# Traffic Accident Recognition in First-Person Videos by Learning a Spatio-Temporal Visual Pattern

Kyung Ho Park
*Data Group*
*SOCAR*
Seoul, Republic of Korea
kp@socar.kr

Dong Hyun Ahn
*School of Cybersecurity*
*Korea University*
Seoul, Republic of Korea
jackie0304@korea.ac.kr

Huy Kang Kim
*School of Cybersecurity*
*Korea University*
Seoul, Republic of Korea
cenda@korea.ac.kr

2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring) | 978-1-7281-8964-2/20/$31.00 ©2021 IEEE | DOI: 10.1109/VTC2021-Spring51267.2021.9448683

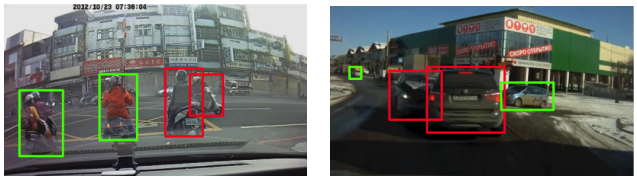
***Abstract*—A camera-based perception of dangerous road situations such as traffic accidents is a significant task in modern autonomous driving and ADAS. The previous approaches have scrutinized a spatio-temporal characteristics of the traffic accident in a sequence of images. However, we figured out the limit of past works that the aforementioned spatio-temporal pattern is only considered in 2D manner, which loses a contextual knowledge of the road situation in 3D space where the accident actually happens. In this study, we propose a novel approach to learn a spatio-temporal pattern of traffic accidents in a sequence of traffic scene images. First, we designed a spatial feature extractor that illustrates the distance among traffic objects in a 3D manner, which contextually describes the road situation better by considering traffic objects' location with their depth information. Second, we proposed an accident detection model and examined the model identified traffic accidents with 0.8560 accuracy and a 0.9080 F1 score. Lastly, we suggested an accident anticipation model, and it outperformed the previously-proposed benchmark anticipation model in a challenging task. We expect further improvement of our approach can contribute to the safe vehicular technology for autonomous driving and ADAS development.**

***Index Terms*—Traffic Accident Recognition, Object Detection, Monocular Depth Estimation, Recurrent Neural Network**

## I. Introduction

Autonomous driving and the Advanced Driving Assistance System (ADAS) started to provide a more convenient and safe mobility experience with vehicular technologies' progress. One of the key challenges in autonomous driving and ADAS development is ensuring safety by recognizing dangerous road circumstances such as traffic accidents. The late recognition might cause the vehicle to be dragged into a dangerous situation and create more dangerous road situations for other vehicles; thus, the academia and industry have endeavored to build a perception system that correctly perceives natural driving scenes and anticipate traffic accidents [1]. Prior researches have utilized various sensors such as camera, depth camera, or LiDAR for traffic accident recognition. Among the aforementioned sensors, past works actively focused on camera as it is cheaper than others but provide a wide range of meaningful information regarding the road situation. With computer vision technology development, past works have focused on traffic accident detection leveraging cameras equipped on a car.

The camera-based traffic accident detection researches discovered spatio-temporal characteristics of the accident from images taken by the camera mounted on a vehicle. First, prior works focused on a spatial pattern that traffic objects (cars, motorbikes) in the image are closely adjoined or overlapped at the accident. Previous studies analyzed traffic accidents occur when several traffic objects collide, or a particular traffic object gets too close to the camera-mounted vehicle; therefore, the distance between traffic objects in the image goes such short in case of an accident. Second, prior works also discovered a temporal pattern of the traffic accident that the distance among traffic objects excessively decreases along the images' sequence. It implies a decrease in the distance among objects does not solely correlate to the accident, but a 'sudden' decrease in the distance much correlates to the accident. Past works employed deep neural networks to recognize this spatial-temporal pattern in images and achieved significant accident detection performance [1]–[3].



