Depth Estimation of Traffic Scenes from Image Sequence Using Deep Learning

Xiaoxu Liu and Wei Qi Yan

Auckland University of Technology, Auckland 1010 New Zealand

Abstract. Autonomous cars can accurately perceive the deployment of traffic scenes and the distance between visual objects in the scenarios through understanding the depth. Therefore, the depth estimation of scenes is a crucial step in the obstacle avoidance and pedestrian protection from autonomous vehicles. In this paper, a method for stereo depth estimation based on image sequences is introduced. In this project, we improve the performance of deep learning-based model by combining depth hints algorithm and MobileNetV2 encoder to enhance the loss function and increases computing speed. To the best of our knowledge, this is the first time MobileNetV2 is applied to depth estimation based on KITTI dataset.

Keywords: Deep learning, automatic car, scene depth understanding, depth estimation

1 Introduction

The process of cognizing and assuming environment based on spatial perception is known as scene understanding [1]. A scene, in the context of autonomous cars, is the environment in which the vehicle is presently operating and contains the location, drivers, event, and their interactions [2, 39, 40, 41, 42]. In order for autonomous vehicles to be driven safely and smoothly in complex urban traffic environments, the perception and understanding of depth in traffic scenes are of paramount importance [47, 48, 49, 50, 51]. Therefore, through robust depth estimation of traffic scene, the autonomous vehicles can become true.

Scene depth information plays an important role in advanced autonomous vehicles. The vehicle-related depth information can accurately perceive the operating environment of the vehicle and obtain the distance between the vehicle and pedestrians or others in traffic environment, so as to realize obstacle avoidance and pedestrian protection functions of autonomous vehicles. Compared with sensors, the driving recorder can obtain the color, texture and other information, the price is relatively low. Therefore, a number of scene understanding tasks are based on the images from driving recorder [43, 44, 45, 46].

The performance of depth estimate of traffic scene in autonomous automobile may be improved to further depth recently due to the advancement of deep learning [3]. Additionally, deep learning has the active benefit of transfer learning, which has benefited the training process of multiple traffic scenarios using a variety of pretrained networks and public datasets. Deep neural networks simulate high-level

abstraction from the visual data and encode the objects, scenes, and events in motion pictures using an efficient representation in order to comprehend them [4]. As a result, deep learning methods offer special benefits when it comes to detect the depth of a picture.

One of the benefits is the end-to-end nature of deep learning, which, on the theory of a particularly exact recognition of individual situations, produces faster global information processing than standard methods. The deep learning approach can successfully meet the accuracy and real-time requirements for autonomous vehicles that must comprehend the information in complex traffic environments [5, 35, 36, 37, 38].

However, for the performance of most depth estimation models based on the KITTI dataset [15, 16, 17, 18, 52], we found that the problem of detailed regions in the scene on the predicted depth map is still existing. One of the reasons is that incomplete feature extraction by using encoder [19, 20, 21, 22], another is that the network focus is on learning the depth to obtain the local minimum of the reprojection loss in the process of self-supervised learning, which cannot attain the global minimum [31].

Therefore, in this paper, we proffer MobileNetV2 structure as an encoder to transfer fine-grained details from high resolution to low resolution. At the same time, we employ the depth hints algorithm to compute an alternative depth value and incorporate it into the objective function to obtain a satisfactory result [31].

In this paper, literature review will be presented in Section 2, our method will be shown in Section 3, our conclusion will be drawn in Section 4.

2 Literature Review

We review the outstanding studies of deep learning and depth estimation in this paper. The characteristics of the end-to-end nature in deep learning [6], strong versatility [7], and active mobility [8] have already demonstrated powerful capabilities in traffic depth understanding. Moreover, the layer-by-layer process of deep learning enables the model to better express the information. Therefore, the method based on deep learning has become the standard solution for image depth estimation.

In the past years, there are already heaps of studies related to depth estimation [23,24,25,26] based on deep learning in indoor and outdoor scenes. Fully convolutional network is one of the most popular structures in deep learning. The improved fully convolutional network [9] was applied to depth estimation. Different from the previous pretrained network, the fixed fully connected layer is employed to obtain the image-to-image conversion. Iro et al. [9] directly removed the fully connected layer and replaced it with a network having a pretrained network structure to return the high-level features to the same size as the original image. The entire net was regarded as an encoder-decoder process. The advantage is to streamline the parameters to make better use of GPUs. Moreover, the improved network can directly process images with any sizes instead of the fixed size of the network input and output like a normal fully connected network.

