# PADENet: An Efficient and Robust Panoramic Monocular Depth Estimation Network for Outdoor Scenes

Keyang Zhou[1], Kaiwei Wang[1] and Kailun Yang[2]

***Abstract*— Depth estimation is a basic problem in computer vision, which provides three-dimensional information by assigning depth values to pixels. With the development of deep learning, researchers have focused on estimating depth based on a single image, which is known as the "monocular depth estimation" problem. Moreover, panoramic images have been introduced to obtain a greater view angle recently, but the corresponding model for monocular depth estimation is scarce in the state of the art. In this paper, we propose PADENet for panoramic monocular depth estimation and re-design the loss function adapted for panoramic images. We also perform model transferring to panoramic scenes after training. A series of experiments show that our PADENet and loss function can effectively improve the accuracy of panoramic depth prediction while maintaining a high level of robustness and reaching the state of the art on the CARLA Dataset.**

## I. INTRODUCTION

Monocular depth estimation aims to predict the corresponding depth value of each pixel from a single RGB image, which is very important for autonomous driving, robotics, and real-world transportation applications. Panoramic images, referred as the mosaic of RGB images collected at $360^\circ$ Field of View (FoV), can greatly increase the perception range of the surrounding scenes. Despite current network structures such as ResNet [1] and DenseNet [2] which have strong extraction capabilities, their number of parameters and inference speed limit their applications in reality. Besides, those models usually have poor performance on panoramic images due to their incapability to address the significant distortions in panoramic images.

On the other side, while panoramic cameras are becoming popular for being integrated in autonomous transportation applications, large-scale omnidirectional image dataset with dense ground-truth depth is lacking, whose acquisition is highly complex [3] and sometimes prohibitive for distorted, wide-FoV data. This poses significant challenges to predict depth for images with large view angle up to $360^\circ$.

To address these problems, this paper proposes PADENet, which concatenates an improved scene understanding module after feature extraction network, and finally upsamples the feature map to the original size. At the same time, considering the lack of the panoramic RGB-Depth dataset, we leverage the equirectangular projected KITTI Dataset [3] for training. Afterwards, network parameters are transplanted to the panoramic image in order to perform depth inference. This paradigm allows to re-use conventional pinhole datasets for obtaining a depth estimation model that adapts comfortably to panoramic imagery, bypassing the prohibitive process of ground-truth acquisition for omnidirectional images.

We adopt unsupervised learning to train our depth estimation model. In the design of loss function, we apply the wrap loss of reconstructed view as a criterion. Besides, due to the distortion characteristics of the panoramic images, we apply window-based loss along with the basic wrap loss. Experiments show that the combined loss function can achieve higher quality of panoramic depth estimation.

The contributions of this paper lie in the following aspects:

- We propose PADENet, which is an effective and robust panoramic monocular depth estimation network.
- We improve the loss function for PADENet by combining the basic wrap loss and window-based loss to promote the quality of predicted depth map.
- We leverage the equirectangular projected KITTI dataset [3] for training, and then transplant the trained model to panoramic images during inference. Our implementations and codes will be available at: `https://github.com/zzzkkkyyy/PADENet`.

## II. RELATED WORK

### A. Monocular Depth Estimation on Rectified Images

Modern monocular depth estimation methods based on deep learning significantly outperform those traditional methods. Here, we mainly review those methods using deep learning. Eigen et al. [4] first used convolutional neural networks into monocular depth estimation by dividing the network into Coarse Network and Fine Network. Jiao et al. [5] enhanced monocular depth estimation with the assistance of semantic segmentation and attention-driven loss. Fu et al. [6] discretized the depth information into several bins and then performed an additional classifier to obtain final results.

While supervised learning dominates the area of monocular depth estimation, the models usually lose generalization and robustness due to dataset bias. Garg et al. [7] first proposed an unsupervised training scheme, which only required left and right views during training. The network predicted disparity maps first and converted them to depth maps later. Godard et al. [8] further improved the methods by putting

This work has been partially funded through the project "Research on Vision Sensor Technology Fusing Multidimensional Parameters" (111303I21805) by Hangzhou SurImage Technology Co., Ltd and supported by Hangzhou KrVision Technology Co., Ltd (krvision.cn). This work has also been funded in part through the AccessibleMaps project by Federal Ministry of Labor and Social Affairs.

