Projecting Your View Attentively: Monocular Road Scene Layout Estimation via Cross-view Transformation (Supplementary Material)

Weixiang Yang[1],Qi Li[1],Wenxi Liu[1*],Yuanlong Yu[1*],Yuexin Ma[2,3],Shengfeng He[4],Jia Pan[5]
[1]*College of Mathematics and Computer Science, Fuzhou University*
[2]*ShanghaiTech University* [3]*Shanghai Engineering Research Center of Intelligent Vision and Imaging*
[4]*School of Computer Science and Engineering, South China University of Technology*
[5]*Department of Computer Science, The University of Hong Kong*

In this supplemental material, we show more examples for our ablation studies (Sec. 1), the state-of-the-art comparison experiments (Sec. 2), as well as the HD map generation (Sec. 3). Besides, we also elaborate the implementation details of our model architecture in Sec. 4.

1. Ablation studies

1.1. Cross-view Transformation

In Fig. 3, we show several exemplar results concerning with the ablation study of the cross-view transformation module in our paper (i.e. Table 4 of the paper). The examples are selected from the results of *KITTI 3D Object* dataset. With the addition of sub-structures (i.e., MLP, cross-view correlation, cycle structure, and feature selection) in the cross-view transformation module, our model can effectively extract the masks of the individual vehicles, remove noises, and refine their shapes.

1.2. Context-aware Discriminator

As shown in Fig. 4, we observe that, although the framework equipped with PatchGAN [3] improves the AP metrics from the baseline, it degrades the IOU metrics, which is consistent with the results reported in our paper (i.e. Table 6 of the paper). This is because the vehicle masks produced by PatchGAN [3] tend to slightly shrink (see the third example in Fig. 4). Our context-aware discriminator manages to improve the IOU and AP metrics simultaneously, which can produce the results that well match the ground-truth masks.

2. Comparison with the State-of-the-arts

We first demonstrate the comparison results from *Argoverse* in Fig. 1. Besides, we illustrate more comparison results for road layout estimation in Fig. 5 from the datasets *KITTI Odometry* and *KITTI Raw*, and vehicle occupancies

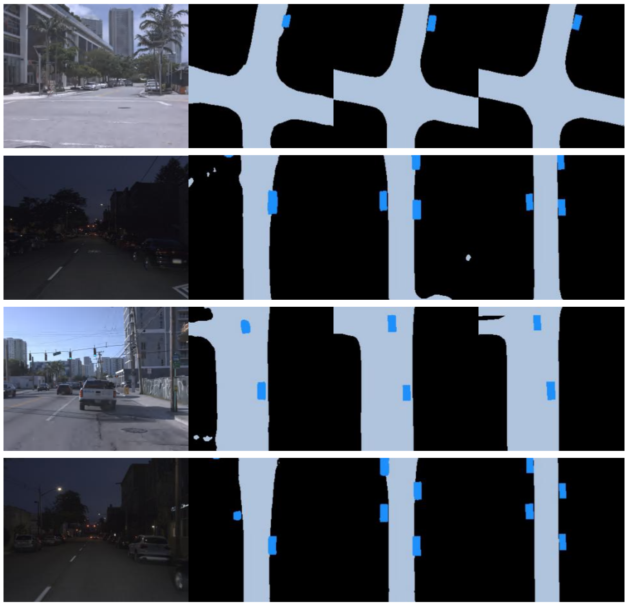


Figure 1. Exemplar results on the joint road layout and vehicle occupancy estimation from *Argoverse*.