(a) Detected objects with red bounding boxes seems to collide, but two objects are located in a different depth at 3D space

(b) Detected objects with red bounding boxes seems to collide, and it has become an accident as two objects are located in a similar depth

Fig. 1: Qualitative examples of showing the limit of prior works on traffic accident detection

Along with the literature review, we figured out a limit of prior works that they considered spatio-temporal patterns only in a 2D space. Previous works highlighted the spatial patterns by considering the distance between 2D bounding boxes, but we discovered this 2D distance does not fully illustrate an accident, which actually happens in a 3D space. Considering Fig. 1, green bounding boxes are detected traffic objects with a particular distance from other objects while red bounding boxes represent traffic objects that seem to collide with other objects. As red bounding boxes at both Fig 1 (a) and (b)

978-1-7281-8964-2/21/$31.00 ©2021 IEEE

are overlapped, previously proposed accident detection models might result in both cases as an accident. However, if we have a closer look, humans can notify the upper case is not an accident as two objects exist in different areas in 3D space, different depths from the point of the camera's view. We can also understand that the lower case is an accident as two objects collided in the same areas in 3D space. Throughout the analysis, we scrutinized the previously-proposed detection model lacks a contextual consideration of depth information compared to humans' understanding; therefore, we resulted in a spatial representation of traffic images can be improved if we consider each object's depth information.

This study proposes a novel accident recognition model by learning a visual pattern of traffic accidents. Compared to previous studies, our approach utilized depth information in the detection model to illustrate a spatial characteristics of traffic images more contextually. We designed a novel feature extraction module that describes the distance among traffic objects in 3D space. While prior works have focused on the distance between bounding boxes of each object, we added a depth estimation module to produce a depth map from a single image. By combining detected objects' bounding boxes and the estimated depth information, the feature extractor model produces a spatial feature map that includes a traffic accident's visual pattern contextually. Furthermore, we established an accident detection model that takes a sequence of feature maps as input and performs a binary classification between accident and non-accident cases. We employed the convolutional LSTM network [4] as a binary classifier to describe a temporal pattern of traffic accident images; thus, our approach successfully let the neural network to learn a spatio-temporal visual pattern of traffic accidents. Lastly, we examined the possibility of accident anticipation by extracting the accident probability from the trained detection model.

Throughout the study, our contributions are as follows:

- We proposed a novel feature extractor that takes a single image as an input and provides a spatial feature map that contextually illustrates the road situation. We combined bounding boxes of traffic objects generated by the object detector and depth map established by the monocular depth estimator. To the best of our knowledge, the proposed feature extractor is the first attempt to describe the distance among traffic objects with depth information, which is similar to the humans' perception.
- We established an accident detection model leveraging a convolutional LSTM network to fully illustrate spatio-temporal characteristics of traffic accidents. We showed the proposed accident detection model achieved a concrete detection performance at the dataset containing driving scenes in a real-world.
- We showed the accident detection model's extracted accident probability can be a useful measurement to anticipate traffic accidents. Leveraging the aforementioned accident probability, we implemented the anticipation model, and it anticipated the traffic accident before 1.208 seconds before the accident occurred.

## II. Literature Review

In this section, we briefly reviewed the related literature on the objection detection (II-A) and the monocular depth estimation (II-B), which are employed in our proposed accident detection and anticipation model.

### A. Object Detection

Object detection is an important computer vision task to identify instances of visual objects of particular classes. The topic itself has been studied since two decades ago, but deep neural networks' recent development excessively accelerated the object detection performance [5]. The object detection models can be categorized into two streams: A two-stage detector and the one-stage detector. The two-stage detectors aim to employ a 'coarse-to-fine' approach by generating a visual representation at the first stage and precisely identify the location and the class of objects at the second stage. The different versions of R-CNN [6] are representative two-stage object detectors that are accurate but comparatively slow. On the other hand, one-stage detectors pursued to completely produce the location and the class of objects in a single step. A series of YOLO [7] models are representative one-stage object detectors that are less accurate than two-stage detectors but accompany a faster inference speed.

