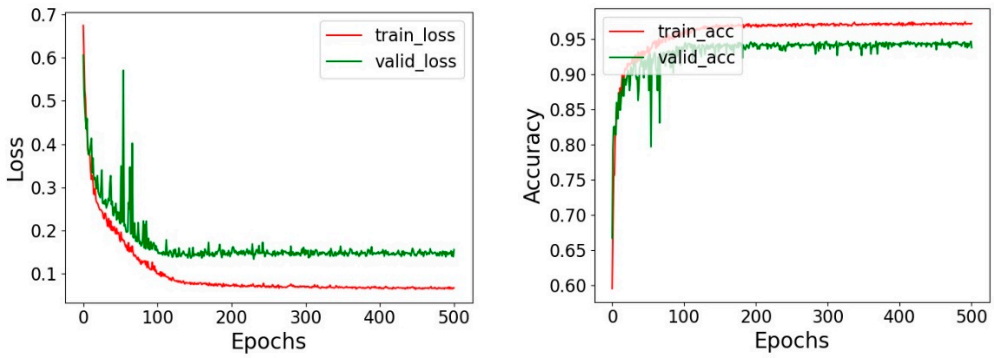
4.4. Ablation Experimental Results

The CNN-based traffic accident classifier utilized in the ablation experiments had a specific architecture and set of parameters optimized for traffic accident classification. Table 4 lists the parameters used for training the CNN-based classifier; these have undergone modifications based on the parameters of the proposed hybrid method. The CNN-based classifier utilized 2D object trajectories of size 224×224 as inputs and consisted of multiple convolutional layers and max-pooling layers. The convolutional layers utilized kernels of size 5×5 instead of 3×3 to obtain a larger receptive field and were followed by max-pooling layers with kernel sizes of 2×2 to preserve the main features while reducing dimensionality. We used Adam for stochastic optimization of the CNN-based classifier. The CNN-based classifier model is small and has fewer training parameters than our model. The loss converged quickly without the need for learning rate decay. The learning rate was set to 1×10^{-5}. The batch size was set at 40. The softmax function was used as the training objective function, allowing the CNN to detect whether accidents were included in the given 2D object trajectories. The CNN-based traffic accident classifier was trained for 500 epochs, each consisting of 100 training steps.

To assess the importance of the ViT and its impact on the overall performance of the proposed hybrid method, ablation experiments were performed. In the ablation experiments, ViT was removed, and the traffic accident classifier based on CNN was evaluated. The results of the ablation experiments are shown in Figure 5. As shown in Figure 5a, when ViT was removed, and the CNN-based traffic accident classifier reached a training loss of 0.07 and a validation loss of 0.15 after 500 epochs of training. Figure 5b shows the accuracy of the CNN-based traffic accident classifier. Even after 500 epochs of training, the highest training accuracy was only approximately 0.97, whereas the validation accuracy reached 0.94.

Table 4. Parameters for training CNN-based traffic accident classifier.

Hyperparameter	Value
Kernel size of convolutional layers	5×5
Kernel size of max-pooling layers	2×2
Input size	224×224
Batch size	40
Learning rate	1×10^{-5}
Total epochs	500
Steps per epoch	100
Optimizer	Adam
Objective function	softmax function



(**a**) (**b**)

Figure 5. Training and validation results of CNN-based traffic accident classifier. (**a**) Loss of CNN-based traffic accident classifier. (**b**) Accuracy of CNN-based traffic accident classifier.

The results of the ablation experiments demonstrated the significance of ViT in the overall performance of the proposed hybrid method for traffic accident classification. When ViT was removed, the performance was poor. The accuracy of the CNN also decreased when compared with the proposed hybrid method. The results suggest that the ViT was crucial for improving the traffic accident classification performance of the proposed hybrid method.

