Potential Risk Estimation with Single Monocular Camera

Kota Shimomura[1], Hiroki Adachi[1], Tsubasa Hirakawa[1],
Takayoshi Yamashita[1], Hironobu Fujiyoshi[1], Masamitsu Tsuchiya[2], Yuji Yasui[2]
[1]Chubu university, [2]Honda R&D
{shimo,ha618,hirakawa}@mprg.cs.chubu.ac.jp, {takayoshi,fujiyoshi}@isc.chubu.ac.jp,
{masamitsu_tsuchiya,yuji_yasui}@jp.honda

Abstract

Object detection, segmentation, position estimation, and identification of white lines on roads are essential components of computer vision for recognizing surrounding vehicles and pedestrians. These tasks are focused on differentiating objects and driving scenes with the help of cameras and LiDAR sensors. However, to enhance the capability of autonomous driving, it is essential to address the possibility of future risks and object variations, which have not been adequately explored. Specifically, identifying the zones where pedestrians and vehicles may suddenly appear is of paramount importance for ensuring driving safety and preventing traffic accidents. In this paper, we propose a novel task that aims to estimate the potential risk regions that can cause traffic accidents. Our focus is on assessing the risk regions from images taken by an in-vehicle camera installed at the front of the vehicle. We define a risk region as an area where pedestrians or vehicles may appear, and we annotate the Cityscapes dataset with risk region annotations. Additionally, we propose an end-to-end network and evaluation metrics for estimating the baseline risk regions. Our results demonstrate that our approach performs exceptionally well in estimating potential risk regions in various scenarios. This research is expected to facilitate the establishment of safety tasks in the driving environment and enable autonomous driving systems to identify potential risk regions and drive safely.

1. Introduction

Autonomous driving-environment recognition comprises multiple components, among which the recognition of objects and scenes around the vehicle is paramount importance. Thanks to advancements in hardware such as cameras and LiDAR, computer vision has made significant progress, leading to the development of highly precise algorithms for various perception tasks on images and videos.

Object detection of pedestrians and vehicles from images taken from an onboard camera mounted at the front of a vehicle [1, 5, 18, 25] can accurately and quickly infer objects of various scales in a scene. Meanwhile, pixel-level semantic segmentation of a scene [4, 19, 27] can also accurately and quickly infer objects of various scales while remaining robust across different domains. Furthermore, it is capable of achieving excellent accuracy in bird's-eye view segmentation and 3D object detection using images from multiple cameras, rather than just a monocular camera [19, 27].

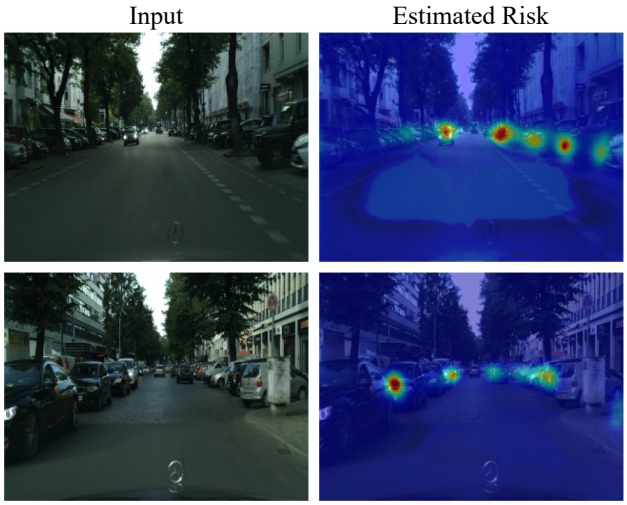

Figure 1. On left side is input image, and on right side is output result of proposed method. Red areas represent higher risk, while blue areas represent lower risk. Notably, proposed method accurately estimates risk regions around stationary vehicles.