By observing the experimental results, FCRN [9] using ResNet-50 performed well in overall depth prediction, but the expression of details is not perfect. Through comparisons, the method [10] can better reflect the detailed information. It is a combination of global and local strategies. This strategy takes use of coarse network to predict the overall trend, and harnesses the fine network to perform local tuning on the overall trend. The depth dimensions obtained by this method are all smaller than the original image. One pixel of the predicted small depth map can represent the overall depth of the current position information, make the RMSE (Root Mean Square Error) smaller, but it will also lose a lot of depth information. This is the reason why it is not as good as FCRN [9] in overall depth estimation, but it handles details and contours better.

Although multilayer neural networks had outstanding performance in depth estimation. However, in the training process of the supervised learning model, it is necessary to obtain in advance the reference standard of the depth value corresponding to a large number of input data as the training samples, so as to carry out the backpropagation of the neural network. However, in reality, it is very costly to obtain the depth information corresponding to the scene. Therefore, a number of studies circumvent the problem of obtaining depth information at a high cost through unsupervised learning methods [27, 28, 29, 30].

An unsupervised CNN for single view depth estimation was proposed [11]. The proposed model takes advantage of a structure similar to FCN without the participation of a fully connected layer. The volume of the model is smaller and computing speed is faster. At the same time, the participation of skip-connect ensures the relative integrity of the output feature details. Moreover, this model has the pre-trained network structure as the encoder part, which can achieve relatively good results in the case of insufficient training data.

Based on the model [11], the algorithm and structure were improved. Different from the FCNs, the disparity corresponding to the current feature size for the outermost 4 layers of the decoder part was estimated [12] that passed to the lower layer of the decoder after upsampling. This can ensure that each layer extracts disparity, which is equivalent to conduct a coarse-to-fine depth prediction. Since most models has taken use of bilinear differences, the range of the gradient always comes from the surrounding four coordinate points. The advantage of coarse-to-fine is that the prediction can make the gradient from a coordinate point rather from the current position.

From the perspective of learning methods, the vigorous development of deep learning was driven by large-scale annotated data, supervised learning promotes the development of deep models towards higher and higher performance. However, a large amount of labeled data often requires huge costs, more and more research work has begun to focus on how to improve the performance of the model without acquiring data labels. Hence, there are heaps of studies focusing on self-supervised learning and unsupervised learning for depth estimation.

Pertaining to stereo matching or binocular depth estimation, a device like LiDAR is extremely bulky and expensive, what it can collect is sparse depth information, what we need is a dense depth map; however, devices based on structured light can

often only perform depth information annotation in indoor scenes, it is difficult to achieve high annotation quality in outdoor traffic scenes. Therefore, self-supervised learning has received more and more attention in stereo matching.

According to the review of these literatures, in this paper, we introduce a depth estimation model which combines MobileNetV2 and depth hints to achieve high-resolution depth estimation images with lower errors.

3 Our Methods

In stereo matching algorithms based on convolutional neural network, supervised learning is basically a regression method which takes use of smooth L1 loss to calculate the errors between the predicted disparity value and the real disparity value to supervise the learning of the network. For the self-supervised learning, the algorithm mainly outputs labels from the feature of the original image, the features of the disparity map to achieve the purpose of training the deep learning model.

In this study, the loss is calculated by using reconstruction. It is assumed that the left image as a reference image is I_{ij}^{l} where (i, j) represent the position coordinates of the pixel point. According to the predicted disparity d and the right image I_{ij}^{r}, the reconstructed left image $\tilde{I}_{ij}^{l}$ is generated through the warping operation. However, in order to avoid that the reconstructed image has a high loss, we will utilize structural similarity index measure (SSIM) as image quality to comprehensively calculate the photometric errors between the reconstructed image and the original image [32].

$$L = \frac{1}{N}\sum_{i,j} \alpha \frac{1-SSIM(I_{ij}^{l}, \tilde{I}_{ij}^{l})}{2} + (1-\alpha) \left\| I_{ij}^{l} - \tilde{I}_{ij}^{l} \right\| \quad (1)$$

where α is the weight of the basic reconstruction loss and similarity loss. Single-scaled SSIM and simplified 3×3 filtering are adopted, α is set to 0.85.

