Generating Pedestrian Views from In-Vehicle Camera Images

Daina Shimoyama, Fumihiko Sakaue and Jun Sato
Nagoya Institute of Technology, Nagoya 466-8555, Japan
{*shimoyama@cv., sakaue@, junsato@*}*nitech.ac.jp*

Keywords: GAN, Semantic Segmentation, Multi-Task Learning, Pedestrian Views, In-Vehicle Camera.

Abstract: In this paper, we propose a method for predicting and generating pedestrian viewpoint images from images captured by an in-vehicle camera. Since the viewpoints of an in-vehicle camera and a pedestrian are very different, viewpoint transfer to the pedestrian viewpoint generally results in a large amount of missing information. To cope with this problem, we in this research use the semantic structure of the road scene. In general, it is considered that there are certain regularities in the driving environment, such as the positional relationship between roads, vehicles, and buildings. We generate accurate pedestrian views by using such structural information on the road scenes.

1 INTRODUCTION

In recent years, many methods have been proposed to predict the behavior of pedestrians on roads for avoiding car accidents and for realizing autonomous driving.

The standard method for predicting the route of a pedestrian is to use RNNs and LSTMs (Alahi et al., 2016). The orientation of the pedestrian's head and the scene structures are also used for improving the accuracy of route prediction (Lee et al., 2017; Yagi et al., 2018).

Furthermore, it is expected that the accuracy of pedestrian behavior prediction can be further improved by using visibility information that indicates what kind of scenery the pedestrian is looking at. For example, if there is a pedestrian crossing in front of a pedestrian's view, we can predict that the pedestrian is likely to head there to cross the street. Thus, in this paper, we propose a method to predict and generate what kind of scenery the pedestrians are seeing from each viewpoint based on the camera images mounted on the vehicle.

The viewpoint transfer can be achieved by first converting the in-vehicle image to a pedestrian viewpoint image, and then inpainting the missing parts in the converted image. Since the 3D structure of the scene is required for viewpoint transfer, we obtain the depth information of each point in the image using a depth estimation method based on monocular images (Godard et al., 2019), and then transfer the in-vehicle image to an arbitrary pedestrian viewpoint image.

(a) in-vehicle image (b) pedestrian viewpoint image

Figure 1: Viewpoint transfer with depth estimation. The viewpoint-transformed image in (b) contains a large amount of missing regions that are not visible in the original camera image in (a).

Fig. 1 (b) is an example pedestrian viewpoint image obtained by converting the in-vehicle image in Fig. 1 (a) estimating the depth information from (Godard et al., 2019). As shown in this figure, the generated image contains a large amount of missing regions that are not visible from the in-vehicle camera viewpoint. Therefore, we need to inpaint the missing parts in the next step.

In recent years, cGAN-based methods (Mirza and Osindero, 2014; Isola et al., 2017) have achieved significant results in such image transformation tasks. However, simple image transformation networks cannot inpaint an image accurately when a large amount of regions are missing in the image like the image in Fig. 1 (b). Thus, in this paper, we propose a method to generate a pedestrian viewpoint image with higher accuracy by simultaneously estimating the scene structure specific to the road environment.

In general, the road environment has certain regularities, such as the positional relationship between

Shimoyama, D., Sakaue, F. and Sato, J.
Generating Pedestrian Views from In-Vehicle Camera Images.
DOI: 10.5220/0011736300003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 4: VISAPP*, pages 385-392
ISBN: 978-989-758-634-7; ISSN: 2184-4321

roads and vehicles, roads and buildings, and intersections and pedestrian crossings. Thus, if we can train the network to learn such regularities, the network may generate more accurate complemented images. In this research, we force the network to learn scene structure recovery as well as image inpainting and to generate accurate complemented images.

2 RELATED WORK

A common method of viewpoint transfer is to place multiple cameras around the 3D scene and use these camera images to generate images from arbitrary viewpoints by interpolation (Kanade et al., 1997; ISHIKAWA, 2008; Lipski et al., 2010; Chari et al., 2012; Sankoh et al., 2018). These methods are used to generate free viewpoint images, such as in sports broadcasting. However, these methods require a large number of cameras densely placed around the scene, and thus these methods cannot be used in road environments.