The availability of large datasets annotated with high accuracy has been a key factor supporting the tasks mentioned above. Various real datasets that compile a wealth of information from images to point clouds are now available [2, 10, 21, 26]. However, due to the high cost associated with performing accurate annotations on a huge dataset, semi-supervised learning [20] and unsupervised learning [6, 8] have been proposed. Additionally, datasets created using

a simulator environment [12, 16, 17], provide a variety of road environments while significantly reducing annotation costs. These contributions have played a significant role in advancing environmental awareness in autonomous driving.

There is still considerable room for improvement in recognizing the interrelationships between future traffic accidents, object variations, and their interactions in the realm of autonomous driving. Specifically, preventing and mitigating traffic accidents caused by pedestrians or vehicles that suddenly emerge from the shadows of side roads or stationary objects would enable drivers to transition from open roads, such as highways, to more complex driving environments like urban regions, thereby providing a safer driving environment. This is because most drivers can empirically identify potential risk regions and consciously or unconsciously adjust their vehicle speed or change course to ensure safety. In high-speed driving environments like freeways, the possibility of pedestrians or other vehicles suddenly appearing is extremely low, but there are many vehicles in close proximity that are traveling at high speeds. Therefore, the primary focus of drivers is on perceiving the surrounding vehicles. Conversely, in urban regions, where speeds are low and there are many intersections and obstacles, drivers pay more attention to gathering information about their surroundings by slowing down their vehicle speed. While safety is maintained through the mutual perception of many drivers and pedestrians, there is a limit to the amount of information that a single driver can acquire, and in complex scenes, potential risks can be overlooked, leading to accidents.

To provide a novel technology for autonomous driving that can perceive and evade potential risks in the surrounding environment, we annotated the Cityscapes dataset [10] to include risk regions. Our ultimate goal is to mitigate the number of traffic accidents and the damage they cause by alerting drivers to potential risk regions, regulating vehicle speed, and altering course accordingly. Currently, many algorithms, such as those for object detection and segmentation, can infer only the explicit states of pedestrians, vehicles, and other objects that appear in a scene. This is due to the fact that annotations for states in a scene can be established with uniform regulations. However, numerous risk regions tend to appear across object boundaries, making them incompatible with algorithms that require precise annotation, such as for object detection and segmentation.

In this paper, we propose a framework that facilitates end-to-end estimation of both risk values and risk regions Figure 1 from input images, alongside a proposed task for estimating potential risk regions through the provision of annotation data. Our experiments are carried out on various risky regions present in a scene, marked by single-point annotation. To summarize, our contributions are as follows:

- We formulate the problem as estimating the probability distributions of risk regions and propose a framework that allows for dense area estimation.
- We create an additional annotation indicating potential risk regions for the Cityscapes dataset. Our annotation represents a risk region as a single-point annotation.
- We propose two novel evaluation metrics that assess the position and priority of risk regions.

2. Related Work

In this section, we present a comprehensive review of the literature on risk analysis during vehicle operation, scene analysis during traffic accidents, and vehicle perimeter recognition using object detection and semantic segmentation.

Risk Analysis. In various regions, analyses have been conducted on the causes of traffic accidents involving pedestrians and vehicles, as well as the environmental information of accident scenes. For instance, Zhen *et al.* [9] focused on the severity of traffic accidents on highways in Southern California, while Zhiyuan *et al.* [22] concentrated on the severity of traffic accidents in North Carolina over a period of approximately five years. The analysis was centered on the severity of traffic accidents, which was found to be influenced by various factors, including the location of occurrence, weather conditions, time of day, pedestrian age and vehicle type, and traffic volume. Notably, weather conditions and the location of traffic accidents were found to be particularly influential, underscoring the importance of environmental recognition around vehicles in the computer vision field. Additionally, a hybrid network that leveraged a wide range of domain data from large-scale traffic data, accident information, and weather information was developed. Using a learned model, sensitivity analysis was performed to investigate the causes of traffic accidents with high contribution rates, and the findings were consistent with the analysis of Zhen *et al.* [9]. However, since these studies utilize environmental information that is already manifest, they cannot be directly applied to the task of estimating potential risk regions, which is the focus of our study.

