Semi-parametric Image Synthesis Supplementary Material

Xiaojuan Qi
CUHK

Qifeng Chen
Intel Labs

Jiaya Jia
CUHK

Vladlen Koltun
Intel Labs

1. Qualitative Results

We provide additional qualitative results. Specifically, we provide comparisons with Pix2pix [2] and the CRN [1] on Cityscapes-coarse, Cityscapes-fine, Cityscapes→GTA5, NYU, and ADE20K. Comparisons on Cityscapes-coarse are shown in Figures 5 and 6. Comparisons on Cityscapes-fine are shown in Figures 7 and 8. Cityscapes→GTA5 comparison is presented in Figures 9 and 10. Comparison on NYU is shown in Figure 11. Comparison on ADE20K is shown in Figure 12. To evaluate the results of pix2pix and CRN with coarse input layouts, we trained pix2pix and CRN on coarse label maps. For the results shown in the paper, we did not apply the synthesized semantic layout to pix2pix and CRN. But we did try this and the results are similar. The performance of pix2pix and CRN on coarse label maps does not change substantially when we use the synthesized semantic layout as their input (for both training and testing).

2. Details on Perceptual Experiment

There are totally 400 users who participated in the all the experiments. All the results are statistically significant with $p < 10^{-3}$ as shown in the main paper. In the unlimited-time comparisons, there are 50 users for Cityscapes, 50 users for ADE20k and 20 users for NYU (the dataset is small). Each comparison is viewed by evaluated by at least 10 users. For the time limited experiments on Cityscapes, we have 180 users who participated in the Cityscape-coarse experiment and 100 users for Cityscape-fine experiment.

3. Image statistics

Mean power spectral statistics over Cityscapes and NYU dataset is shown in Figure. 1. The mean power spectrum of images synthesized by our approach is much more similar with that of real images. The mean power spectrum of images synthesized by Pix2pix and CRN with many local maxima that do not exist in real images.

4. Random example

We show randomly selected results in Figure. 17. Even for randomly picked results, our results realistic in most cases.

5. Diversified Synthesis

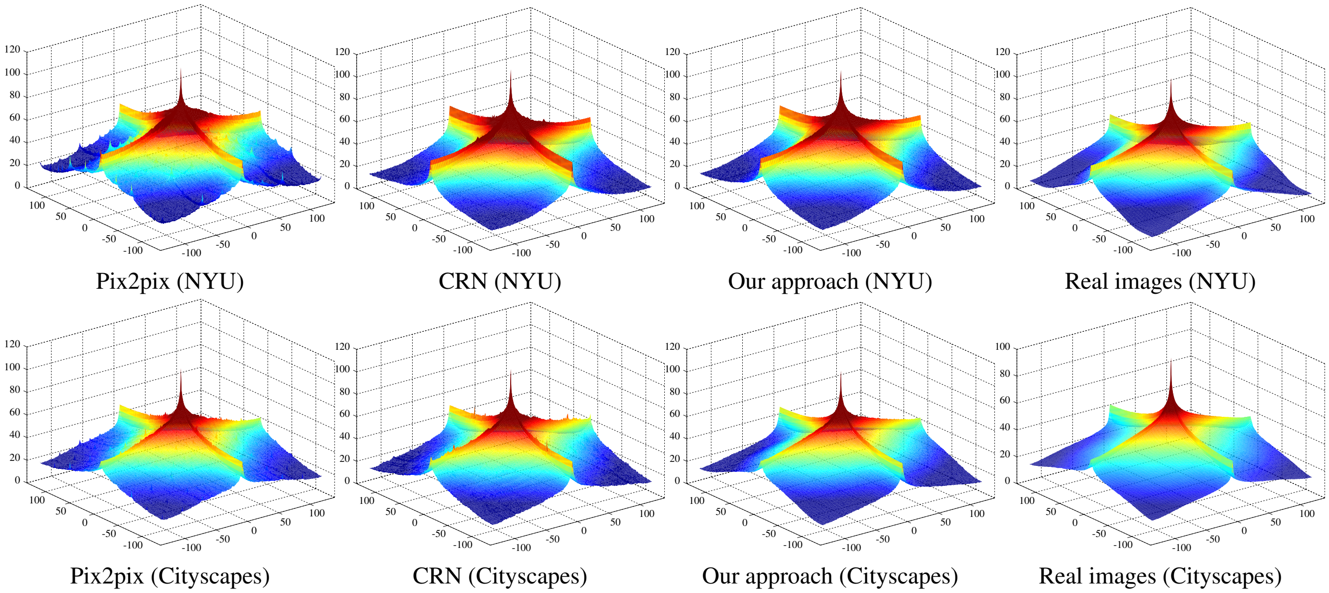
Results of diversified synthesis on Cityscapes are shown in Figure 13. Results on ADE20K are shown in Figures 14 and 15.