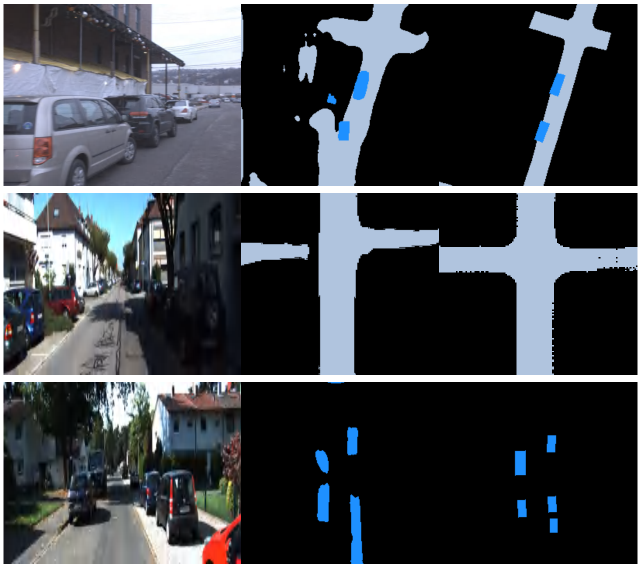


Figure 2. Failure cases on *Argoverse*, *KITTI 3D Object*, and *KITTI Odometry*.

Methods	FPS
Mono3D [1]	0.24
OFT [7]	<5
MonoOccupancy [4]	15
MonoLayout [5]	32
Ours	35

Table 1. A comparative study of inference time on NVIDIA GeForce GTX 1080Ti GPU for different methods on the images of *KITTI RAW* dataset.

in Fig. 6 from the dataset *KITTI 3D Object*.

3. Panorama HD Map Generation

For generating HD map, we use trivial image stitching techniques to showcase the results. In particular, we simply leverage the camera parameters and poses of each frame provided by *Argoverse* to montage the estimations of sampled images to obtain the HD map. To show that our approach can be applied for generating panorama HD map, we illustrate two more examples in Fig. 7 and Fig. 8.

4. Network Details and Runtime Performance

Encoder. Our encoder is built on top of the ResNet-18 [2], which takes in the RGB images with the size $1024 \times 1024 \times 3$ as input, and produces $32 \times 32 \times 512$ feature maps as the encoded features. In particular, we use the ResNet-18 [2] architecture without bottleneck layers.

Cycle structure. The MLP contains two fully-connected layers and ReLU activation.

Cross-view transformer. Each item of the key K, the query Q, and the value V in CVT utilizes a single convolutional layer (kernel size 1×1, stride $= 1$, padding $= 0$), respectively. $\mathcal{F}_{conv}$ of CVT applies one convolutional layer (kernel size 3×3, stride $= 1$, padding $= 1$) to compress the features.

Decoder. The decoder consists of four deconvolution (i.e., transposed convolution) blocks. Each block increases the spatial resolution by a factor of 2, and decreases the number of channels to 64, 32, 16, and 2, respectively. It outputs the feature map with the size $256 \times 256 \times 2$, which spatially corresponds to a rectangular region of $40m \times 40m$ area on the ground space.

Discriminator. $\mathcal{F}_D$ consists of 5 convolution blocks. Except that the last convolutional layer, each block contains a convolution layer along with the batch normalization and LeakyReLU. $\mathcal{F}'_D$ consists of 3 convolution blocks. Except that the last convolutional layer, each block has a convolution, spectral normalization, and LeakyReLU.

Runntime Information. During training, our model takes around 50 epochs (2.5h to 10.5h) to converge and requires 5.029 GB GPU memory. During inference, we process each image for 0.0286 sec. on average and 0.79 GB GPU memory.

Timing analysis. We also show the inference time of our method comparing to other competing methods in Table 1.

5. Failure cases

Fig. 2 shows a few scenarios in which we cannot produce accurate road layout estimation. First, our model may produce inaccurate prediction in the cases of occlusion and sharp turns (see the first and second rows in Fig. 2). Second, our model may be confused by multiple close vehicles (see the third row in Fig. 2).