In order to solve this problem, we consider viewpoint transfer from images obtained by a single in-vehicle camera. Some methods have been proposed for generating new views from a single viewpoint image by using geometric transformations, such as projective transformations (Kazuki Ichikawa and Jun Sato, 2008). By estimating the scene depth information, we can further improve the geometric viewpoint transfer. Eigen et al. proposed a deep learning-based method for estimating scene depth from monocular images (Eigen et al., 2014). An unsupervised learning method for depth estimation is also proposed by using a pair of stereo cameras during the network training (Garg et al., 2016; Godard et al., 2017).

Although the depth information enables us to transfer the image point in one view to the image point in a new view, the geometric transformation alone can only produce image information that exists in the original image, and it cannot generate images that do not exist in the original image, such as the image obtained by looking sideways.

In recent years, image inpainting has been developed as a technique for filling in the missing image information, and it has been shown that images can be restored accurately even when many defective pixels are scattered throughout the image (Bertalmio et al., 2000; Liao et al., 2021). However, since these image inpainting methods use the similarity and regularity of only 2D image features, they require non-defective pixels to be scattered throughout the image. Therefore, the existing image inpainting methods do not work properly when we have large missing regions like the image in Fig. 1 (b), which are caused by the difference in viewing direction before and after the viewpoint transformation.

In order to solve this problem, we in this paper use multi-task learning to simultaneously learn two tasks, that is inpainting the in-vehicle images and inferring the structural information of the hidden scene. Multitask learning is a method that improves the performance of each task by learning multiple related tasks at the same time. Examples of multi-task learning include Faster R-CNN (Ren et al., 2016) and YOLO (Redmon and Farhadi, 2018), which simultaneously perform object class recognition and object location estimation, and Mask R-CNN (He et al., 2017), which simultaneously performs semantic segmentation in addition to object recognition using Faster R-CNN (He et al., 2017). The network structure for multi-task learning can take various forms depending on the number and types of tasks, but the basic structure consists of a task-sharing layer that learns features common to each task and a task-specific layer that learns features specific to each task.

In this paper, we propose a method for learning image inpainting and structural inference simultaneously by using multi-task learning, and performing viewpoint transfer based on the inference of the structural information of the invisible scene.

3 GENERATING PEDESTRIAN VIEWPOINT IMAGE

In this research, viewpoint transfer images with missing regions are complemented using a network based on conditional GAN (cGAN). However, viewpoint transfer with significantly different viewpoints can result in very large missing regions in images. Therefore, a simple image inpainting method cannot complete missing images with high quality. In this research, we propose a method to generate pedestrian viewpoint images with high quality by recovering the scene structure unique to road scenes while performing image completion. For this objective, we propose two methods: a method based on multi-task learning (Method 1) and a method using Semantic Loss with multi-task learning (Method 2).

3.1 Generating Pedestrian Viewpoint Images Using Multi-Task Learning

In a road scene, there are objects unique to the road scene, such as roads, cars, and buildings, each of which has a similar general shape. In addition, road

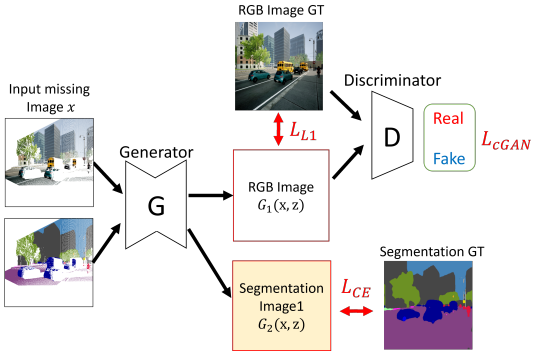

Figure 2: Network structure of method 1. Generator receives missing RGB images and missing label images, and outputs generated RGB images that complement missing RGB images and generated label images that complement missing label images.

scenes have various unique properties, such as a road stretching across the bottom of the image and buildings likely to line both sides of the road. By learning these unique properties of road scenes, we can expect more accurate missing image completion.