[1]K. Zhou and K. Wang are with State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, China {zhoukeyang, wangkaiwei}@zju.edu.cn

[2]K. Yang is with Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany kailun.yang@kit.edu

forward three losses: reconstruction loss, image smoothness loss and left-right disparity map matching loss. Additionally, multiple clues are applied in training process. Diaz et al. [9] utilized the semantic segmentation to discretized depth information. Ren et al. [10] added scene understanding modules based on scene classification and coarse depth estimation to refine the predictions. However, these approaches cannot be directly applied for panoramic images due to the significant distortions and complexities.

### *B. Monocular Depth Estimation on Panoramic Images*

Despite the fact that there are many effective models dealing with monocular depth estimation on rectified images, monocular depth estimation on panoramic images is less explored due to the lack of high-quality datasets. In order to overcome these shortages, Gregoire et al. [11] proposed an effective method using the projected KITTI dataset to solve the problem of lack of training samples. Alisha et al. [12] presented CylindricalSfMLearner for estimating motion structure and used it to assist panoramic depth estimation. Keisuke et al. [13] proposed to replace the normal convolution into a distortion-aware convolution in order to learn the distorted feature better. Nikolaos et al. [14] proposed UResNet and RectNet, which were specially designed for panoramic depth estimation by adjusting the convolution's receptive field and kernel size. However, we find that most works were aimed for estimating indoor panoramic depth information. In this work we extend previous works to address outdoor panoramic monocular depth prediction.

## III. METHODOLOGY

### *A. Equirectangular Projection*

To obtain data suitable for training a panoramic depth estimation model, we present the method of projection transformation to adapt existing rectilinear image datasets into equirectangular ones. The coordinates of rectified images represent horizontal and vertical directions in the rectangular coordinate system, while coordinates of equirectangular images represent longitude and latitude directions in the spherical coordinate system. Therefore, equirectangular projection is to perform the conversion from rectangular coordinates to equirectangular coordinates. The illustration of such projection is shown in Fig. 1.

Given the equirectangular coordinate of a pixel in original image equals to ($\varphi$, $\Phi$), and the corresponding rectangular coordinate equals to ($x$, $y$, $z$), the projection can be written as Equation 1:

$$\begin{aligned} \phi &= \arctan\left(\frac{x}{z}\right) \\ \varphi &= \arcsin\left(\frac{y}{\sqrt{x^2+y^2+z^2}}\right) \end{aligned} \tag{1}$$

For the simplicity of the projection equation, we define the projection process as function $F$. Then Equation 1 can be rewritten as Equation 2:

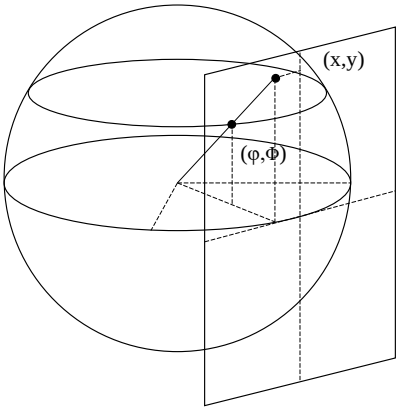


Fig. 1: Illustration of the equirectangular projection.

$$\begin{pmatrix} \phi \\ \varphi \end{pmatrix} = F \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{2}$$

We perform the mapping process given the height and width of the projected equirectangular images, which are equal to *h_equi* and *w_equi*, respectively. Here, the maximum horizontal and vertical view angle are *fov_h* and *fov_w*, respectively. The x-axis and y-axis coordinates are *X* and *Y*, respectively. Thus, the mapping equation between view angle and equirectangular coordinates is depicted in Equation 3:

$$\begin{aligned} X &= \frac{\phi \cdot w\_equi}{fov\_w} - \frac{w\_equi}{2} \\ Y &= \frac{\phi \cdot h\_equi}{fov\_h} - \frac{h\_equi}{2} \end{aligned} \tag{3}$$