6. Ablation Study

We study the contribution of two major components in our model: the external memory and simulation of artifacts during the construction of the simulated canvas during training (Section 5.2). We train two ablated models from scratch on the Cityscapes dataset with coarse label maps. For the w/o memory setting, no canvas is produced. (The content of the canvas comes from the external memory.) So the synthesis network only gets the semantic layout as input, no canvas. The architecture of the synthesis network is the same. We perform the same pairwise comparisons for our complete model and the ablated models. The results are reported in Table 1. Our complete model is more realistic than the ablated models. A qualitative comparison is provided in Figure 16. Without the memory module, the result lacks the rich texture, structure, and color of natural images. Without data augmentation, color and illumination can be unnatural, as in Figure 16(a,c), boundaries may not be cleanly blended, as in Figure 16(b,d), and some regions are not effectively inpainted, as in Figure 16(e,f). Cars all seem to be floating over the road without proper shadows.

	Preference rate
Ours>w/o memory	80.6%
Ours>w/o simulated canvas	65.3%

Table 1. Ablation study. We compare our complete model to two ablated models: one without memory and one without injection of artifacts in the construction of the simulated canvas (Section 5.2). We conduct a perceptual experiment via randomized A/B tests on Amazon Mechanical Turk.



Pix2pix (NYU) CRN (NYU) Our approach (NYU) Real images (NYU)

Pix2pix (Cityscapes) CRN (Cityscapes) Our approach (Cityscapes) Real images (Cityscapes)

Figure 1. Mean power spectra over the NYU and Cityscapes dataset. Magnitude is on a logarithmic scale. We compare the mean power spectra of images synthesized by Pix2pix, CRN, and our approach to the mean power spectrum of real images from the NYU and Cityscapes test set. The mean power spectrum of images synthesized by our approach is virtually indistinguishable from the mean power spectrum of real images, while the mean power spectra of images synthesized by Pix2pix and the CRN are characterized by spurious spikes. Zoom in for details.

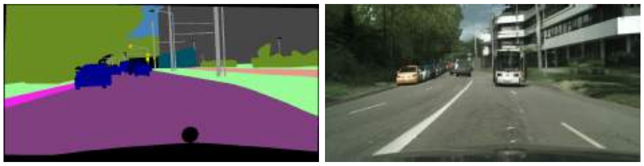


Figure 2. First failure mode: synthesized image may not conform precisely to the input semantic layout.

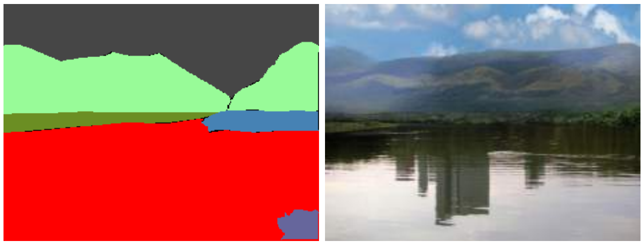


Figure 3. Second failure mode: contextual inconsistency.

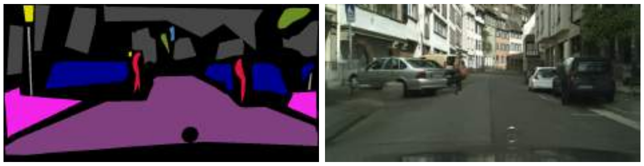


Figure 4. Third failure mode: improper image segment patch retrieved. Zoom in for details.