References

[1] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 2

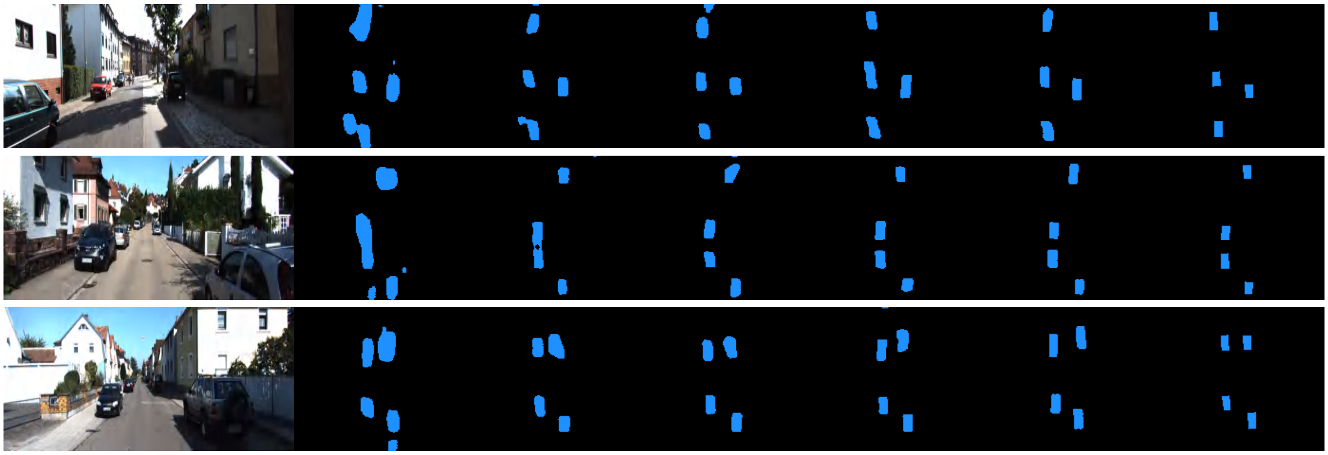
[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1, 3

[4] Chenyang Lu, Marinus Jacobus Gerardus, Van De Molengraft, and Gijs Dubbelman. Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks. *IEEE Robotics & Automation Letters*, 4(2):445–452, 2019. 2, 4

[5] Kaustubh Mani, Swapnil Daga, Shubhika Garg, N. Sai Shankar, Krishna Murthy Jatavallabhula, and K. Madhava Krishna. Monolayout: Amodal scene layout from a single image. 2020. 1, 2, 4, 5