Since the mapping process is linear, it can also be represented as matrix $G$, written as Equation 4:

$$\begin{pmatrix} X \\ Y \\ 1 \end{pmatrix} = G \cdot \begin{pmatrix} \phi \\ \varphi \end{pmatrix} \tag{4}$$

Finally, the transformation is combined together to obtain the equirectangular projection as Equation 5 shows:

$$\begin{pmatrix} X \\ Y \\ 1 \end{pmatrix} = G \cdot F \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{5}$$

Besides, if rectangular projection is required, we only need to perform the inverse transformation as Equation 6 shows:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = F^{-1}\left(G^{-1} \cdot \begin{pmatrix} X \\ Y \\ 1 \end{pmatrix}\right) \tag{6}$$

### *B. Dataset Transformation*

One main difficulty in panoramic monocular depth estimation is the lack of corresponding outdoor dataset. Inspired by [11], we use the equirectangular projection method mentioned in Sec. III-A to convert the rectilinear KITTI Dataset and use the projected dataset for training the panoramic

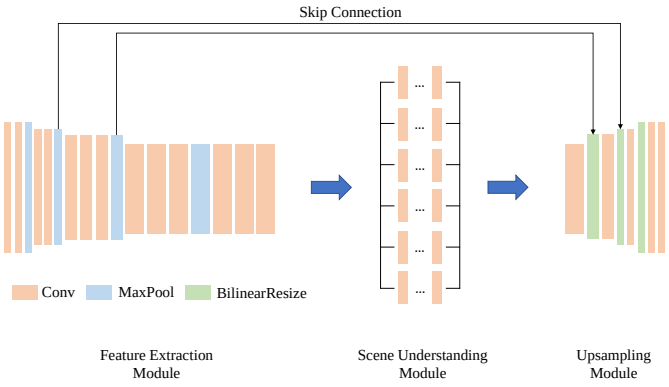


Fig. 2: Overview of the proposed PADENet architecture.

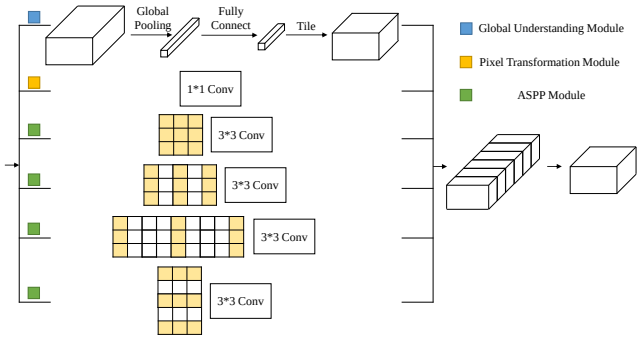


Fig. 3: The proposed Scene Understanding Module for multi-scale feature extraction.

depth estimation model. Since it is hard to obtain precise depth information in the real-world panoramic scenes, one alternative choice is to use the CARLA [15] automotive environment simulator to generate the synthetic panoramic scenes and corresponding depth ground-truth for validation.

Specifically, we convert the original KITTI Dataset [3] to equirectangular images whose resolution is $960 \times 320$, which is a subregion of $360°$-full panoramic image. We find that the training model based on the projected KITTI Dataset [3] can still maintain a high-level accuracy and robustness when testing on $360°$-full panoramic image.

### C. Network Architecture

The structure of the entire network is shown in Fig. 2, which can be divided into three parts: Feature Extraction Module, Scene Understanding Module, and Upsampling Module. In Feature Extraction Module, we use the VGG [16] network structure. In Scene Understanding Module, we design a parallel module for multi-scale extraction of the information for the input feature map, which will be described in the following paragraphs. In Upsampling Module, we use bilinear upsampling and convolution layers instead of deconvolution layers to restore the initial resolution enhanced by skip-connection mechanism. These improvements are to ensure that the feature information of the original image are well delivered to the Upsampling Module as much as possible.