The structural information of the scene can be represented by semantic label images obtained from the semantic segmentation. However, the semantic label images obtained from the input missing images also have missing regions and are incomplete. Therefore, we propose a method that learns two tasks simultaneously, one is a task to complement the missing RGB images and the other is a task to complement the missing semantic label images. These two tasks are related to each other, but are different. The first task focuses on the scene appearance, while the second task focuses on the scene structure. By adopting the multi-task learning of complementing the missing RGB images and missing label images, it may be possible to share features for complementing missing information and to incorporate structural information of the scene more efficiently into the learning process. We call this method 1 in this paper.

The network structure of the proposed method 1 is shown in Fig. 2. The generator receives missing RGB images and missing label images, and outputs generated RGB images that complement the missing RGB images and generated label images that complement the missing label images. The RGB image pairs and label image pairs are used as inputs for training, and the training is performed based on the following evaluation equation:

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN} + \lambda_{RGB}\mathcal{L}_{RGB} + \lambda_{Label}\mathcal{L}_{Label} \quad (1)$$

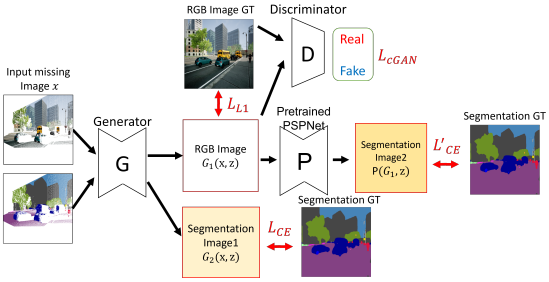

Figure 3: Network structure of method 2. We evaluate the quality of complemented RGB images by combining semantic loss with adversarial loss to further improve the performance of the generator.

In-vehicle Pedestrian 1 Pedestrian 2

Figure 4: Examples of images obtained from Airsim. We generated images seen at the same time from in-vehicle camera viewpoints and various pedestrian viewpoints.

$$\mathcal{L}_{cGAN} = \boldsymbol{E}_{x_1,y_1}[\log D(x_1,y_1)] + \boldsymbol{E}_{x_1,x_2,z}[\log(1-D(x_1,G_1(x_1,x_2,z)))] \quad (2)$$

$$\mathcal{L}_{RGB} = \boldsymbol{E}_{x_1,x_2,y_1,z}[||y_1 - G_1(x_1,x_2,z)||_1] \quad (3)$$

$$\mathcal{L}_{Label} = \boldsymbol{E}_{x_1,x_2,y_2,z}[-\sum_C y_2 \log G_2(x_1,x_2,z)] \quad (4)$$

where x_1 is the missing RGB image, x_2 is the missing label image, y_1 is the target RGB image, y_2 is the target label image, and z is the input noise. $G_1(x_1,x_2,z)$ is the complemented RGB image generated by G, and $G_2(x_1,x_2,z)$ is the complemented label image generated by G. Note that x_2 and y_2 are generated from x_1 and y_1 by using PSPNet (Zhao et al., 2017).

3.2 Generation of Pedestrian Viewpoint Images Using Semantic Loss

We next explain a method for further improving the generated pedestrian viewpoint images by using semantic loss. Since the viewpoint transformation causes a large amount of missing regions in the image, it is very difficult to evaluate the complemented image. Thus, we evaluate the quality of the complemented RGB image by combining a semantic loss with an adversarial loss to further improve the performance of the generator. We call this Method 2.

The network structure of the proposed method 2 is shown in Figure 3. The generator takes a missing RGB image and a missing label image as inputs, and outputs a complemented RGB image and a complemented label image. The complemented RGB image

is then input to the pre-trained PSPNet to perform semantic segmentation, and the L1 norm of the obtained semantic label image and its ground truth image is added as the semantic loss. Since the semantic loss evaluates the structural correctness of the generated RGB images, it constrains the generated RGB images based on higher-level evaluations, and we can expect better training of the generator.

The following evaluation equations are used to train the generator in Method 2.