### B. Monocular Depth Estimation

Monocular depth estimation is the task of estimating depth from a single image without any information. Given a single image as an input, the model predicts a depth map, although the scene image has various texture and structural variations [8]. Prior studies presented numerous approaches to resolve the monocular depth estimation problem. Early works tried to combine the information from multiple scales that aggregate the feature maps for better pixel-level depth prediction [9], [10]. Furthermore, numerous researches employed Conditional Random Fields (CRF) to achieve a more precise depth estimation result [11], [12]. However, prior studies mentioned above had a limit that the trained monocular depth estimation models were not robust to the unseen images. To hedge this limit, Godard *et al.* [13] proposed a depth estimation model in a self-supervised manner by exploiting epipolar geometry constraints, and achieved a state-of-the-art depth estimation performance at the KITTI benchmark dataset [14].

## III. Proposed Methodology

In this section, we elaborate on the proposed accident recognition methods. The overall architecture of our methodology is illustrated in Fig. 2. In §III-A, we described a spatial feature extractor that takes a raw image as an input and produces a feature map containing contextual information regarding the traffic accident. In §III-B, we explained the accident detection model. Given a sequence of feature maps, the model analyzes temporal characteristics of spatial feature maps and performs a binary classification between accident and non-accident. Finally, §III-C illustrates that how we leveraged the aforementioned accident detection model to anticipate the accident before its occurrence.

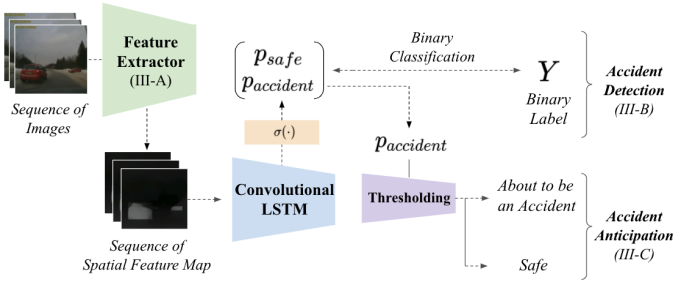



Fig. 2: Overall Architecture of the proposed accident detection and anticipation model

### A. Spatial Feature Extractor

The spatial feature extractor is a module providing a feature map from the raw image. We established the spatial feature extractor to describe an essential cue of accident detection in a more contextual manner: the distance between traffic objects in 3D space. If the distance among objects in 3D space is small, we expected the model could infer the scene as an accident. On the other hand, the model could infer the scene as a safe circumstance if the distance among objects in 3D space is large. Following this analogy, we provide a single raw image into an object detector and monocular depth estimator, which detect traffic objects' location and calculate the scene's depth, respectively. Then, we combined the aforementioned submodules' results to generate a feature map containing objects' location with their depth. We employed YOLOv3 [15] for the object detector and utilized model weights pretrained with the KITTI dataset [14]. Given a single image, the object detector produces bounding boxes of detected traffic objects such as cars or motorbikes. Furthermore, we employed a monocular depth estimation model proposed in [13], as the model achieved a state-of-the-art depth estimation performance on the KITTI dataset. The monocular depth estimator takes a single image and results in a pixel-wise depth of the scene. Finally, we combined the result of the object detector and monocular depth estimator. We analyzed depth information out of detected bounding boxes does not contain essential clues of the accident; thus, we substituted depth values out of detected bounding boxes as zero. On the other hand, we let the depth values inside of the bounding boxes as they were. As a result, we could extract a spatial feature map illustrating the distance among detected traffic objects in a 3D manner.