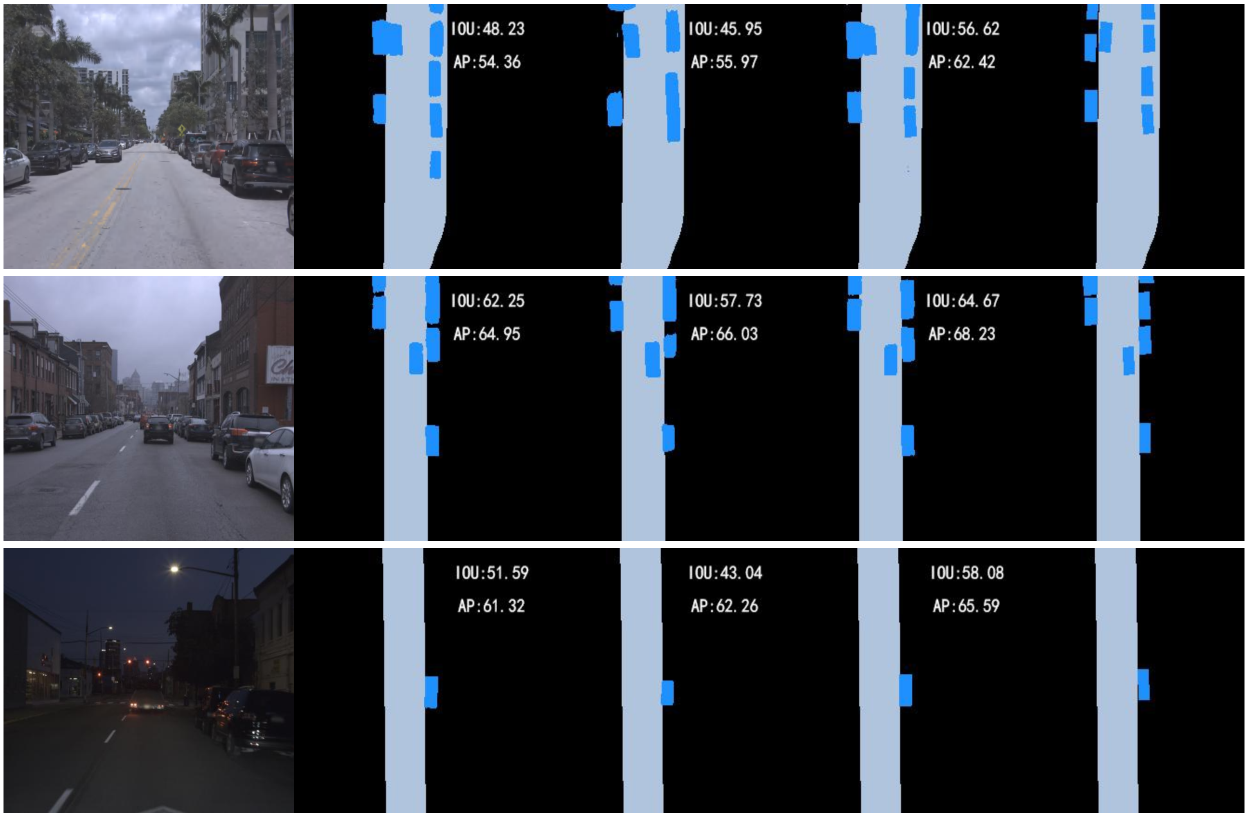
[6] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics and Automation Letters*, 5(3):4867–4873, 2020. 4

[7] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection, 2018. 2



Front view Baseline + MLP + Correlation + Cycle structure + Feature selection Ground-truth

Figure 3. Exemplar results of the ablation study for the cross-view transformation module.



Front View Baseline PatchGAN [3] Ours Ground-truth

Figure 4. Exemplar results of ablation study for the context-aware discriminator. We provide the corresponding IOU and AP metrics.