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN} + \lambda_{RGB}\mathcal{L}_{RGB} + \lambda_{Label}\mathcal{L}_{Label} + \lambda_{Sem}\mathcal{L}_{Sem} \quad (5)$$

$$\mathcal{L}_{cGAN} = \boldsymbol{E}_{x_1,y_1}[\log D(x_1,y_1)] + \boldsymbol{E}_{x_1,x_2,z}[\log(1-D(x_1,G_1(x_1,x_2,z)))] \quad (6)$$

$$\mathcal{L}_{RGB} = \boldsymbol{E}_{x_1,x_2,y_1,z}[||y_1 - G_1(x_1,x_2,z)||_1] \quad (7)$$

$$\mathcal{L}_{Label} = \boldsymbol{E}_{x_1,x_2,y_2,z}[-\sum_C y_2 \log G_2(x_1,x_2,z)] \quad (8)$$

$$\mathcal{L}_{Sem} = \boldsymbol{E}_{x_1,x_2,y_1,z}[-\sum_C G_P(y_1)\log G_P(G_1(x_1,x_2,z)] \quad (9)$$

where, $G_P(\cdot)$ represents the label image generated by PSPNet.

4 DATASET

In order to train the network of the proposed method, it is necessary to prepare pairs of in-vehicle images and pedestrian viewpoint images. To create such pairs, images from two different viewpoints must be acquired at the same time, but it is very difficult to obtain a large number of such real image pairs. Thus, we in this research generate a synthetic image dataset by using Airsim.

Airsim (Shah et al., 2017) is an outdoor scene simulator that can simulate vehicle and drone views on a map built on the Unreal Engine. Airsim can simulate various conditions such as weather and location, and can acquire RGB images, depth images, and segmentation images of the scene. In this research, we used Airsim to obtain images from in-vehicle camera viewpoints and various pedestrian viewpoints at the same time, as shown in Fig. 4, to create paired images for network training.

5 EXPERIMENTS

We next show the experimental results of the proposed method. We prepared 13,200 pairs of missing images and ground truth images of various pedestrian viewpoints as described in the previous section. We used 12,000 sets as training data and 1,200 sets as test data, and trained and tested the proposed method. The learning rate and the number of epochs vary depending on the proposed method, and we used the values shown in Table 1 for each method.

We compare our two methods with the existing method, pix2pix (Isola et al., 2017). The accuracy of each method is evaluated quantitatively as well as qualitatively by using synthetic images. We also evaluate our method by using real images to show the efficiency of the network trained with synthetic images.

5.1 Synthetic Image Experiments

We first show the results from synthetic images. Fig. 5 shows the pedestrian viewpoint images obtained from test in-vehicle images by using the proposed method and the existing method. From these images, we find that the objects such as vehicles, buildings, and roads are generated more accurately by using method 1 and method 2.

In particular, in the images in the first row of Fig. 5, we find that method 1 and method 2 clearly generate the white lines and sidewalks in front of the road, while the existing method fails to generate the missing areas in front of the road. In the image in the third row, we find that method 1 and method 2 generate better images reproducing the boundary line between the road and the grass, as well as the car. Furthermore, even when the viewpoints are very different and the information in the input image is very limited, as in the example in the fifth row, the proposed method can recover the buildings and road regions fairly accurately, while the existing method fails to recover the scene. This is because the proposed method recovers semantic label images that represent the structural information of the scene.

5.2 Accuracy Evaluation

We next evaluate the accuracy of the proposed method quantitatively. As evaluation metrics, we used LPIPS (learned perceptual image patch similarity) for the generated RGB images and mIoU (mean intersection over union) for the generated label images. We also evaluated mIoU of label images obtained by inputting the generated RGB images to PSPNet. Table 2 shows LPIPS of the generated RGB images, mIoU of generated label images, and mIoU of label images obtained from the generated RGB images. These two mIoUs are indicated by mIoU1 and mIoU2 respectively. The table shows that the proposed method improves on all

Table 1: Details of learning networks for the proposed method.

Method	Generator Learning rate	Discriminator Learning rate	epoch number
existing method	2.0×10^{-4}	1.0×10^{-6}	300
method 1	2.0×10^{-4}	2.0×10^{-6}	400
method 2	2.0×10^{-4}	1.0×10^{-6}	400