7. Failure Modes

We have observed two failure modes in the results produced by the presented approach. The first is that the synthesized image may not conform precisely to the input semantic layout. This is illustrated in Figure 2. The second failure mode is that different segments composited in the synthesized image may be contextually inconsistent. This is illustrated in Figure 3. The third failure mode is that proper image segment cannot be retrieved from the memory. This is illustrated in Figure 4(the car).

References

[1] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017. 1

[2] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1

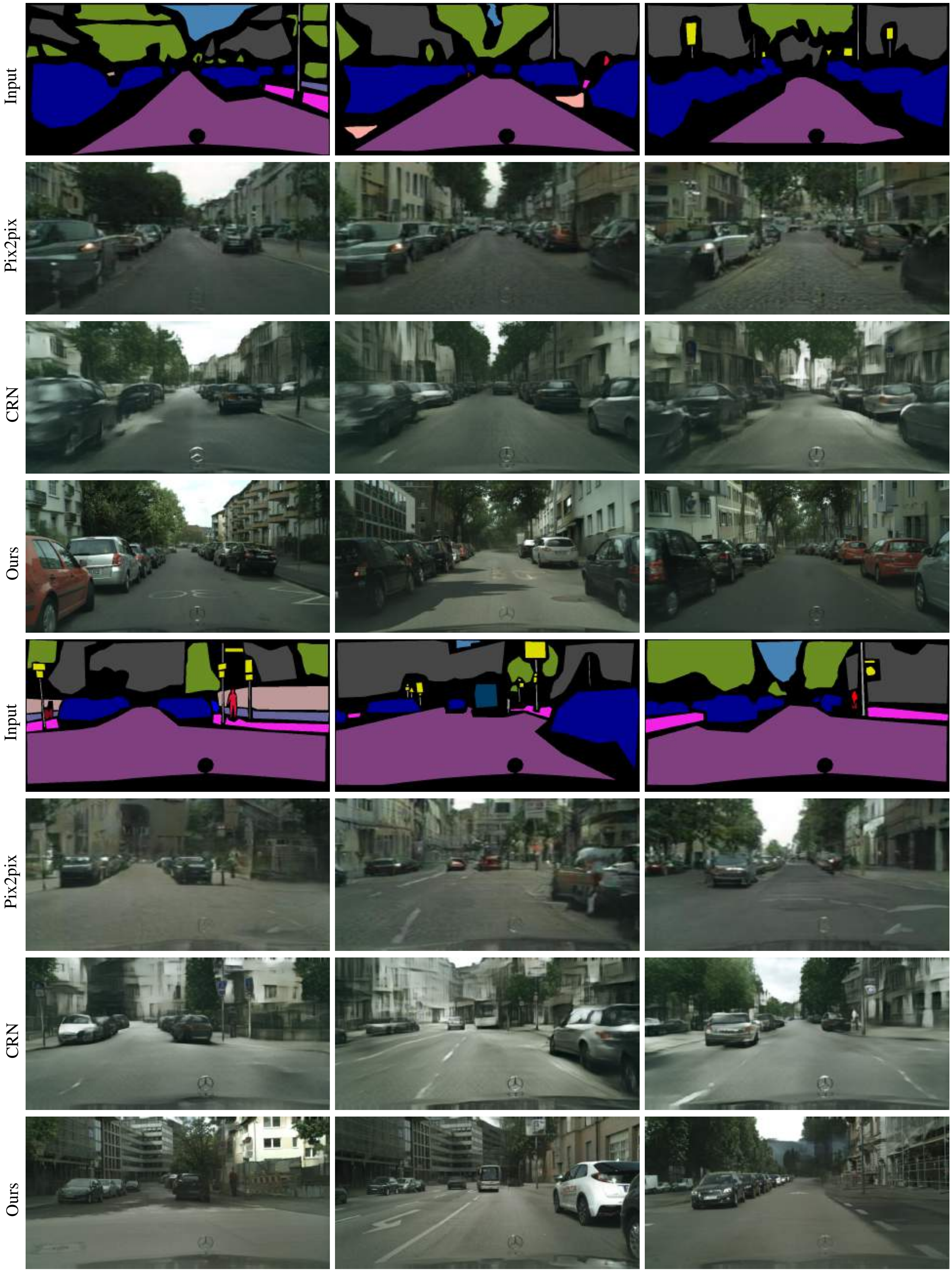


Figure 5. Comparison on Cityscapes-coarse.