Scene Understanding Module, inspired by PSPNet [17] and DORN [6], is illustrated in Fig. 3. We design the structure so that it can learn the global information and specific details at the same time using different parallel convolutions, where all useful information are aggregated into the Upsampling Module. As shown in Fig. 3, Scene Understanding Module is divided into three modules: Global Understanding Module, Pixel Transformation Module, and Atrous Spatial Pyramid Pooling (ASPP) Module. Global Understanding Module uses global pooling and fully connection to obtain a global feature vector of the feature map; Pixel Transformation Module learns the transformation of each local pixel feature through $1 \times 1$ convolutions; ASPP Module is originally used in DeepLab [18], which sets different dilation rates to cover different sizes of receptive fields. Considering that horizontal distortion of the panoramic image at the pole position is larger than vertical distortion, we design three sets of dilation rates in the horizontal direction and two sets of dilation rates in the vertical direction. The outputs are concatenated and passed to a $1 \times 1$ convolution layer to adjust the channels. In this way, a wide variety of scales is covered without losing original spatial resolution.

### D. Loss Function

In monocular depth estimation, there are two main solutions, namely supervised learning and unsupervised learning, depending on whether dataset contains ground-truth depth. We eventually adopt unsupervised learning as our training strategy since the training data contains only binocular images, and adapt the original loss function based on the characteristics of the panoramic image.

*1) Basic Unsupervised Learning Loss:* The loss function of unsupervised learning can be divided into the following three terms, in which $p$ is the original grid coordinate of the image and $d$ is the corresponding disparity value:

*Reconstruction loss.* It describes the difference between the left (right) view reconstructed by the corresponding disparity map and the real left (right) view as Equation 7 indicates:

$$L_{rect} = \frac{1}{N} \sum \left| I_{ij} - I_{G \cdot F\left(F^{-1}(G^{-1} \cdot p_{ij}) + [d,0,0]^T\right)} \right| \quad (7)$$

*Smoothness loss.* It describes the smooth gradient regularization term of disparity map as Equation 8 shows:

$$L_{smooth} = \frac{1}{N} \sum |\partial_x d_{ij}| \, e^{-\|\partial_x I_{ij}\|} + |\partial_y d_{ij}| \, e^{-\|\partial_y I_{ij}\|} \quad (8)$$

*Left-right consistency loss.* It describes the difference between the reconstructed disparity map and the original disparity map as Equation 9 shows:

$$L_{lr} = \frac{1}{N} \sum \left| d_{ij} - d_{G \cdot F\left(F^{-1}(G^{-1} \cdot p_{ij}) + [d,0,0]^T\right)} \right| \quad (9)$$

What should be aware of is that the disparity wrapping process corresponds to a curve in the panoramic image.

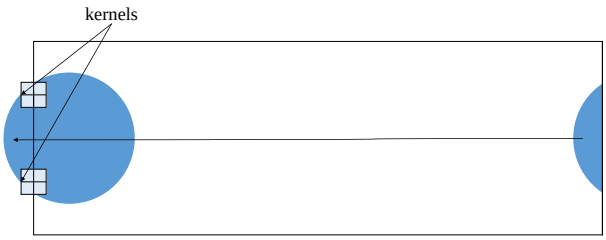


Fig. 4: Illustration of the ring padding for continuous panoramic depth estimation.

Therefore, when calculating the reconstruction loss and left-right consistency loss, images should be projected to the corresponding Cartesian coordinate system and be re-projected back after wrapping.

*2) Window-based Loss:* The basic unsupervised learning uses pixel-wise matching as the loss function, which may make network fall into the local minima and cause large prediction error in smooth areas. Inspired by ActiveStereoNet [19], we select a window around each matching pixel as the region of matching candidates, whose shape is set to $(2k+1) \times (2k+1)$. We then calculate the matching loss for each pixel in the window and summarize by weight as the final loss. With this design, each pixel in the window is regarded as a matching target, thus reducing the problem of local minima.