Object Detection. Recent studies on object detection [1, 3] have achieved high accuracy and speed, while also demonstrating robustness against occlusion and weather changes, which is crucial for autonomous driving technology. However, safety support systems for vehicle operations that rely solely on object detection are limited to detecting only the objects that appear in camera footage, which differs from the primary goal of this research, which is to prevent accidents by alerting drivers to estimated risk regions. Furthermore, it should be noted that object detection necessitates precise annotation of bounding boxes, which can be a costly

(a) Input (b) Single-Point Annotation (c) Ground Truth

Figure 2. Examples of labels used in our framework. (a) shows input image. (b) shows single-point annotation at pixel level with respect to center of risk region. (c) is ground truth obtained by enlarging annotation based on single point and applying gaussian filter.

and time-consuming process, especially when it comes to annotating risk regions.

Semantic Segmentation. Semantic segmentation faces a fundamental problem analogous to object detection, in that it monitors object regions at the pixel level. However, risk regions that span across different objects lack well-defined boundaries, which poses a challenge for semantic segmentation in risk region tasks. Kozuka *et al.* [14] propose a method for determining risk and safe regions by leveraging prior knowledge of semantic segmentation. However, risk regions such as shaded regions around stationary vehicles and side roads are classified into a single vehicle or road class in segmentation. Moreover, risk regions that straddle different objects are treated as safe regions, potentially reducing the amount of information about sparse risk regions.

3. Annotation for Risk Region Estimation

The availability of openly accessible datasets with annotations for risk regions is crucial for achieving the task of estimating potential risk regions. However, due to the abstract and varying nature of risk regions, accurately annotating a large dataset can be very costly, and it is difficult to provide a uniform and precise annotation [14]. Building on previous research, we focused on estimating risk regions from images captured by in-vehicle cameras mounted in front of a vehicle. We defined a risk region as a region where pedestrians and vehicles may suddenly appear and annotated the center of the risk region with a single point in the Cityscapes dataset [10]. An example is shown in Figure 2 (b). Note that although the annotation size is large for visibility, the annotations are made for a single pixel.

3.1. Annotation Cost

As listed in Table 1, we allocated annotators with prior driving experience in each city to annotate the train, val, and test sets in the Cityscapes dataset. We ensured that each image had at least one risk region selected for annotation. In contrast to [14], we also annotated one point per image in the test set and did not create pairwise labels using the predicted segmentation masks and prior knowledge. The annotation process for the entire dataset took 3385 minutes or approximately 41 seconds per image.

Data	Images	Workers
Train	2975	5
Val	500	3
Test	1525	4

Table 1. Annotation for Cityscapes. Images refers to number of images, while Workers indicates number of annotators.

4. Risk Estimation

Our objective is to estimate multiple potential risk regions from input images, with the output being a pixel-wise relative risk value z, as illustrated in Figure 1. However, since most potentially risky regions span across object boundaries and do not appear on a uniform object, pixel-level identification using supervised segmentation and object detection with bounding boxes is challenging due to the high cost of annotation. To overcome this challenge, we propose a framework inspired by the saliency prediction task [11, 13, 28] that can densely estimate multiple risk regions from input images in an end-to-end fashion.

4.1. Risk Estimation Network

To accomplish the task of estimating potential risk regions, we adopt a network architecture [7] with an encoder-decoder structure, which is inspired by the most relevant work on risk region estimation [15]. The entire network is depicted in Figure 3. The Inception module's [23] concept is to serve as a multi-scale feature extractor with various receptive fields. Therefore, the module is comprised of a combination of multiple convolutional layers with different kernel sizes, enabling it to obtain various receptive fields. The network comprises an encoder with a series of modules

and downsampling and a decoder with comparable modules, upsampling, and a residual connection for integrating high-resolution features.