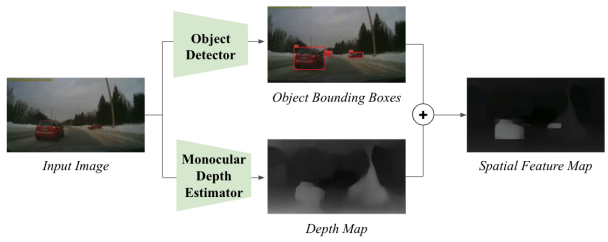



Fig. 3: The architecture of the feature extractor

### B. Traffic Accident Detection

Following prior researches, we scrutinized that traffic scene image's temporal pattern contributed to the precise accident detection performance. We adapted the importance of temporal dynamics of traffic scenes; therefore, we employed recurrent neural networks as an accident detection model by providing a sequence of spatial feature maps extracted from the series of raw images. Especially, we utilized convolutional LSTM networks because the model sustains a spatial pattern at the sequence of input features while it effectively learns the temporal dynamics of the sequence [4]. We analyzed the convolutional LSTM networks can effectively learn the accident's spatio-temporal dynamics when we provide a sequence of spatial features. We applied the sigmoid activation function at the result of convolutional LSTM networks, and let the model perform a binary classification to identify whether the input sequence of feature maps is accident or non-accident. Ground-truth labels of each input sequence of feature maps are one-hot encoded, and the model optimizes its parameters to minimize the classification loss. Throughout the experiment, we optimized the size of the input sequence and evaluated the accident detection performance with quantitative evaluation metrics, and the detailed results are elaborated in Section IV.

### C. Traffic Accident Anticipation

We discovered the accident detection model could be utilized to anticipate the accident before it occurs. A key takeaway of the accident anticipation is extracting the accident probability from the result of the accident detection model, which is convolutional LSTM networks. As the accident detection model solves a binary classification between accident and non-accident, the model provides a two-dimensional vector that contains a probability of an accident and non-accident respective to each dimension. Following the one-hot encoding rule, which implies an accident and non-accident, we can extract the accident probability from the aforementioned two-dimensional vector. We leveraged this accident probability to anticipate the occurrence of an accident. If the accident probability is high at a particular sequence of feature maps, we interpret the input sequence implies the accident can happen very soon. On the other hand, if the accident probability is low, we inferred the input sequence implies the input sequence describes a safe road situation. Throughout the analysis, we set the accident anticipation model to estimate a sequence of spatial features as an accident when the accident probability goes larger than the preset threshold. As the accident probability lies between zero and one due to the activation function, we empirically set the particular threshold as 0.8.

## IV. Experiment

### A. Experiment Setup

*1) Dataset:* Throughout the experiment, we employed the dataset published in [2]. The dataset contains video clips taken from the camera mounted on a car, and the dataset includes both accident and non-accident video clips. As the accident detection model takes a sequence of images, we applied a

preprocessing as described in Fig. 4. We set a particular size of the window and slid along with the timestamps to establish a sequence of images.

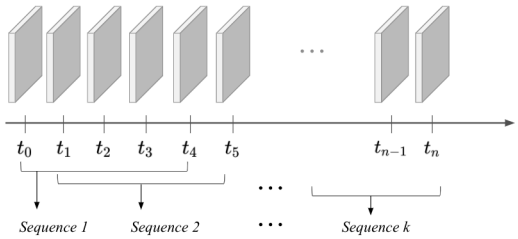



Fig. 4: Visualized illustration of the data preprocessing. We set a length of the window, and slid along with the video.

*2) Ground Truth Confirmation:* As the accident detection model necessitates a binary label of accident and non-accident, we manually checked the timestamp where the accident happened following the previously-proposed practices in [1], [3]. Thus, we first manually annotated the accident timestamp at each video clip. If a sequence of images includes the image after the accident timestamp, we labeled the sequence as an accident. On the other hand, we labeled a sequence of images as non-accident if it only includes images recorded before the accident timestamp. As a result, we established a preprocessed dataset which consists of a pair of sequences of images and binary labels.