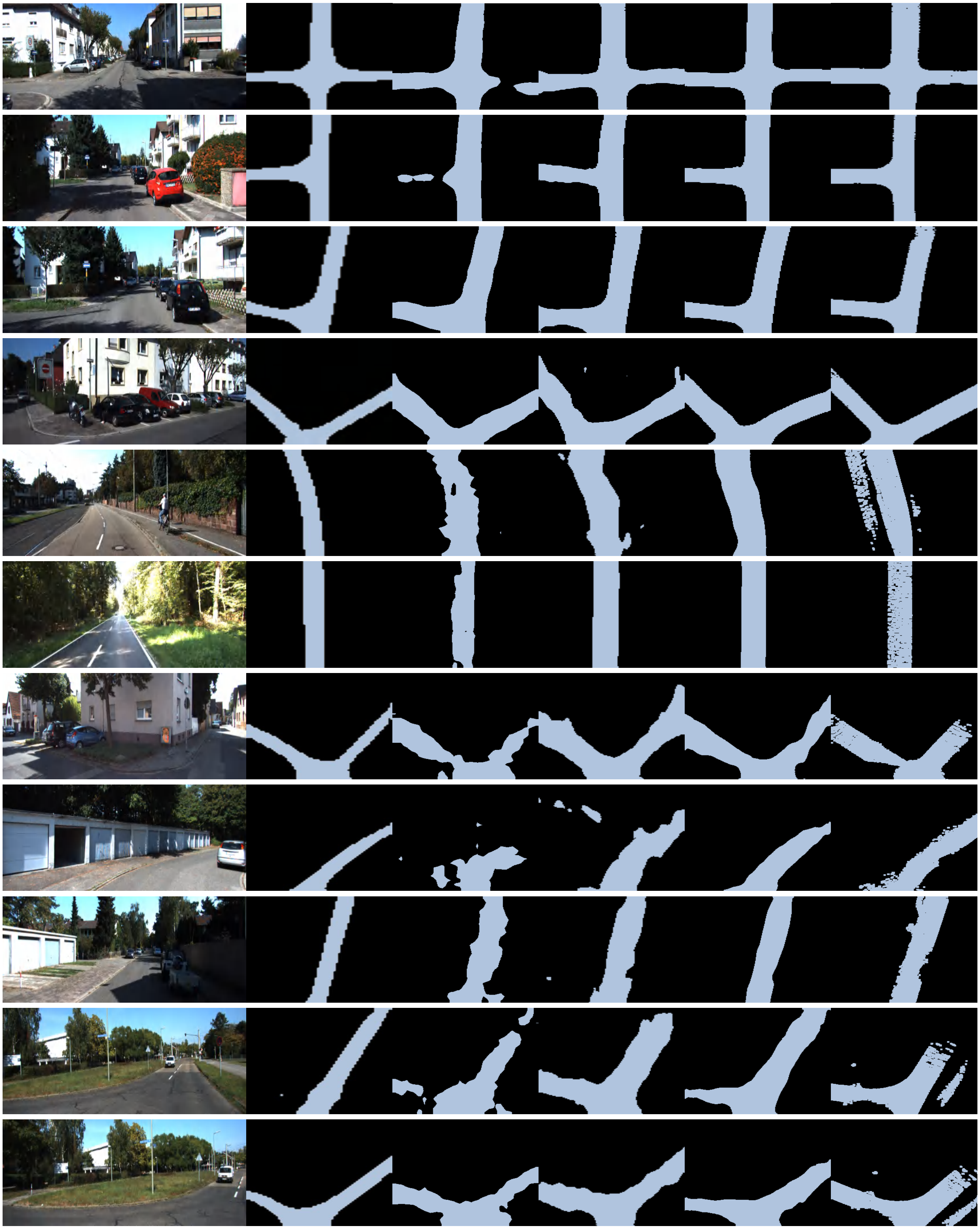


Figure 5. Comparison results of road layout estimation on *KITTI Odometry* and *KITTI Raw*. Although ground-truths contain noises, our estimation results demonstrate the better coverage for road layout than others.