We let I represents pixel's value. Specifically, each pixel's matching loss is defined in Equation 10:

$$C_{ij} = \left| I_{ij} - I'_{xy} \right| \tag{10}$$

Meanwhile, each pixel's weight is defined in Equation 11:

$$w_{ij} = e^{-\frac{\left| I_{ij} - I'_{xy} \right|}{2}} \tag{11}$$

Finally, all pixel losses in the window are added with a corresponding weight, outputing the improved reconstruction loss as Equation 12 shows:

$$L_{ij} = \frac{\sum_{x=i-k}^{i+k} \sum_{y=j-k}^{j+k} w_{xy} C_{xy}}{\sum_{x=i-k}^{i+k} \sum_{y=j-k}^{j+k} w_{xy}} \tag{12}$$

*3) Fusion Strategies and Improvements:* In Encoder-Decoder mechanism, every level of decoder is a part of refinement. Therefore, different levels of loss are often added when calculating loss. We find that for low-resolution maps, the information has already been greatly pooled, thus window mechanism cannot improve the accuracy of the disparity prediction. For high-resolution maps, window mechanism is still effective because the receptive field's size is relatively small. We also verify that using the original loss function at low-resolution outputs and window-based loss at high-resolution outputs can achieve better results in the experiments.

### E. Transplanting to Panoramic Images

Because network parameters are independent of the input size, different shapes of input images will not affect the process of forward inference. Here, we use the projected KITTI images for training, and directly input the complete panoramic images by using the same parameters when performing panoramic depth estimation.

However, when we directly input the full panoramic image, the left and right edge objects may not be aligned and break the continuity. Inspired by Yang et al. [20], we adopt ring padding instead of normal padding aimed for panoramic images, which can ensure the continuity of convolution operation. In this way, the ring padding strategy can effectively keep the continuity of depth prediction, which helps to attain 360° seamless estimation, eliminating the blind spots for surrounding sensing. Details are expressed in Equation 13 and Fig. 4:

$$u_{pad} \equiv u (mod\ w) \tag{13}$$

where $u$ is the x-axis coordinate of a pixel, $u_{pad}$ the wrapped x-axis coordinate and $w$ the width of the image.

## IV. EXPERIMENTS

### A. Datasets

*KITTI*: The KITTI Dataset [3] is a widely-used autonomous driving dataset. Full KITTI Dataset [3] contains 42382 rectified stereo pairs from 61 scenes, where each image has a $1242 \times 375$ resolution. However, we find that the increasing amount of image pairs doesn't help the convergence process much. Therefore, we simply choose a subset of KITTI Dataset [3] used in object detection. We call it "KITTI Detection Dataset". The KITTI Detection Dataset contains 7481 training image pairs and 7518 testing image pairs. Here we only use its training split to yield the depth estimation model. We take the first 7200 image pairs for training and the following 281 image pairs for validation.

*CARLA*: As there is no available depth-annotated automotive panoramic imagery, we use a dataset which is generated from CARLA automotive environment simulator for testing and call it "synthetic CARLA Dataset" [15]. The synthetic CARLA Dataset contains 200 panoramic images which cover the whole 360° view angle. Each image has a ground-truth depth map, which facilitates the quantitative evaluation of our model. We follow the conventions for depth evaluation [3], and filter out the pixels whose distance is less than 0m or greater than 50m.

### B. Implementation Details

We use PyTorch [21] and a NVIDIA GeForce GTX 1080Ti GPU for our model's training and validation. Due to the limited hardware, we set batch size to 1 when training the PADENet and find that the model still behaves well. We use original loss, window-based loss and fused loss for comparison. The model is trained for 30 epochs on the projected KITTI dataset for each parameter setting. The initial learning rate is 1e-4 and is updated to 2e-5 in the

TABLE I: Basic loss v.s Window-based loss.

| Methods | Abs. Rel. | Sq. Rel. | RMSE | RMSE log | Acc: $\delta < 1.25$ |
|---|---|---|---|---|---|
| Garanderie et al. [11] | 0.231 | 6.377 | 3.598 | 0.463 | 0.716 |
| Ours with basic loss | 0.203 | 5.274 | 3.571 | 0.446 | 0.738 |
| Ours with window loss at all levels | **0.200** | **4.427** | **3.550** | **0.443** | **0.752** |