4.2. Loss Function

Exponentially Weighted MSE Loss. Compared with the general saliency prediction task, the risk regions present in vehicle camera images possess a specific characteristic where the ground truth is distributed in the center of the image, specific to the vehicle driving scene. Furthermore, the ground truth is sparsely distributed in proportion to the image region, with many zero regions. To address these challenges, we introduce the exponentially weighted mean squared error (MSE) loss [24]. This loss function is exponentially weighted on the basis of the magnitude of the predictions, which is effective in addressing the tendency to predict 0 as a result of sparse ground truths compared with conventional MSE. The exponentially weighted MSE loss is defined as

$$\mathcal{L}_{Ew-MSE} = \frac{1}{N}\sum_{i=1} exp(-z_i)(I_i - z_i)^2, \qquad (1)$$

where $I \in \mathbb{R}^{(1\times h\times w)}$ is the ground truth, $z \in \mathbb{R}^{(1\times h\times w)}$ is the model output, and N is the number of pixels.

Total Variation Distance. In the risk region estimation, using loss functions that are designed for pixel-by-pixel regression and classification makes it difficult to utilize global information around a risk region, as pixel-by-pixel prediction does not account for the relationships between pixels. This limitation is especially critical for risk regions that span the boundaries of different objects. However, by treating a risk region as a probability distribution of risk regions in an image, it is possible to introduce a loss function that measures the distance between probability distributions, thus addressing the aforementioned problem. This is the most innovative aspect of our method, which differs significantly from the approach in [14], which classifies risk and safe regions. To transform the predicted risk regions and ground truth into a probability distribution, we employ linear normalization, which preserves the initial proportions, unlike softmax normalization, which de-emphasizes the maximum value of elements for arrays within $[0, 1]$. As a result, it is possible to effectively monitor the error between the estimated most at-risk region and the corresponding ground truth. The use of linear normalization for probability distribution distance loss has been shown to be superior through experimental evaluation [28]. The total variation is defined as

$$\mathcal{L}_{Tv-Dist} = \sum_i |f(z_i) - f(I_i)|. \qquad (2)$$

The definition of z_i and I_i by linear normalization is as follows.

$$f(z_i) = \frac{x_i^z}{\sum_{i=1}^N x_i^z}, \qquad f(I_i) = \frac{x_i^I}{\sum_{i=1}^N x_i^I}, \qquad (3)$$

where $x := \{x_i\}_{i=1}^N$ is the set of unnormalized values for either the estimated risk probability x^z or ground truth x^I.

Relative Risk Loss. The objective of our training is to estimate multiple risk regions in a scene while ensuring that there are enough safe regions that contradict the risk regions. To this end, we propose the use of relative risk loss, which comprises exponentially weighted MSE loss and total variation distance. The risk region estimation network is optimized using the proposed relative risk loss. The relative risk loss is defined as

$$\mathcal{L}_{risk} = \lambda\mathcal{L}_{Ew-MSE} + \mathcal{L}_{Tv-Dist}, \qquad (4)$$

where λ is a temperature parameter that limits the contribution of exponentially weighted MSE to the overall loss function as a hyperparameter.

4.3. Label Preprocessing

The information provided by single-point annotation on the pixel level for risk regions is limited compared with the information available from the entire image region. Risk regions typically straddle the boundaries of different objects. If we also consider the distance information in an image, we can see that risk areas that are farther apart are relatively smaller, and those that are closer together are relatively larger. Therefore, incorporating depth information to expand risk regions can enhance the amount of information in the ground truth. To achieve this, we employ the distance bias in outdoor images discussed in Section 3 of Chen *et al.* [7]. According to their study, classifying lower points in an image as closer to the depth yields a recall of 85.8%, while classifying points closer to the center of an image as having more depth yields a recall of 71.4%. In Cityscapes, the high points in an image usually represent empty regions, while the low points indicate the front of the vehicle. We use this a priori knowledge to obtain relative distance information by taking the absolute value of the distance on the x-axis and the value on the y-axis with respect to the center and highest pixel in an image and then expanding the annotation at the pixel level. As the risk level is not the same within an enlarged risk region, we use a gaussian filter to represent the continuously changing risk region.