Our network architecture extends the shared encoders [32]. A sequence of three images were fed into the model, where the first pair of images was employed to predict depth, the remaining two images were applied to predict pose. The difference is that, for depth network, we added MobileNetV2 as our encoders as shown in Fig. 1.

Although most KITTI-based depth estimation currently employs ResNet as encoder, compared to ResNet using standard convolution to extract features, MobileNetV2 utilizes the combination of depth-wise convolutions with point-wise convolutions can exponentially reduce the time complexity and space complexity. Moreover, in order to suit for depth-wise convolution, the inverted residual block applied by MobileNetV2 can extract features in higher dimensions. Regarding pose estimation, we make use of axis-angle representations to predict the rotation and scale the rotation and translation outputs by 0.01. All three pairs of images provided to the pose and depth network share with the same parameters [31].

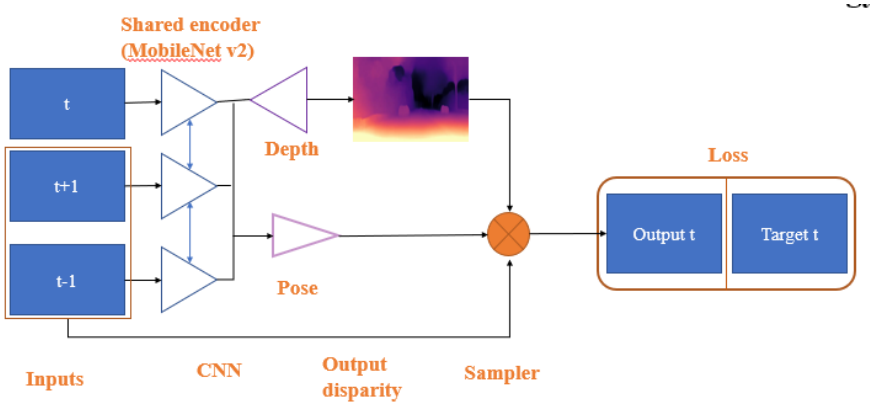

Fig. 1. The network structure of depth estimation based on network architecture and MobileNet v2 module.

To avoid the network gets stuck in a local minimum and fail to seek the global minimum, we employ the Semi-Global Matching (SGM) algorithm to generate a depth hint. We are use of depth hint for regression if the reprojected image generated with depth hint is more accurate than the network estimated [31,33]. The SGM algorithm sets a global energy function related to the disparity map, composed of the disparity of each pixel to minimize this energy function,

$$E(D) = \sum_P (C(P, D_P) + \sum_{q \in N_p} P_1 I\left[\left|D_P - D_q\right| = 1\right] + \sum_{q \in N_p} P_2 I\left[\left|D_p - D_q\right| > 1\right]) \quad (2)$$

where D is the disparity map, p and q are the pixels in the image, N_p is the adjacent pixel point of the pixel point P_d, $C(P, D_P)$ is the cost of the pixels if the disparity of the current pixel is D_P, P_1 and P_2 are penalty coefficients, which are applicable if the disparity value in the adjacent pixels of pixel P and the disparity difference of P is equal to 1 and greater than 1, respectively.

The steps of the SGM algorithm are listed as follows:

Step 1 (Pre-processing): Employ Sobel operator to process the source image, map the image processed by the Sobel operator to a new image, and obtain the gradient information of the image for subsequent calculation costs.

Step 2 (Cost calculation): Use the sampling method to calculate the gradient cost of the pre-processed image gradient information and apply the sampling method to calculate the SAD cost of the source image.

Step 3 (Dynamic planning): There are four paths by default, and the parameters P1 and P2 of path planning are set and SAD Window size.

Step 4 (Post-processing): There are four parts: Uniqueness detection, sub-pixel interpolation, left-right consistency detection, and connected area detection.