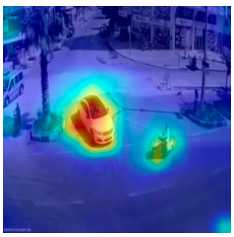
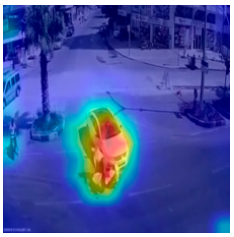
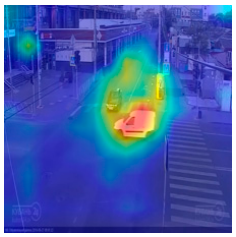
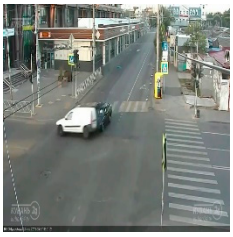
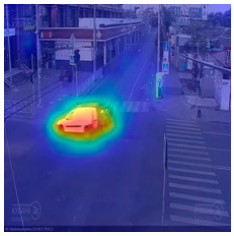
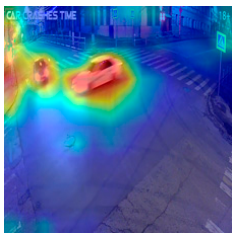
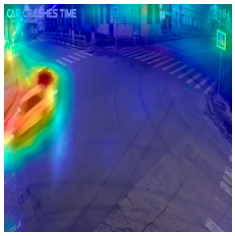
4.5. Visual Interpretation of ViT

ViT was added to the CNN-based traffic accident classification model, and its ability in improving the classification performance was demonstrated in ablation experiments. However, we still needed to ascertain how the ViT model utilized a multi-head attention mechanism to focus on objects in traffic scenes and the area where accidents occurred in CCTV frames. The multi-head attention mechanism was a critical component of the ViT. The attention mechanism allows ViT to consider different parts of the input CCTV frames simultaneously and assign them different attention weights.

As shown in Table 5, the multi-head attention mechanism in ViT was similar for the no-accident and accident classes. Specifically, for the no-accident visualization results, the multi-head attention was focused on objects on the road that were potentially dangerous, which were given high multi-head attention values and which are represented in deep red. Other objects on the road, such as signs, manhole covers, and streetlights, were given lower multi-head attention values and are represented in green. Yellow represents intermediate multi-head attention values. In the accident visualization results, the multi-head attention was focused on where the accident occurred, giving the area high multi-head attention values, which are represented as deep red. Other objects on the road were given lower multi-head attention values and are represented as green or yellow, depending on their relevance to the accident. Overall, the multi-head attention mechanism in ViT allows the

traffic accident classification model to focus on the most relevant parts of the input CCTV frames for the traffic accident classification task.

Table 5. Visualization results of ViT in CCTV frames.

No Accident		**Accident**	
CCTV Frames	**Visualization Result**	**CCTV Frames**	**Visualization Result**
			
			
			

5. Discussion

This paper establishes a new CADP dataset that can be used for traffic accident classification tasks and details experiments conducted with the new CADP dataset. Although the experimental process only covered a single dataset, this dataset had low CCTV video clarity, varying CCTV frame sizes, and poor camera angles. In addition, the new CADP dataset includes challenging CCTV frames such as snowy, rainy, and nighttime conditions. The LSTM-based method considering different temporal resolutions (LSTMDTR) [41] uses CCTV frames as training data, but the CCTV frames are from a fixed camera angle, which helps LSTMDTR solve traffic accident classification tasks. The Vision Transformer-Traffic Accident (ViT-TA) [42] method classifies traffic accidents based on a first-person camera view. However, it is difficult to determine the exact location of traffic accidents. The stacked autoencoder [43] also uses a large number of CCTV frames for training, increasing its computational cost and resulting in lower model efficiency. In contrast, our proposed hybrid method has two main advantages. First, by using CCTV frames from our new CADP dataset as training data where each accident is captured by a different camera with its own angle, the proposed hybrid method is enabled to accurately classify whether each frame contains any traffic accident or not, after determining the location of the traffic accidents. Second, this paper adopts a hybrid method that uses object trajectories to solve the traffic accident classification task. The proposed hybrid method analyzes a fusion of features from a frame and its corresponding object trajectories, enhancing the inference of the relationship between the frame and the object trajectories. Most existing traffic

accident classification methods utilize single-camera datasets that are not publicly available and are sample-limited [41,42]. If multiple cameras are not available, the accuracy of the proposed hybrid method may be affected, as it is trained to consider multiple camera angles. However, compared to existing traffic accident classification methods, the proposed hybrid method remains versatile and capable of handling challenging traffic scenarios and can effectively prevent overfitting in limited samples. In the absence of multiple cameras, other training data can be used. For example, simulation data generated by a simulator can be used to imitate conditions where multiple cameras may be present, and transfer learning can be employed to fine-tune a pre-trained model on a smaller dataset, reducing the reliance on large amounts of training data. The proposed hybrid method requires fewer computational resources than the previous methods, and by relying on novel hybrid models and better CADP datasets, it remains competitive. This paper addresses the lack of a hybrid method in the field of traffic accident classification. In the future, we plan to conduct more experiments to validate the superiority of the proposed hybrid method, for example, by comparing it to other hybrid models and state-of-the-art traffic accident classification research.