5. Experiments

In this section, we verify the ability of our new framework to estimate potential risk regions for unknown images. We also compare our framework with the most relevant method, [15]. To conduct a unified comparison experiment, we employ the encoder-decoder structure [7] as the network

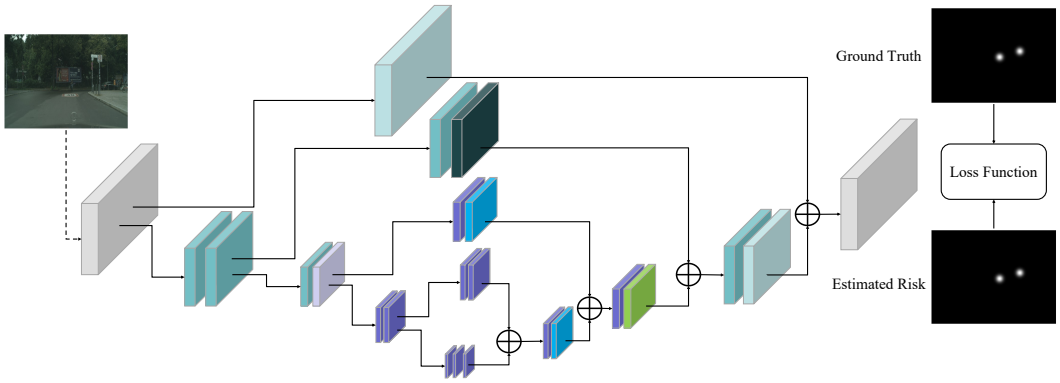

Figure 3. Overview of our network. Blocks with same color are composed of common Inception module. Symbols indicate element-wise addition. In case of pre-training for semantic segmentation tasks, final layer head is replaced.

for risk region estimation. The network replaces the final layer with a head for the semantic segmentation task and is pre-trained on Cityscapes train set. The head is replaced for region-at-risk estimation. Except for the head in the final layer, the network utilizes the pre-trained weights and fine-tunes the weights for the entire network during training.

Our Baseline. To expand the annotation of a single image, we utilized depth information to enlarge multiple single-point annotations into a continuously changing risk region. It was then transformed using a gaussian filter to create the ground truth. The network was optimized using the important equation Eq. (4) from Sec. 4.2.

Comparative Method. We adopted the method proposed in [15] as a comparative approach, which utilizes a differential pair consisting of pixels in the risk region indicated by single-point annotation and pixels randomly selected from the image. Additionally, an equal pair was created by utilizing two pixels randomly selected from the safe region generated using the segmentation results and prior information. When generating safe regions, we designated the semantic segmentation class (e.g., person, rider, car, truck, bus, train, motorcycle, and bicycle) to risk region classes, with 60 pair labels produced per image. The optimization of the network was carried out using the loss function [15].

5.1. Evaluation Metrics

We aimed to conduct a quantitative evaluation of the potential risk region estimation results of our framework. We evaluated our proposed method from two different perspectives: capability to estimate risk regions in a scene, and ability to assess the relative risk value. For this purpose, we utilized two commonly used evaluation metrics, namely the area under the curve (AUC) and the correlation coefficient (CC), which have been used in previous studies [15, 28].

AUC: The performance of binary classification for estimated risk regions was evaluated using AUC. A binary map was created using a positive set representing the ground truth risk regions and a negative set. By varying the threshold from 0 to 1, the estimated results can be converted into risk regions and backgrounds, and a receiver operating characteristic (ROC) curve can be generated. The AUC was calculated using the ROC curve.

CC: The linear correlation coefficient (CC) is a statistical metric that quantifies the linear correlation between two stochastic variables. To evaluate the performance, the predicted risk region z and ground truth I are considered as two random variables, and CC is defined as

$$CC = \frac{cov(z, I)}{\sigma(z) \times \sigma(I)}, \tag{5}$$

where $cov(-,-)$ is the covariance, and $\sigma(-)$ is the standard deviation.