TABLE II: Comparison among different settings of window shape.

| Methods | Abs. Rel. | Sq. Rel. | RMSE | RMSE log | Acc: $\delta < 1.25$ |
|---|---|---|---|---|---|
| Ours with window loss at all levels | 0.200 | 4.427 | 3.550 | 0.443 | 0.752 |
| Ours with window loss at last 1 level | 0.180 | 4.063 | 3.290 | 0.420 | 0.757 |
| Ours with window loss at last 2 levels | **0.174** | 3.865 | **3.132** | **0.411** | **0.769** |
| Ours with window loss at last 2 levels with dilated rate 2 | **0.174** | **3.384** | 3.230 | 0.419 | 0.750 |

last 5 epochs for fine-tuning. Adam optimizer is used during training.

### C. Quantitative Results

We conduct 2 groups of experiments, one of which is to validate the effectiveness of our PADENet and the proposed window-based loss compared to original unsupervised loss, while the other is to find the best fusion parameters setting for loss function. All those experiments are conducted by training on the projected KITTI Detection Dataset and testing on the synthetic CARLA Dataset. The evaluation metrics include Absolute Relative Error (Abs. Rel.), Square Relative Error (Sq. Rel.), Root Mean Square Error (RMSE and RMSE log) and $\delta$ threshold (Acc). We test our model's quantitative metrics on the valid split of synthetic CARLA Dataset.

Since only Garanderie et al.'s work [11] use synthetic CARLA Dataset for the quantitative evaluation about outdoors panoramic depth estimation, here we just compare our results with theirs. As displayed in Table I, our results in different evaluation metrics both outperform Garanderie et al.'s work [11] by large margins, reaching the new state of the art of outdoor panoramic monocular depth estimation. Besides, the network inference frame rate can reach 50fps on the GTX 1080Ti GPU processor, which makes the model possible to infer in real time, which is critical for autonomous driving systems.

### D. Quantitative Analysis

According to the experimental results, we find that PADENet can improve the quality of panoramic depth estimation significantly. Additionally, we test different fusion strategies in order to get better performance.

As is shown in Fig. 2, PADENet has 4-level output maps from low to high resolutions. We set the first 4-k layers to calculate the original unsupervised losses, and the last k layers to calculate window-based losses. We perform experiments by setting k = 1 and 2 respectively due to hardware limitations. The window size is set to 11 according to previous comparison experiments. In addition, we adopt a similar idea of dilated convolution, and set the dilation rate to 1 and 2. respectively.

As is displayed in Table II, it can be seen that final depth prediction reaches better performance when k is set to 2. We also find that if dilate rate is changed from 1 to 2, while the result measured in square relative error is slightly better, the performance on RMSE and accuracy is worse. This shows that dilated rate is not very relevant. Finally, we use the trained model with k = 2 as our final predicting model. Some representative prediction samples on the projected KITTI Detection Dataset are shown in Fig. 5.

The qualitative panoramic depth estimation effect on the synthetic CARLA Dataset [15] is shown in the Fig. 6. We find that the trained model on the Projected KITTI Detection Dataset can be transplanted well into 360° panoramic images. At the same time, it has an excellent depth estimation effect for cars, boxes, telephone boxes and other objects in the CARLA Dataset [15]. Comparing the results of training with full window-based loss in Table II, we find that the fusion strategy improves the quality of depth prediction. At the same time, the continuity of some slender objects has also been greatly improved. This is because the calculation process of window-based loss aggregates all pixels in the candidate window. Overall, the proposed PADENet enables to attain fully dense, seamless and precise depth estimation in 360°, beneficial for surrounding sensing of autonomously driving vehicles.

## V. CONCLUSIONS

Monocular depth estimation is a traditional problem in computer vision, while recently it gains striking progress due to the rapid development of deep learning. Meanwhile, panoramic monocular depth estimation receives increasing attention these days because of its 360° view angle. In this paper, we proposed a network called PADENet. Additionally, we fuse the original unsupervised loss and window-based loss. All these methods and innovations lead to much higher performance, allowing PADENet to achieve the new state of the art of panoramic monocular depth estimation, which prove the effectiveness of our proposals.