We apply the root mean squared error (RMSE), absolute relative error and squared relative error with the 1.25 as the threshold as the evaluation methods.

$$ABsRel = \frac{1}{N}\sum \frac{|d_i - d_i^*|}{d_i} \quad (3)$$

$$SqRel = \frac{1}{N}\sum \frac{|d_i - d_i^*|^2}{d_i} \quad (4)$$

where d_i and d_i^* are the ground truth and predicted depth at pixel i and N is the total number of pixels.

$$\text{RMSE}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left|d_i - d_i^*\right|^2} \quad (5)$$

where d_i is the real depth information, d_i^* is the predicted depth value, and m is the total number of pixels.

4 Experimental Results

We run the experiments based on the KITTI dataset which consists of calibrated stereo video registered to LiDAR measurements of a city, captured from a moving car. We totally obtained 1,000 data points for training and testing as shown in Fig.2. The dataset is split into training set and test set at a ratio of 7:3, the resolution of the training images are all 320×1024.

In this experiment, we apply depth hints as a substitution of depth ground truth. If the loss using depth hints is smaller than the network using ground truth for regression, we reply on depth hints to optimize the network. The visualization of depth hint maps is shown in Fig. 3.

The result shows in Fig. 4 that the model is able to well identify the distance in the scene. Moreover, the boundaries of vehicles, walls, traffic light and other objects displayed in the output images are of high definition. The depth information of the main objects in the color maps are also relatively high. In our preliminary experimental results, RMSE is up to 4.083.

In order to ensure the stability and reliability of our network, we take use of Root Mean Squared Error (RMSE), Absolute Relative Error (AbsRel), and Squared Relative Error (SqRel) as evaluation metrics to compare the performance of the networks with different encoders when training the same data set as shown in Table.1. The results show that the MobileNetV2 encoder performs better than ResNet-18 and ResNet-50 based on all three evaluation methods with and without using depth hints. In the case of employing the MobileNetV2 as the backbone, the network with depth hints outperforms the other networks. This may indicate that in this dataset, MobileNetV2 combined with depth hints is able to achieve outstanding performance.