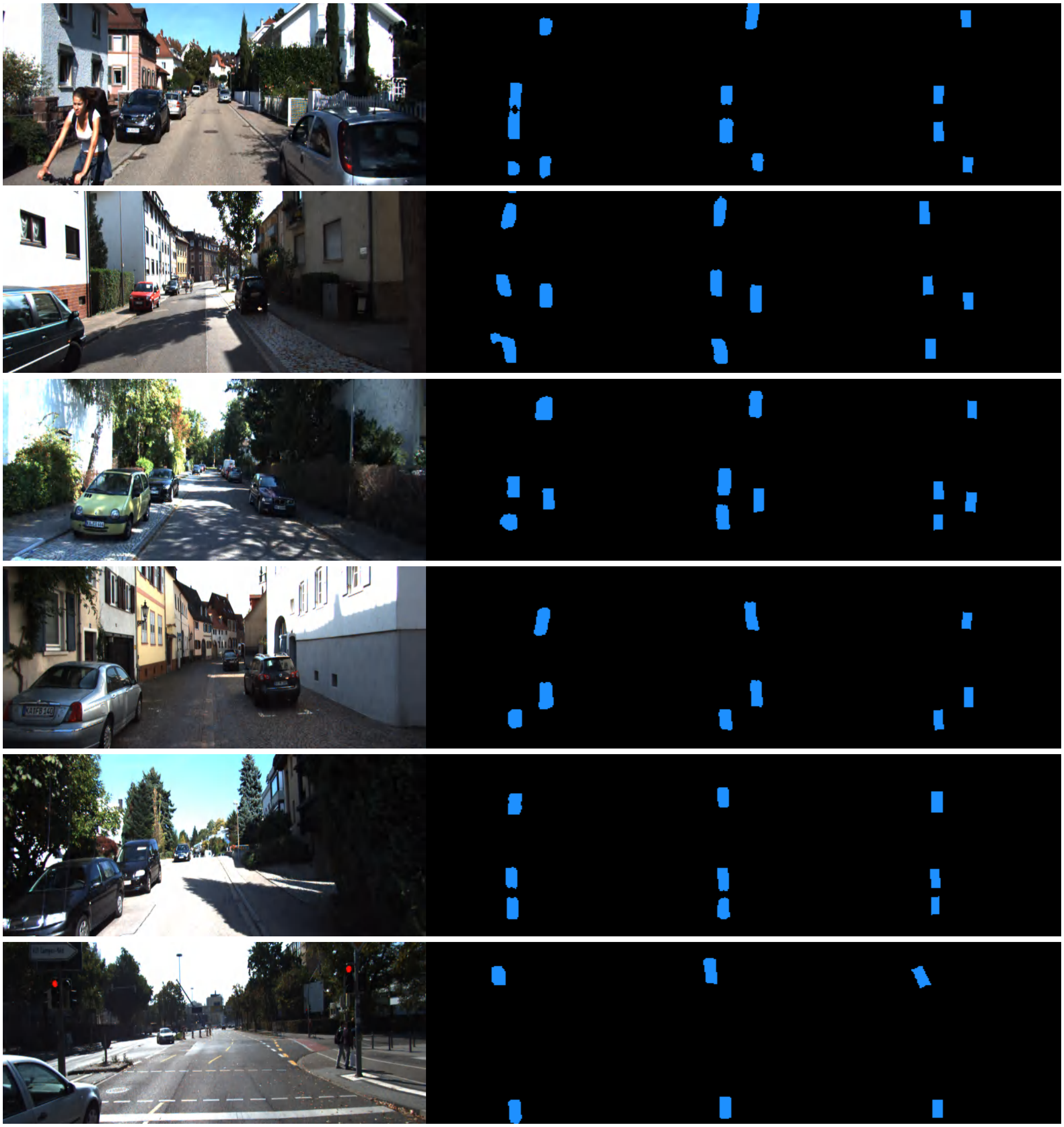
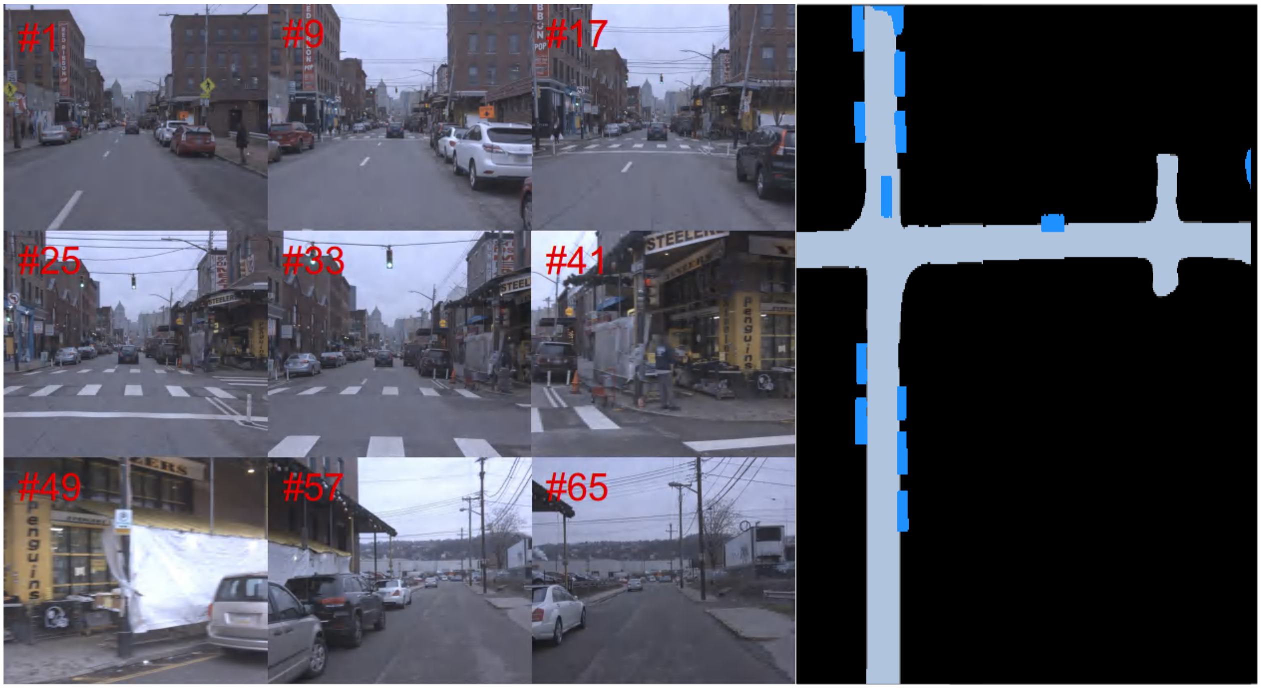