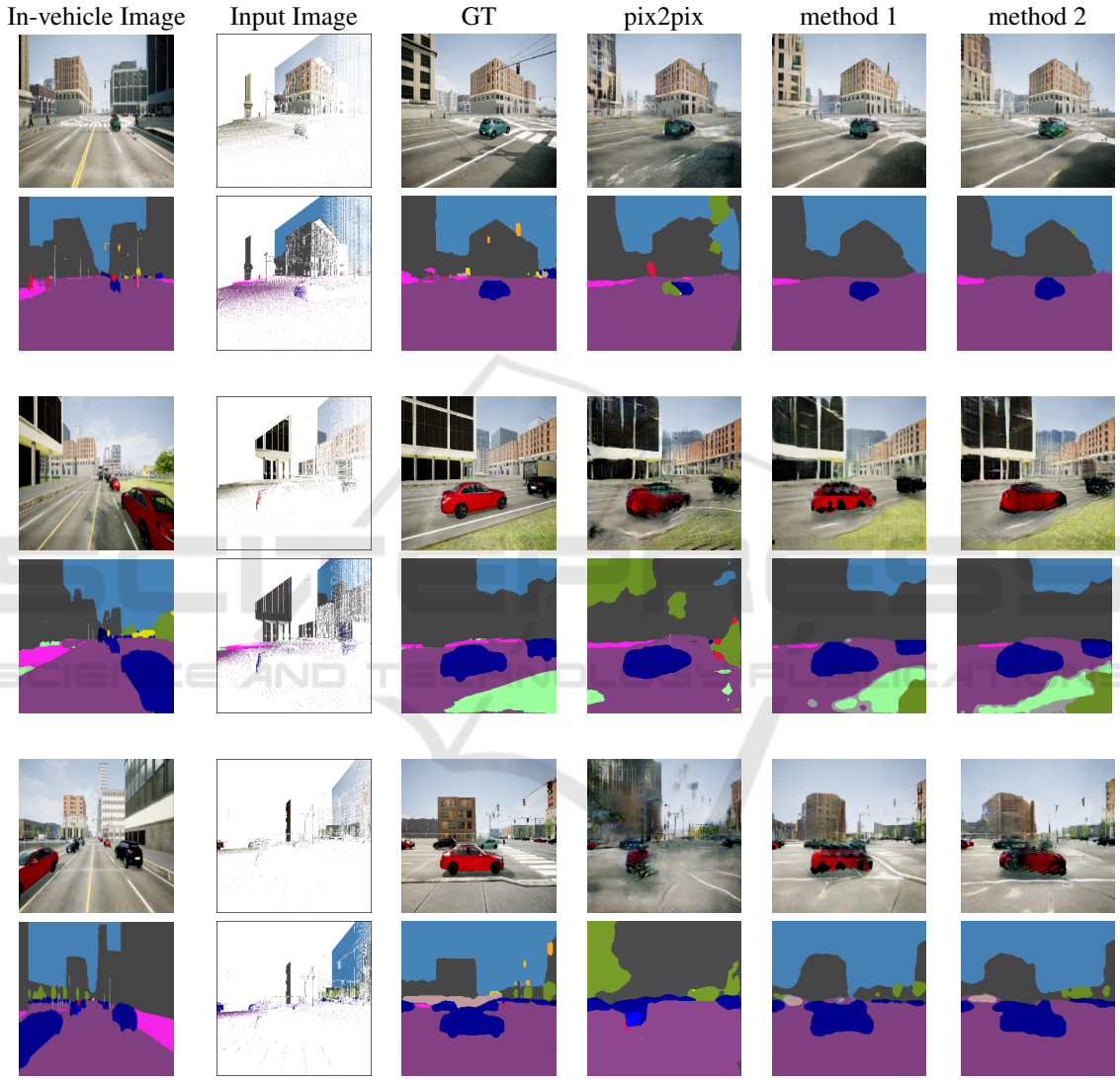

Figure 5: Synthetic image experiments. The rightmost three columns show pedestrian viewpoint images and segmentation images obtained from in-vehicle images in the leftmost column by using the proposed method and the existing method.

metrics compared to the existing method, confirming the effectiveness of the proposed method.

We next evaluate our method by using pre-trained YOLO (Redmon and Farhadi, 2018). In this experiment, we evaluated the closeness of the generated images to the ground truth images by comparing the detection results of vehicles and person output by pre-trained YOLO between the ground truth images and the generated images.