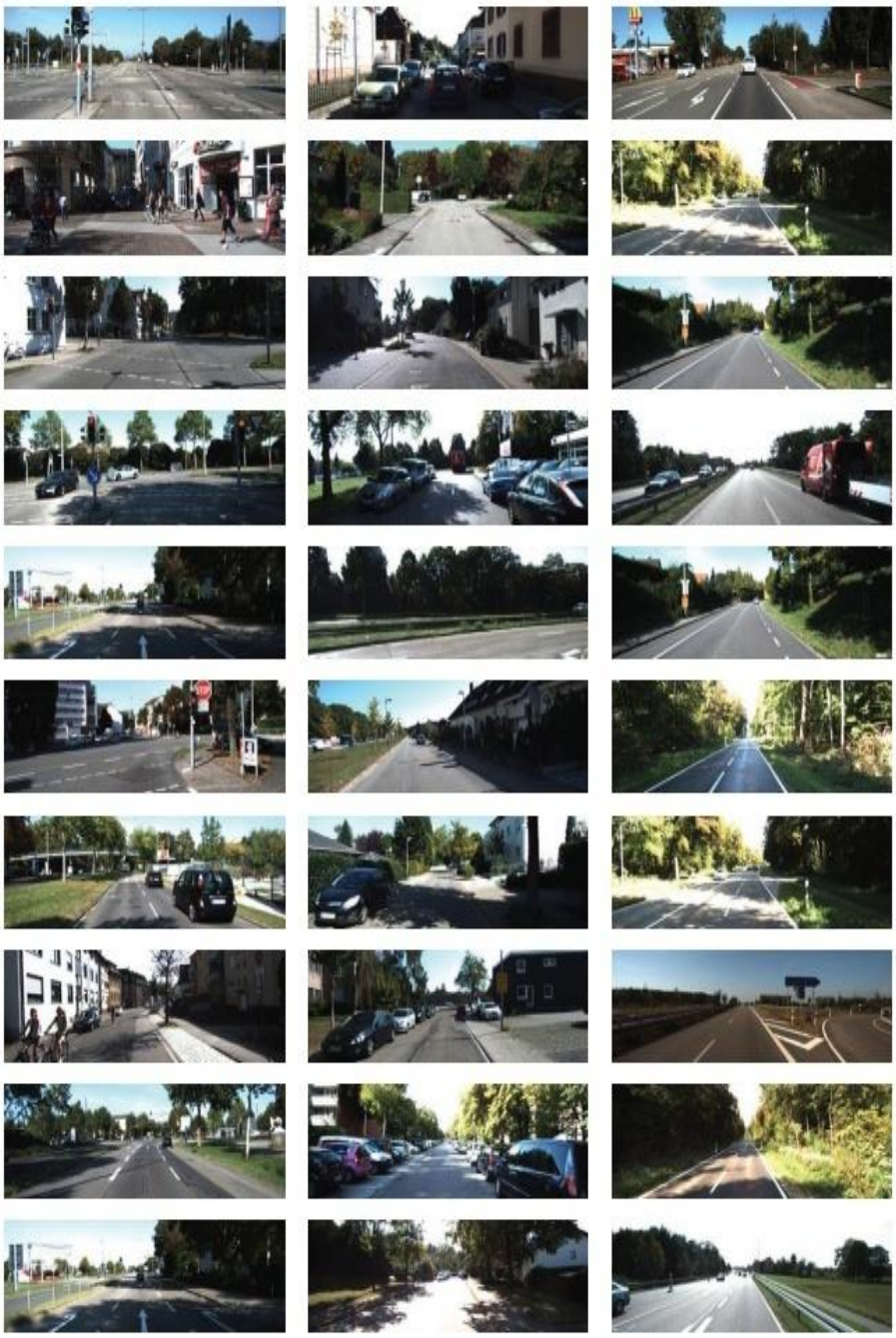

Fig.2. The original RGB images as the training data

Fig. 3. The visualization of depth hints

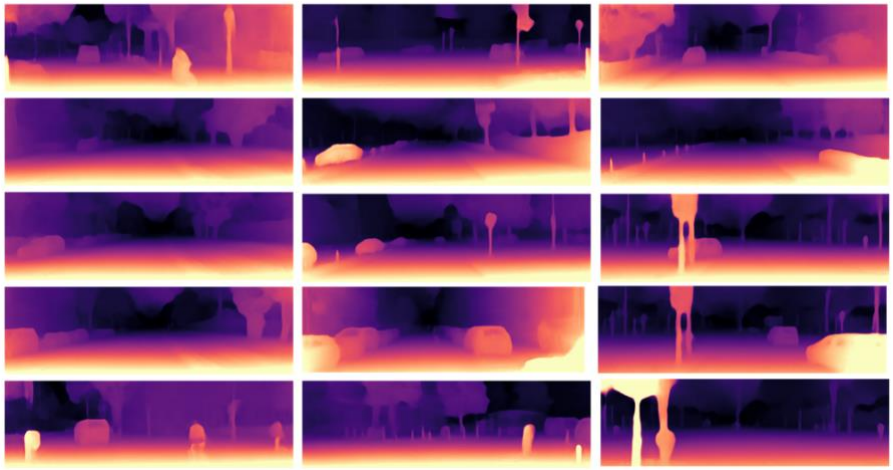

Fig. 4. The depth estimation of traffic scenes with color maps at pixel level (Brighter colors indicate closer distances, darker colors show greater distances)

Table 1. Comparisons of multiple deep neural networks

Training modality	AbsRel	SqRel	RMSE
ResNet-18	0.132	0.86	4.518
ResNet-50	0.121	0.777	4.467
MobileNetV2	0.115	0.736	4.293
ResNet-18 + depth hints	0.131	0.81	4.384
ResNet-50 + depth hints	0.120	0.726	4.395

5 Conclusion

We initially demonstrate a MobileNetV2 method combined with depth hints to infer high-resolution depth maps from 2D images. Through comparisons, we see that in this dataset, MobileNetV2 combined with depth hints performed better than other encoders. At present, the RMSE of this model has reached 4.083.

In the near future, in order to expand this work, we will generate a depth information data set of traffic scenes in New Zealand and conduct depth estimation of traffic scenes based on this data set. Moreover, we will further optimize this algorithm to obtain a higher resolution depth estimation map.