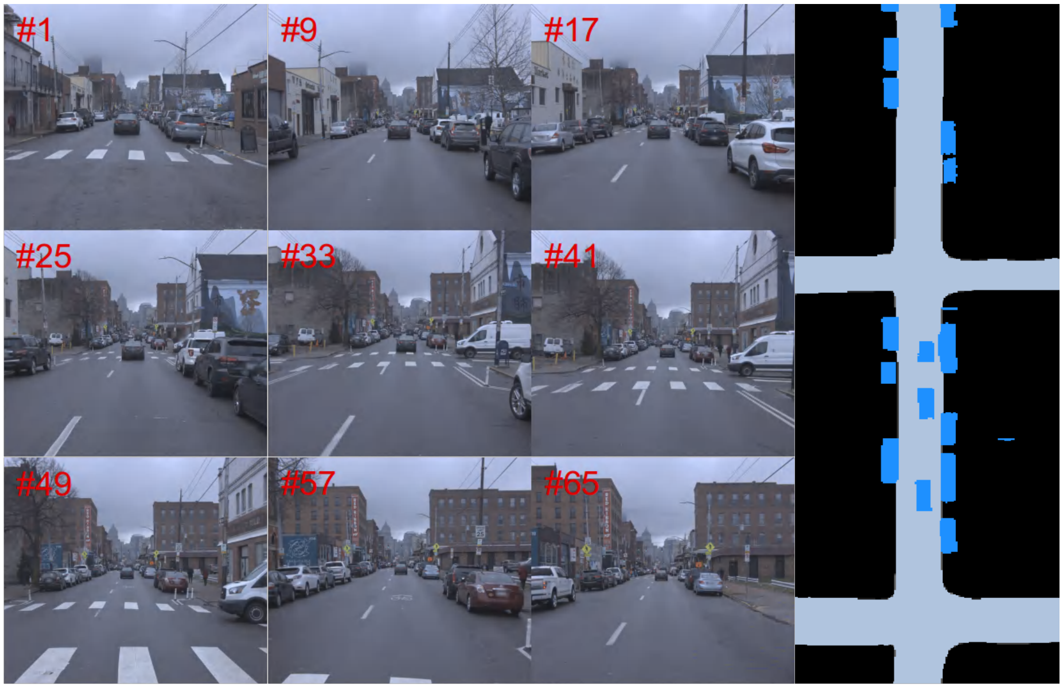
Figure 6. Vehicle occupancy estimation results on *KITTI 3D Object*.



Key frames of *Argoverse*

Panorama HD map

Figure 7. We fuse the estimated road layout from the image sequences of *Argoverse* to produce a panorama HD map.



Key frames of *Argoverse*

Panorama HD map

Figure 8. We fuse the estimated road layout from the image sequences of *Argoverse* to produce a panorama HD map.