6. Conclusions

This paper proposes a CCTV frame-based hybrid method for classifying traffic accidents. First, in the Trajectory Generator stage, all dynamic objects on roads in the CCTV frames are detected through the YOLOv5 network, and 2D object bounding boxes are drawn around them. Then, the trajectories of all the 2D object bounding boxes are obtained using the Deep SORT algorithm. Second, in the Traffic Accident Classifier stage, a CNN extracts high-level features from the trajectories, and the ViT extracts more high-level features from the CCTV frames. Finally, the proposed hybrid method utilizes the *Feature Fusion Network* to extract the fusion features from the output of the second stage. It then determines whether a traffic accident has occurred. An ablation experiment was conducted to evaluate the contribution of ViT to the proposed hybrid method. The results demonstrated that the accuracy of the CNN-based traffic accident classification model, without ViT, was approximately 2% lower than that of the proposed hybrid method. Moreover, the proposed hybrid method is not limited to the CCTV frames used in this paper. With further research and development, the proposed hybrid method can be extended to other data sources for traffic accident classification, such as fish-eye cameras or black boxes in vehicles.

Author Contributions: Conceptualization, Y.Z. and Y.S.; methodology, Y.Z. and Y.S.; software, Y.Z. and Y.S.; validation, Y.Z. and Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Korea Institute of Police Technology (KIPoT) grant funded by the Korea government (KNPA) (No. 092021D75000000, AI driving ability test standardization and evaluation process development).

Data Availability Statement: Data available in a publicly accessible repository that does not issue DOIs. Publicly available datasets were analyzed in this study. This data can be found here: [https://ankitshah009.github.io/accident_forecasting_traffic_camera, accessed on 21 July 2022].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sun, D.; Ai, Y.; Sun, Y.; Zhao, L. A Highway Crash Risk Assessment Method based on Traffic Safety State Division. *PLoS ONE* **2020**, *15*, e0227609. [CrossRef]
2. Bokaba, T.; Doorsamy, W.; Paul, B.S. Comparative Study of Machine Learning Classifiers for Modelling Road Traffic Accidents. *Appl. Sci.* **2022**, *12*, 828. [CrossRef]
3. Pessach, D.; Shmueli, E. A Review on Fairness in Machine Learning. *ACM Comput. Surv. (CSUR)* **2022**, *55*, 1–44. [CrossRef]
4. Jain, A.K.; Mao, J.; Mohiuddin, K.M. Artificial Neural Networks: A Tutorial. *Computer* **1996**, *29*, 31–44. [CrossRef]
5. Alkheder, S.; Taamneh, M.; Taamneh, S. Severity Prediction of Traffic Accident Using An Artificial Neural Network. *J. Forecast.* **2017**, *36*, 100–108. [CrossRef]