Figure 6 shows the object detection results obtained from YOLO. It is clear that the detection results for the images generated by method 2 are closer to the detection results for the ground truth images than those generated by the existing method. In par-

Table 2: Accuracy of pedestrian viewpoint image generation. We evaluated the accuracy of generated pedestrian viewpoint images by using LPIPS, and the accuracy of segmentation images by using mIoU (mIoU1). We also evaluated mIoU of label images obtained by inputting the generated pedestrian viewpoint images to PSPNet (mIoU2).

	existing method	method1	method2
LPIPS (↓)	0.423	0.356	**0.352**
mIoU1 (↑)	-	0.276	**0.368**
mIoU2 (↑)	0.1474	0.235	**0.366**

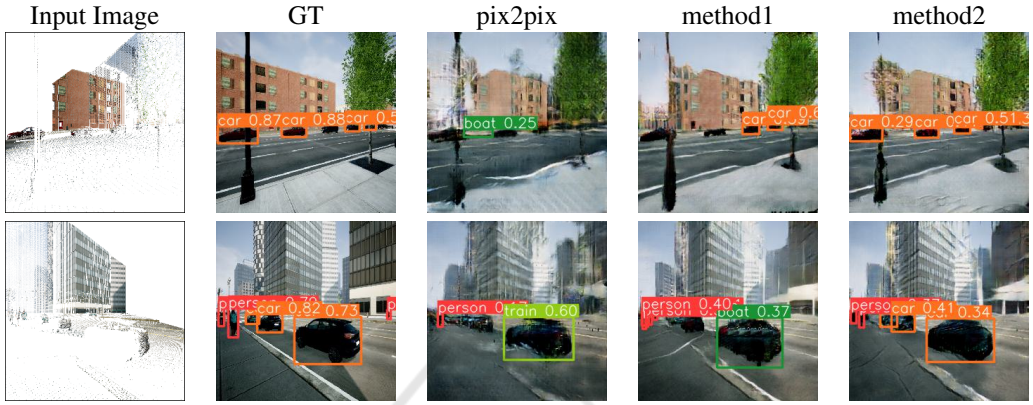

Figure 6: Evaluation of the quality of generated images using YOLO. The rightmost three columns show objects extracted by using YOLO from images generated by using the proposed method and the existing method.

ticular, YOLO cannot detect several cars in the images generated by the existing method, whereas it can detect cars accurately in the images generated by the proposed method 2. These results show that the proposed method can generate images that are closer to the ground truth images.

5.3 Real Image Experiments

We next show the results from real images. In this experiment, we evaluated our method by using 1200 real images in the Cityscapes dataset (Cordts et al., 2015). We tested our networks and pix2pix trained by using synthetic images as before.

The second column in Fig. 7 shows the input missing images viewed from the pedestrian viewpoint, which are generated from the in-vehicle images shown in the first column in Fig. 7 by using the depth estimation net (Godard et al., 2019). The third column shows the pedestrian views generated from the existing method, and the fourth and fifth columns show those from the proposed method trained on the synthetic images.

These images show that the images generated by the proposed method with semantic loss reproduce roads and objects much more accurately than those by the existing method, even in real images. Especially, in the images in the second and third rows of Fig. 7, it can be seen that the proposed method can predict plausible buildings in the large missing area where no information exists in the input images. Although the actual images cannot be obtained and the comparison with the ground truth images is not possible, we find that the proposed method can generate realistic images comparable to human imagination.

6 CONCLUSION

In this paper, we proposed a method for generating pedestrian viewpoint images from in-vehicle images by using deep learning.

We first proposed a method based on multi-task learning, which simultaneously learns a task to complement missing RGB images and a task to complement missing label images. We next extended our method by adding semantic loss derived by using the pre-trained semantic segmentation network. We also constructed the training dataset for our network by generating synthetic pairs of road environment images using Airsim simulator.

Experimental results show the effectiveness of the proposed method using structural information in the scene. However, since sufficient accuracy has not yet been obtained for real images, further improvement in performance is necessary for future work.