References

1. Li ,Y., Tong, G., Yang, J., Zhang, L. Peng, H.: 3D point cloud scene data ac-quisition and its key technologies for scene understanding. Laser & Optoelectron-ics Progress, 040002 (2019)
2. Liu, L., Ouyang, W., Wang, X., Fieguth, P.W., Chen, J., Liu, X., & Pietikäinen, M.: Deep learning for generic object detection: A survey. International Journal of Computer Vision (2018).
3. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T.: The rise of deep learning in drug discovery. Drug Discovery Today, 1241–1250 (2019)
4. Husain, F., Dellen, B., & Torras, C.: Scene Understanding Using Deep Learning. Academic Press, 373-382 (2017)
5. Yang, S., Wang, W., Liu, C., & Deng, W.: Scene understanding in deep learning-based end-to-end controllers for autonomous vehicles. IEEE Transactions on Sys-tems, Man, and Cybernetics: Systems, 53-63 (2019)
6. Lecun, Y., Muller, U., Ben, J., Cosatto, E., & Flepp, B.: Off-road obstacle avoid-ance through end-to-end learning. International Conference on Neural Information Processing Systems, 739-746 (2005)
7. Ohsugi, H., Tabuchi, H., Enno, H., & Ishitobi, N.: Accuracy of deep learning, a machine-learning technology, using ultra-wide-field fundus ophthalmoscopy for detecting hematogenous retinal detachment. Scientific Reports, 7(1): 9425 (2017)
8. Li, F., Deng J., & Li, K.: ImageNet: Constructing a largescale image database. Journal of Vision, 9(8): 1037-1038 (2009)
9. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., & Navab, N.: Deeper depth prediction with fully convolutional residual networks. International Conference on 3D Vision (3DV) (2016).
10. Eigen, D., & Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. IEEE International Confer-ence on Computer Vision, pp. 2650-2658, (2014).
11. Garg, R., Kumar, B.V., Carneiro, G., & Reid, I.: Unsupervised CNN for single view depth estimation: Geometry to the Rescue. ECCV, pp. 740-756 (2016).
12. Godard, C., Aodha, O. & Gabriel, J.: Unsupervised monocular depth estimation with left-right consistency. IEEE CVPR, pp. 270-279 (2017).
13. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K. & Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 01, pp. 1-1. (2020).

14. Miangoleh, S.M., Dille, S., Mai, L., Paris, S., & Aksoy, Y.: Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. IEEE CVPR. pp. 9685-9694, (2021).
15. Chaoqiang, Z., et al.: Monocular depth estimation based on deep learning: An overview. Science China Technological Sciences, pp.1-16, (2020).
16. Ochs, M., Kretz, A. & Mester, R.: SDNet: Semantically guided depth estimation network. German Conference on Pattern Recognition, pp.288-302, (2019).
17. Darabi, A. & Maldague, X.: Neural network based defect detection and depth es-timation in TNDE. NDT & E International, pp. 165-175, (2012)
18. Garg, R., Bg, V., Carneiro, G., et al.: Unsupervised CNN for single view depth es-timation: Geometry to the rescue. European Conference on Computer Vision, pp.740-756 (2016)
19. Ramirez, P., Poggi, M., Tosi, F., et al.: Geometry meets semantics for semi-supervised monocular depth estimation. Asian Conference on Computer Vision, pp.298-313 (2018)
20. Repala, V. & Dubey, R.: Dual CNN models for unsupervised monocular depth es-timation. International Conference on Pattern Recognition and Machine Intelli-gence, pp.209-217 (2019)
21. Honauer, K., Johannsen, O. & Kondermann, D., et al.: A dataset and evaluation methodology for depth estimation on 4D light fields. Asian Conference on Com-puter Vision, pp. 19-34 (2016)
22. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. IEEE Conference on Computer Vision and Pattern Recognition, pp. 5162-5170 (2015)
23. Dan, X. et al. Multiscale continuous CRFs as sequential deep networks for mo-nocular depth estimation. IEEE Conference on Computer Vision and Pattern Recognition, pp. 5354-5362 (2017).
24. Liu, J., Li, Q., Cao, R., et al.: MiniNet: An extremely lightweight convolutional neural network for real-time unsupervised monocular depth estimation. ISPRS Journal of Photogrammetry and Remote Sensing, pp.255-267 (2020).
25. Hu, J., Zhang, Y., Z., & Takayuki, O.: Visualization of convolutional neural net-works for monocular depth estimation. International Conference on Computer Vi-sion, pp. 3869-3878 (2019).
26. Ding, X., Wang, Y., Zhang, J., et al.: Underwater image dehaze using scene depth estimation with adaptive color correction. OCEANS, pp.1-5 (2017)
27. Torralba, A., & Aude, O.: Depth estimation from image structure. IEEE Transac-tions on Pattern Analysis and Machine Intelligence, pp. 1226-1238 (2002).
28. Song, W., et al.: A rapid scene depth estimation model based on underwater light attenuation prior for underwater image restoration. Pacific Rim Conference on Multimedia, pp.1-9 (2018).
29. Rajagopalan, A., Chaudhuri, S. & Mudenagudi, U.: Depth estimation and image restoration using defocused stereo pairs. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.1521-1525, (2014)
30. Chen, P., et al.: Towards scene understanding: Unsupervised monocular depth es-timation with semantic-aware representation. IEEE Conference on Computer Vi-sion and Pattern Recognition, pp. 2624-2632 (2019)
31. Watson, J., Firman, M., Brostow, G. J., & Turmukhambetov, D.: Self-supervised monocular depth hints. IEEE International Conference on Computer Vision, pp. 2162-2171 (2019).