6. Zaidi, S.S.A.; Ansari, M.S.; Aslam, A.; Kanwal, N.; Asghar, M.; Lee, B. A Survey of Modern Deep Learning Based Object Detection Models. *Digit. Signal Process.* **2022**, *126*, 103514. [CrossRef]
7. Mitchel, T.W.; Wulker, C.; Kim, J.; Ruan, S. Quotienting Impertinent Camera Kinematics for 3D Video Stabilization. In Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 29 October–1 November 2019.
8. Deng, R.; Yang, H.; Asad, Z.; Zhu, Z.; Wang, S.; Wheless, L.E.; Fogo, A.B.; Huo, Y. Dense Multi-Object 3D Glomerular Reconstruction and Quantification on 2D Serial Section Whole Slide Images. *Med. Imaging 2022 Digit. Comput. Pathol.* **2022**, *12039*, 83–90.
9. Feng, X.; Wu, H.M.; Yin, Y.H.; Lan, L.B. CGTracker: Center Graph Network for One-Stage Multi-Pedestrian-Object Detection and Tracking. *J. Comput. Sci. Technol.* **2022**, *37*, 626–640. [CrossRef]
10. Yin, G.; Yu, M.; Wang, M.; Hu, Y.; Zhang, Y. Research on Highway Vehicle Detection Based on Faster R-CNN and Domain Adaptation. *Appl. Intell.* **2022**, *52*, 3483–3498. [CrossRef]
11. Chung, T.Y.; Cho, M.; Lee, H.; Lee, S. SSAT: Self-Supervised Associating Network for Multiobject Tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 7858–7868. [CrossRef]
12. Ćorović, A.; Ilić, V.; Đurić, S.; Marijan, M.; Pavković, B. The Real-Time Detection of Traffic Participants Using YOLO Algorithm. In Proceedings of the 2018 IEEE Telecommunications Forum (TELFOR), Belgrade, Serbia, 20–21 November 2018; pp. 1–4.
13. Ulutan, O.; Rallapalli, S.; Srivatsa, M.; Torres, C.; Manjunath, B.S. Actor Conditioned Attention Maps for Video Action Detection. In Proceedings of the 2020 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; pp. 527–536.
14. Bai, C.; Gong, Y.; Cao, X. Pedestrian Tracking and Trajectory Analysis for Security Monitoring. In Proceedings of the 5th IEEE Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 12–14 June 2020; pp. 1203–1208.
15. Yang, D.; Wu, Y.; Sun, F.; Chen, J.; Zhai, D.; Fu, C. Freeway Accident Detection and Classification Based on the Multi-Vehicle Trajectory Data and Deep Learning Model. *Transp. Res. Part C Emerg. Technol.* **2021**, *130*, 103303. [CrossRef]
16. Song, W.; Li, D.; Sun, S.; Zhang, L.; Xin, Y.; Sung, Y.; Choi, R. 2D&3DHNet for 3D Object Classification in LiDAR Point Cloud. *Remote Sens.* **2022**, *14*, 3146.
17. Tian, Y.; Song, W.; Chen, L.; Fong, S.; Sung, Y.; Kwak, J. A 3D Object Recognition Method from LiDAR Point Cloud Based on USAE-BLS. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 15267–15277. [CrossRef]
18. Qiu, L.; Li, S.; Sung, Y. 3D-DCDAE: Unsupervised Music Latent Representations Learning Method Based on A Deep 3D Convolutional Denoising Autoencoder for Music Genre Classification. *Mathematics* **2021**, *9*, 2274. [CrossRef]
19. Ramaswamy, S.L.; Chinnappan, J. RecogNet-LSTM+CNN: A Hybrid Network with Attention Mechanism for Aspect Categorization and Sentiment Classification. *J. Intell. Inf. Syst.* **2022**, *58*, 379–404. [CrossRef]
20. Niu, X.X.; Suen, C.Y. A Novel Hybrid CNN–SVM Classifier for Recognizing Handwritten Digits. *Pattern Recognit.* **2012**, *45*, 1318–1325. [CrossRef]
21. Hearst, M.A.; Dumais, S.T.; Osuna, E.; Platt, J.; Scholkopf, B. Support Vector Machines. *IEEE Intell. Syst. Appl.* **1998**, *13*, 18–28. [CrossRef]
22. Yin, D.; Dong, L.; Cheng, H.; Liu, X.; Chang, K.W.; Wei, F.; Gao, J. A Survey of Knowledge-Intensive NLP with Pre-Trained Language Models. *arXiv* **2022**, arXiv:2202.08772.
23. Chen, R.; Hua, Q.; Chang, Y.S.; Wang, B.; Zhang, L.; Kong, X. A Survey of Collaborative Filtering-Based Recommender Systems: From Traditional Methods to Hybrid Methods based on Social Networks. *IEEE Access* **2018**, *6*, 64301–64320. [CrossRef]
24. Fang, J.; Lin, H.; Chen, X.; Zeng, K. A Hybrid Network of CNN and Transformer for Lightweight Image Super-Resolution. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–23 June 2022; pp. 1103–1112.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. Available online: https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (accessed on 10 August 2022).
26. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
27. Wojke, N.; Bewley, A.; Paulus, D. Simple Online and Real-Time Tracking with A Deep Association Metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649.
28. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
29. Shah, A.P.; Lamare, J.B.; Nguyen-Anh, T.; Hauptmann, A. CADP: A Novel Dataset for CCTV Traffic Camera-Based Accident Analysis. In Proceedings of the 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 27–30 November 2018; pp. 1–9.
30. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple Online and Realtime Tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468.
31. Pereira, R.; Carvalho, G.; Garrote, L.J.; Nunes, U. Sort and Deep-SORT Based Multi-Object Tracking for Mobile Robotics: Evaluation with New Data Association Metrics. *Appl. Sci.* **2022**, *12*, 1319. [CrossRef]
32. Pramanik, A.; Pal, S.K.; Maiti, J.; Mitra, P. Granulated RCNN and Multi-Class Deep SORT for Multi-Object Detection and Tracking. *IEEE Trans. Emerg. Top. Comput. Intell.* **2021**, *6*, 171–181. [CrossRef]

33. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*. Available online: https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (accessed on 11 September 2022).
34. Le, T.N.; Ono, S.; Sugimoto, A.; Kawasaki, H. Attention R-CNN for Accident Detection. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), Melbourne, Australia, 7–11 September 2020; pp. 313–320.
35. Kapania, S.; Saini, D.; Goyal, S.; Thakur, N.; Jain, R.; Nagrath, P. Multi Object Tracking with UAVs Using Deep SORT and YOLOv3 RetinaNet Detection Framework. In Proceedings of the 1st ACM Workshop on Autonomous and Intelligent Mobile Systems (AIMS'20), New York, NY, USA, 22 January 2020; pp. 1–6.
36. Fang, J.; Qiao, J.; Bai, J.; Yu, H.; Xue, J. Traffic Accident Detection via Self-Supervised Consistency Learning in Driving Scenarios. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 9601–9614. [CrossRef]
37. Pirsiavash, H.; Ramanan, D. Detecting Activities of Daily Living in First-Person Camera Views. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 2847–2854.
38. Huang, X.; He, P.; Rangarajan, A.; Ranka, S. Intelligent Intersection: Two-Stream Convolutional Networks for Real-Time Near-Accident Detection in Traffic Video. *ACM Trans. Spat. Algorithms Syst. (TSAS)* **2020**, *6*, 1–28. [CrossRef]
39. Wei, J.; Li, C.F.; Hu, S.M.; Martin, R.R.; Tai, C.L. Fisheye Video Correction. *IEEE Trans. Vis. Comput. Graph.* **2011**, *18*, 1771–1783. [CrossRef]
40. Taccari, L.; Sambo, F.; Bravi, L.; Salti, S.; Sarti, L.; Simoncini, M.; Lori, A. Classification of Crash and Near-Crash Events from Dashcam Videos and Telematics. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems, Maui, HI, USA, 4–7 November 2018; pp. 2460–2465.
41. Jiang, F.; Yuen, K.K.R.; Lee, E.W.M. A Long Short-Term Memory-Based Framework for Crash Detection on Freeways with Traffic Data of Different Temporal Resolutions. *Accid. Anal. Prev.* **2020**, *141*, 105520. [CrossRef]
42. Kang, M.; Lee, W.; Hwang, K.; Yoon, Y. Vision Transformer for Detecting Critical Situations and Extracting Functional Scenario for Automated Vehicle Safety Assessment. *Sustainability* **2022**, *14*, 9680. [CrossRef]
43. Singh, D.; Mohan, C.K. Deep Spatio-Temporal Representation for Detection of Road Accidents Using Stacked Autoencoder. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 879–887. [CrossRef]
44. Maha Vishnu, V.C.; Rajalakshmi, M.; Nedunchezhian, R. Intelligent Traffic Video Surveillance and Accident Detection System with Dynamic Traffic Signal Control. *Clust. Comput.* **2018**, *21*, 135–147. [CrossRef]
45. Gotmare, A.; Keskar, N.S.; Xiong, C.; Socher, R. A Closer Look at Deep Learning Heuristics: Learning Rate Restarts, Warmup and Distillation. *arXiv* **2018**, arXiv:1810.13243.