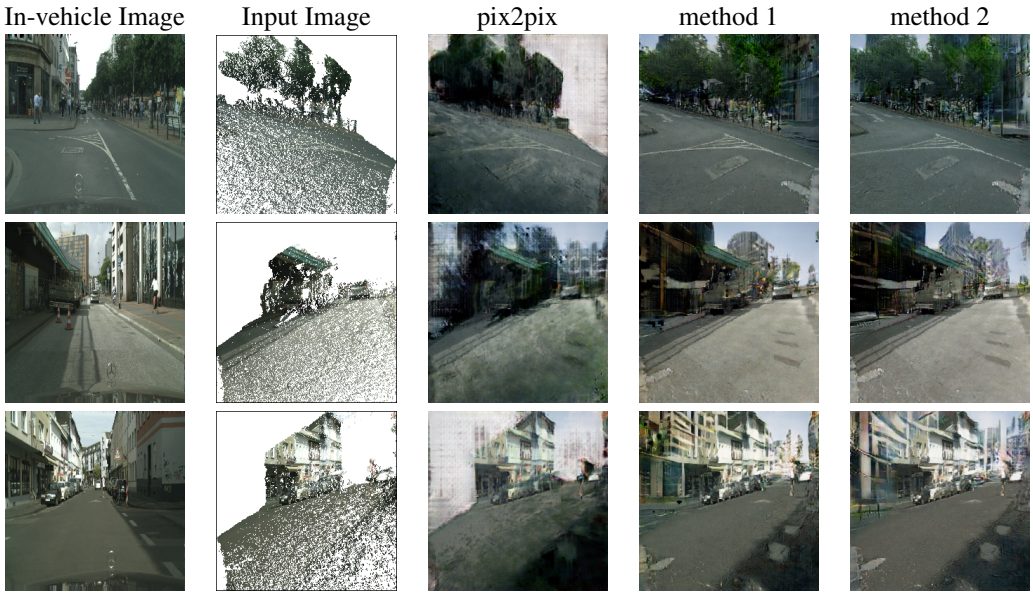

Figure 7: Real image experiments. The rightmost three columns show pedestrian views obtained from in-vehicle images in the leftmost column by using the proposed method and the existing method.

REFERENCES

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., and Savarese, S. (2016). Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971.

Bertalmio, M., Sapiro, G., Caselles, V., and Ballester, C. (2000). Image inpainting. In *Proc. ACM SIG-GRAPH*, pages 417—-424.

Chari, V., Agrawal, A., Taguchi, Y., and Ramalingam, S. (2012). Convex bricks: A new primitive for visual hull modeling and reconstruction. In *2012 IEEE International Conference on Robotics and Automation*, pages 770–777. IEEE.

Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2015). The cityscapes dataset. In *CVPR Workshop on The Future of Datasets in Vision*.

Eigen, D., Puhrsch, C., and Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27.

Garg, R., Bg, V. K., Carneiro, G., and Reid, I. (2016). Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European conference on computer vision*, pages 740–756. Springer.

Godard, C., Mac Aodha, O., and Brostow, G. J. (2017). Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279.

Godard, C., Mac Aodha, O., Firman, M., and Brostow, G. J. (2019). Digging into self-supervised monocular depth prediction.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

ISHIKAWA, A. (2008). Free viewpoint video generation for walk-through experience using image-based rendering. *ACM Multimedia 2008, Vancouver, Canada, Oct.-Nov.*

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

Kanade, T., Rander, P., and Narayanan, P. (1997). Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multimedia*, 4(1).

Kazuki Ichikawa and Jun Sato (2008). Image synthesis for blind corners from uncalibrated multiple vehicle cameras. In *2008 IEEE Intelligent Vehicles Symposium*, pages 956–961.

Lee, N., Choi, W., Vernaza, P., Choy, C. B., Torr, P. H., and Chandraker, M. (2017). Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 336–345.

Liao, L., Xiao, J., Wang, Z., Lin, C., and Satoh, S. (2021). Image inpainting guided by coherence priors of semantics and textures. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

Lipski, C., Linz, C., Berger, K., Sellent, A., and Magnor, M. (2010). Virtual video camera: Image-based viewpoint navigation through space and time. In *Computer*

Graphics Forum, volume 29, pages 2555–2568. Wiley Online Library.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

Sankoh, H., Naito, S., Nonaka, K., Sabirin, H., and Chen, J. (2018). Robust billboard-based, free-viewpoint video synthesis algorithm to overcome occlusions under challenging outdoor sport scenes. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1724–1732.

Shah, S., Dey, D., Lovett, C., and Kapoor, A. (2017). Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*.

Yagi, T., Mangalam, K., Yonetani, R., and Sato, Y. (2018). Future person localization in first-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7593–7602.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890.