### B. Result on Accident Detection

First, we examined the performance of accident detection model by learning a spatio-temporal pattern of traffic images. We acquired 40 video clips consisting of randomly selected 20 samples of accident video and 20 samples from non-accident video. We established a pair of sequences and binary labels following the aforementioned preprocessing method. Note that we differentiated the sequence's length as 5, 10, 20 to figure out an optimal sequence length. We composed the training set maintains the ratio between accident and non-accident as 5:5 to stabilize the model training process. But we composed the ratio between accident and non-accident as 2:8 at the test set as we empirically assumed the accident is not as frequent as safe circumstances in reality. We trained the accident detection model with a categorical cross-entropy loss and optimized it with an Adam optimizer. The trained model is evaluated on the test along with four evaluation metrics: Accuracy, Precision, Recall, and F1 Score. The experiment result is illustrated in Table I. Note that we could not compare our detection model with prior studies as our model is the only study that cast accident recognition task as a binary classification problem.

Considering the result shown in Table I, we evaluated the proposed accident detection model precisely learned a spatio-temporal pattern of traffic accidents. Especially, we figured out the sequence with the length of 5 achieved the best accident detection performance. We analyzed the effectiveness of the shortest sequence length derives from the accident's characteristics; the traffic accident occurs all of a sudden consuming short seconds. When we employ a recurrent form of neural networks, it is essential to design an input sequence to include many portions of accident-correlated information to highlight the accident's pattern to the model. If the input sequence's length is too large, the portion of accident-correlated feature maps at the total sequence goes small. We analyzed this small portion of accident-correlated images diluted a traffic accident's pattern; thus, the detection model could not successfully learn the pattern. On the other hand, the portion of accident-correlated images in a sequence goes large when the sequence's size is small; thus, the accident's pattern is comparatively highlighted to the model. Following the analogy on the sequence length, we scrutinized the proposed accident detection model accomplished a significant detection performance with a small length of input sequences.

TABLE I: Accident detection performance. The bold numbers imply the best detection performance.

| Window Size | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| **5** | **0.8560** | **0.9706** | **0.8529** | **0.9080** |
| 10 | 0.8110 | 0.7449 | 0.6016 | 0.8498 |
| 20 | 0.6659 | 0.5279 | 0.4860 | 0.6835 |
| Average | 0.7776 | 0.7268 | 0.6468 | 0.8137 |

### C. Result on Accident Anticipation

We examined the accident anticipation model's performance by comparing our model with a benchmark model: Dynamic-Spatial-Attention (DSA) Recurrent Neural Network [2]. We employed the DSA as a benchmark because both our approach and the DSA take a sequence of feature vectors as input and result in the accident probability. Second, the DSA model is known to accomplish stable accident anticipation performances in various datasets. We implemented our approach by freezing the trained accident detection model's parameters and extracting the accident probability from the model's produced two-dimensional feature vector. We experimentally evaluated our approach's effectiveness with the following metrics: Time-to-Accident (TTA) and computation overhead.

*1) Comparative Analysis on Time-to-Accident:* The TTA evaluates the accident anticipation performance by measuring the time between the first accident-detected timestamp and the actual accident. The larger TTA implies the model better contributes to the safety by predicting the accident before its occurrence earlier. Following our study's motivation, we selected a challenging accident anticipation task where traffic objects are located near each other in the image, but the actual locations in a 3D world are hard to be recognized. We sampled an accident video, processed them into the sequence with a length of 5, which achieved the best accident detection performance. The input sequences are provided to each model, and we identified an accident if the estimated accident probability exceeds a preset threshold of 0.8. We measured the TTA from each model and examined our approach outperformed the benchmark model as illustrated in Fig. 5.