However, there are still some points that we can work up with in the future. For instance, window-based loss can be updated in order to fit the panoramic images better. Besides, considering that the resolution of panoramic images is usually much larger than normal rectified images, more efficient backbone can be introduced in the model. In the future, we aim to optimize our model and utilize new techniques such

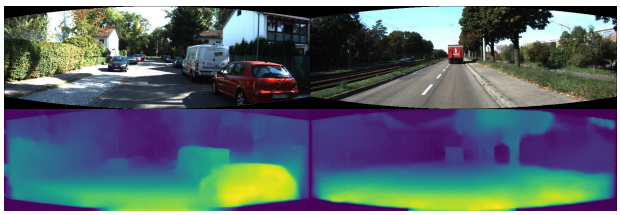

Fig. 5: Qualitative examples of using PADENet on the projected KITTI Detection Dataset.

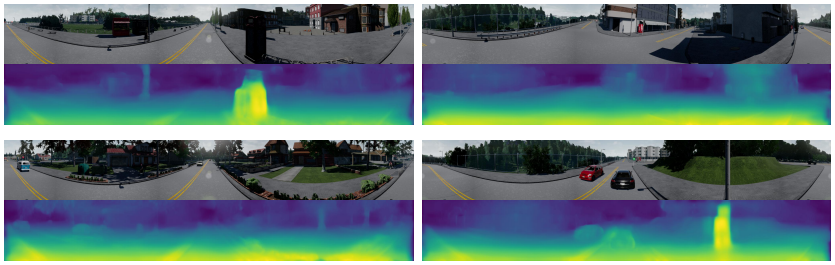

Fig. 6: Qualitative examples of using PADENet on CARLA Dataset.

as GANs [22] to further improve the quality of panoramic monocular depth estimation, which can be deployed on real-time vision sensors and to assist the real-world autonomous transportation applications.

## REFERENCES


[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[3] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

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

[5] J. Jiao, Y. Cao, Y. Song, and R. Lau, "Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 53–69.

[6] 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.

[7] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 340–349.

[8] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.

[9] R. Dıaz and A. Marathe, "Soft labels for ordinal regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] H. Ren, M. El-khamy, and J. Lee, "Deep robust single image depth estimation neural network using scene understanding," *arXiv preprint arXiv:1906.03279*, 2019.

[11] G. Payen de La Garanderie, A. Atapour Abarghouei, and T. P. Breckon, "Eliminating the blind spot: Adapting 3d object detection and monocular depth estimation to 360 panoramic imagery," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 789–807.

[12] A. Sharma and J. Ventura, "Unsupervised learning of depth and ego-motion from panoramic video," *arXiv preprint arXiv:1901.00979*, 2019.

[13] K. Tateno, N. Navab, and F. Tombari, "Distortion-aware convolutional filters for dense prediction in panoramic images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 707–722.

[14] N. Zioulis, A. Karakottas, D. Zarpalas, and P. Daras, "Omnidepth: Dense depth estimation for indoors spherical panoramas," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 448–465.

[15] A. Dosovitskiy, G. Ros, F. Codevilla, A. M. Lopez, and V. Koltun, "Carla: An open urban driving simulator," *arXiv: Learning*, 2017.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[17] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[18] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv: Computer Vision and Pattern Recognition*, 2017.

[19] Y. Zhang, S. Khamis, C. Rhemann, J. Valentin, A. Kowdle, V. Tankovich, M. Schoenberg, S. Izadi, T. Funkhouser, and S. Fanello, "Activestereonet: End-to-end self-supervised learning for active stereo systems," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 784–801.

[20] K. Yang, X. Hu, L. M. Bergasa, E. Romera, X. Huang, D. Sun, and K. Wang, "Can we pass beyond the field of view? panoramic annular semantic segmentation for real-world surrounding perception," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 446–453.

[21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[22] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," *arXiv: Learning*, 2017.