5.2. Potential Risk Estimation

We list the result of quantitative evaluation of potential risk region estimation in Tables 2 and 3. It is evident that the proposed method achieved superior accuracy in both AUC and CC for all Cityscapes sets as compared with the comparative method. Notably, the improvement in the CC metric score that our method the adroitness to incorporate the prioritization of risk regions juxtaposed with the comparative technique. This can be attributed to the incorporation of our loss function, which addresses the ground truth of sparse risk regions and enhances the overall performance. The fine-tuning column depicts the outcomes of learning from scratch and utilizing the pre-trained weights from the segmentation task. The performance improved for both Cityscapes val and test sets by using the pre-trained model. These findings suggest that incorporating semantic segmentation knowledge as a pre-training task can be effective in our method.

Table 4 presents experimental results that compare the effect of λ in the loss function. As indicated in the table, the best performance was attained with $\lambda = 0.5$.

Method	Fine Tuning	AUC	CC
Point Supervision [15]		0.545	0.111
	✓	0.511	0.102
Ours		0.540	0.158
	✓	**0.562**	**0.218**

Table 2. Ablation analysis with Cityscapes val set.

Method	Fine Tuning	AUC	CC
Point Supervision [15]		0.544	0.105
	✓	0.512	0.109
Ours		0.542	0.183
	✓	**0.564**	**0.216**

Table 3. Ablation analysis with Cityscapes test set.

λ	0.1	0.3	0.5	0.7	0.9
AUC	0.517	0.520	**0.564**	0.549	0.554
CC	0.089	0.120	**0.216**	0.184	0.142

Table 4. Analysis of temperature parameter λ on Cityscapes test.

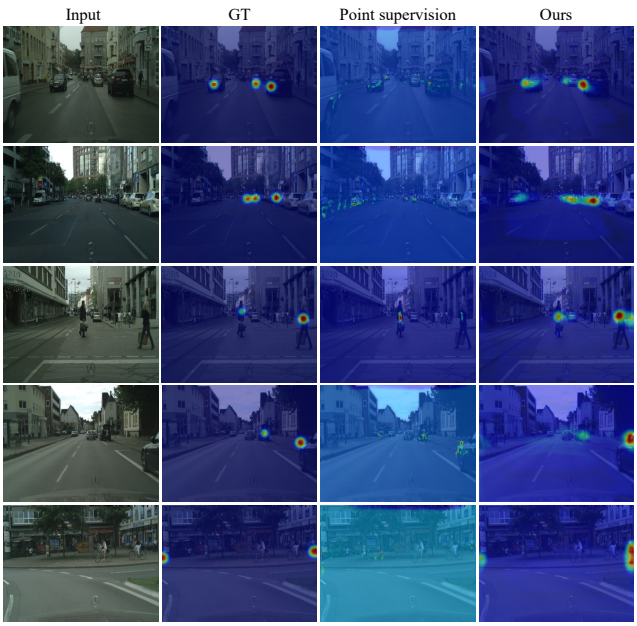

Figure 4. Qualitative comparison using our method and Comparative method [15]. Images are from Cityscapes test set.

5.3. Model Visualization

We evaluated the effectiveness of our proposed method by visualizing the results of potential risk region estimation separately for our method and a comparative method. The visualization results of the risk region estimation for Cityscapes test are shown in Figure 4. From first to fourth rows indicate the input images, the ground truth, the estimation results of [15], and those of our proposed method, respectively.

In the first line, our method estimated the risk region not only between the vehicles stopped on the right side of the road but also for oncoming vehicles. These results imply that the risk level around the near vehicles is higher than around the far vehicles. In the second line, our method enabled us to estimate a widespread risk region between several stopped vehicles on the right side. Furthermore, the degree of the risk region was higher for vehicles in the near distance than in the far ones. On the other hand, the comparative method was difficult to estimate the risk region for faraway and stopped vehicles.