Around timestamp 80 in Fig. 5, our model confidentially predicted an accident while the benchmark model provided

lower accident probability. We analyzed the better anticipation performance of our approach comes from the consideration of the depth information. We interpreted the benchmark model confused to understand the circumstance clearly and suffered to infer an accident as the given information is limited in a 2D space. On the other hand, our approach could precisely understand the road situation in a 3D space leveraging the depth information; thus, it certainly anticipated an accident. In a nutshell, we figured out the consideration of depth information contributes to the precise understanding of the road situation for the accident anticipation. Nevertheless, we acknowledge further rooms to improve our approach, necessitating external validation on other datasets and comparative experiments with the previously-proposed approaches.

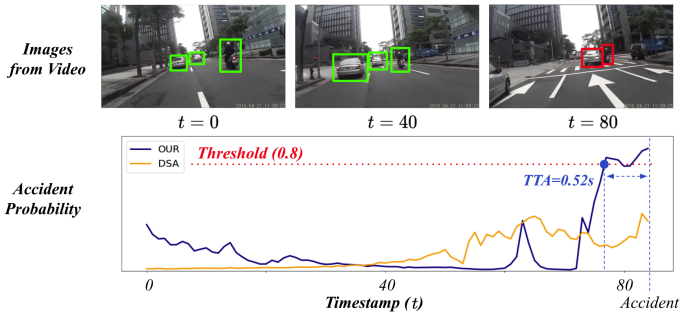



Fig. 5: Comparative experiment results from the challenging accident anticipation task. Our model anticipated the accident while the benchmark model failed to recognize it.

*2) Comparative Analysis on Computation Overhead:* As the computation overhead is essential to the accident anticipation model's deployment, we evaluated the computation overhead by measuring the inference time consumed for a single video. We employed a single video and compared the inference time of our approach and benchmark model in Table II. We figured out the proposed approach accompanies a larger computation overhead compared to the benchmark model. We scrutinized our model has larger computational complexity due to the spatial feature extractor. While the benchmark model extracts information only from the 2D space, our model additionally extracts the depth information from the 3D space with a monocular depth estimator, which adds to more computational complexity. Therefore, we figured out future studies shall reduce the proposed model's computation overhead for the real world's fast accident anticipation.

TABLE II: Inference time of our model and the benchmark model. The video has 100 frames during 4 seconds.

| Model | Inference Time |
|---|---|
| OUR | 11.99s |
| DSA [2] | 3.43s |

## V. Conclusion

Our study proposed an accident detection model and the anticipation model by learning spatio-temporal characteristics of traffic accidents. First, we proposed a novel feature extractor consisting of an object detector and monocular depth estimator to illustrate the accident's spatial characteristics. Second, we modeled a temporal pattern of the aforementioned spatial feature maps by employing convolutional LSTM networks as an accident detection model. The experiment on the benchmark dataset showed our approach achieved a precise detection performance. Lastly, we presented the accident anticipation model by extracting an accident probability from the detection model. The proposed model outperformed the benchmark model in sample videos but accompanied a larger computation overhead. Throughout the analogy, we acknowledged that the proposed approach necessitates further strict validations on the performance and improvement of computational complexity. We expect our approach can contribute to traffic accident detection and anticipation for a safe driving environment shortly by resolving the aforementioned rooms for improvements.

## References


[1] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised traffic accident detection in first-person videos," *arXiv preprint arXiv:1903.00618*, 2019.

[2] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 136–153.

[3] T. Suzuki, H. Kataoka, Y. Aoki, and Y. Satoh, "Anticipating traffic accidents with adaptive loss and large-scale incident db," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3521–3529.

[4] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, pp. 802–810, 2015.

[5] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *arXiv preprint arXiv:1905.05055*, 2019.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[8] A. Bhoi, "Monocular depth estimation: A survey," *arXiv preprint arXiv:1901.09402*, 2019.

[9] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, pp. 2366–2374, 2014.

[10] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3917–3925.

[11] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5354–5362.

[12] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011.

[13] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 3828–3838.

[14] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[15] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.