32. Godard, C., Mac Aodha, O., & Brostow, G. J. Unsupervised monocular depth es-timation with left-right consistency. IEEE Conference on Computer Vision and Pattern Recognition, pp. 270-279 (2017).
33. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual infor-mation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(2): 328–341 (2008).
34. Godard, C., Mac Aodha, O., Firman, M., & Brostow, G. J.: Digging into self-supervised monocular depth estimation. IEEE International Conference on Com-puter Vision, 3828-3838 (2019).
35. Liu, X., Yan, W.: Traffic-light sign recognition using Capsule network. Springer Multimedia Tools and Applications (2021)
36. Liu, X., Yan, W.: Vehicle-related scene segmentation using CapsNets. IEEE IVCNZ (2020)
37. Liu, X., Nguyen, M., Yan, W.: Vehicle-related scene understanding using deep learn. Asian Conference on Pattern Recognition (2019)
38. Liu, X.: Vehicle-related Scene Understanding Using Deep Learning. Master's Thesis, Auckland University of Technology, New Zealand (2019)
39. Mehtab, S., Yan, W.: FlexiNet: Fast and accurate vehicle detection for autonomous vehicles-2D vehicle detection using deep neural network. ACM ICCCV: (2021)
40. Mehtab, S., Yan, W.: Flexible neural network for fast and accurate road scene per-ception. Multimedia Tools and Applications (2021)
41. Mehtab, S., Yan, W., & Narayanan, A.: 3D vehicle detection using cheap LiDAR and camera sensors. IEEE IVCNZ (2021)
42. Yan, W.: Computational Methods for Deep Learning: Theoretic, Practice and Ap-plications. Springer (2021)
43. Yan, W.: Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics. Springer (2019)
44. Gu, Q., Yang, J., Kong, L., Yan, W., & Klette, R.: Embedded and real-time vehicle detection system for challenging on-road scenes. Optical Engineering 56 (6), 06310210 (2017)
45. Ming, Y., Li, Y., Zhang, Z., & Yan, W.: A survey of path planning algorithms for au-tonomous vehicles. International Journal of Commercial Vehicles (2021).
46. Shen, D., Xin, C., Nguyen, M., & Yan, W.: Flame detection using deep learning. In-ternational Conference on Control, Automation and Robotics (2018)
47. Xin, C., Nguyen, M., & Yan, W.: Multiple flames recognition using deep learning. Handbook of Research on Multimedia Cyber Security, 296-307 (2020)
48. Luo, Z., Nguyen, M., & Yan, W.: Kayak and sailboat detection based on the im-proved YOLO with Transformer. ACM ICCCV (2022)
49. Le, R., Nguyen, M., & Yan, W.: Training a convolutional neural network for trans-portation sign detection using synthetic dataset. IEEE IVCNZ (2021)
50. Le, R., Nguyen, M., & Yan, W.: Training a convolutional neural network for transportation sign detection using synthetic dataset. IEEE IVCNZ (2021)
51. Pan, C., & Yan, W.: Object detection based on saturation of visual perception. Multimedia Tools and Applications 79 (27-28), 19925-19944 (2020)
52. Geiger, A., Lenz, P., & Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3354_3361 (2012)