Next, we discuss the third line in Figure 4 including pedestrians. In this case, our method estimated region around the vehicle but also the pedestrians as risk regions. Our method has a higher risk for pedestrians in the near distance, while the comparative method has a higher risk for pedestrians in the far distance in the center of the image. In the fourth line, where a vehicle is overtaking from behind, we can estimate that the vehicle about to overtake is in the risk region. In the fifth line, is an intersection scene. Our method estimated the risk region by considering the possibility of vehicles jumping out of the road leading to the outside of the camera's view. On the other hand, the comparative method tended to estimate objects (e.g., vehicles and pedestrians) in the image as the risk region, and failed to estimate the risk region considering the possibility of vehicles jumping out from outside FOV.

For these reasons, our method was capable of estimating risk regions that capture the context of the entire scene.

6. Conclusion and Discussion

Most recent research on environmental awareness has concentrated on object detection and segmentation to improve the capability of recognizing apparent objects. In contrast, this study focuses on estimating potential risk regions that are in conflict with revealed objects. A dataset with additional single-point annotation contributes to the establishment of a novel potential risk-region estimation task in computer vision. We take risk estimation as a probability distribution prediction task and use a linear normalization-based loss function. With the proposed loss function, our model outperforms the previous method.

Our baseline method has the potential to improve performance, we expect further improvement by using larger networks or time series data. Overall, we hope that our work serves as a preliminary step towards reducing the number of traffic accidents and the resulting damages caused by possible future changes in risks and objects, and that it will contribute to the development of a potential risk-region prediction task in computer vision.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection, 2020. 1, 2

[2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing. 2

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016. 1

[5] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022. 1

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 1

[7] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 3, 4

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, June 2021. 1

[9] Zhen Chen and Wei (David) Fan. A multinomial logit model of pedestrian-vehicle crash severity in north carolina. *International Journal of Transportation Science and Technology*, 8(1):43–52, 2019. 2

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2, 3

[11] Richard Droste, Jianbo Jiao, and J. Alison Noble. Unified Image and Video Saliency Modeling. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, 2020. 3

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[13] Samyak Jain, Pradeep Yarlagadda, Shreyank Jyoti, Shyamgopal Karthik, Ramanathan Subramanian, and Vineet Gandhi. Vinet: Pushing the limits of visual modality for audio-visual saliency prediction, 2021. 3

[14] Kazuki Kozuka and Juan Carlos Niebles. Risky region localization with point supervision. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017. 3, 4

[15] Kazuki Kozuka and Juan Carlos Niebles. Risky region localization with point supervision. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 246–253, 2017. 3, 4, 5, 6

[16] Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. SPIGAN: Privileged adversarial learning from simulation. In *International Conference on Learning Representations*, 2019. 2

[17] Moritz Menze, Christian Heipke, and Andreas Geiger. Joint 3d estimation of vehicles and scene flow. In *ISPRS Workshop on Image Sequence Analysis (ISA)*, 2015. 2

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 1

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 1

[20] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. In *arXiv:2005.04757*, 2020. 1

[21] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[22] Zhiyuan Sun, Yuxuan Xing, Xin Gu, and Yanyan Chen. Influence factors on injury severity of bicycle-motor vehicle crashes: A two-stage comparative analysis of urban and suburban areas in beijing. *Traffic Injury Prevention*, 23(2):118–124, 2022. 2

[23] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 3

[24] Igor Vozniak, Philipp Müller, Lorena Hell, Nils Lipp, Ahmed Abouelazm, and Christian Müller. Context-empowered vi-

sual attention prediction in pedestrian scenarios. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 950–960, 2023. 4

[25] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-YOLOv4: Scaling cross stage partial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13029–13038, June 2021. 1

[26] Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[27] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Neural Information Processing Systems (NeurIPS)*, 2021. 1

[28] Sheng Yang, Guosheng Lin, Qiuping Jiang, and Weisi Lin. A dilated inception network for visual saliency prediction. *IEEE Transactions on Multimedia*, 22(8):2163–2176, 2019. 3, 4, 5