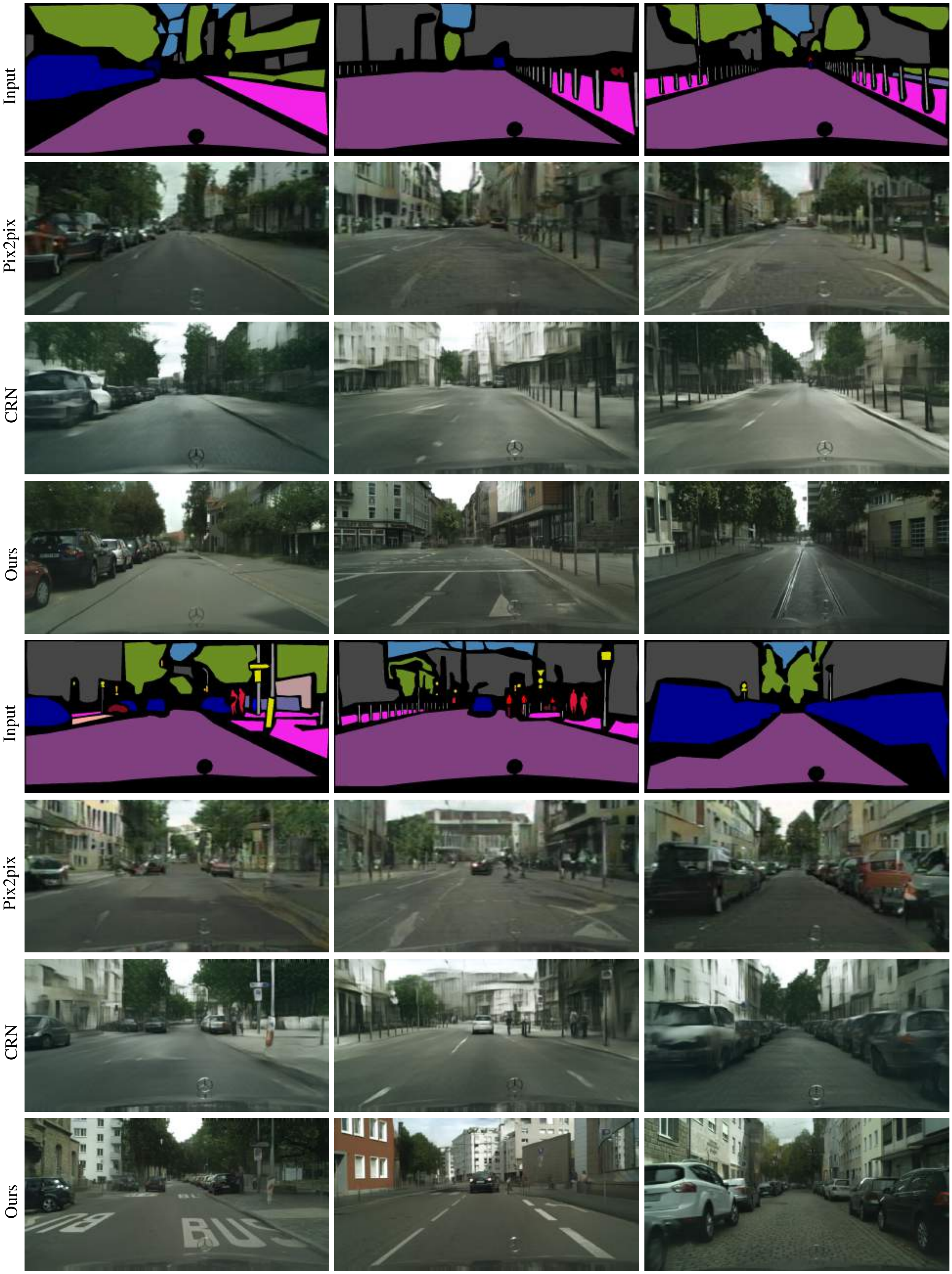


Figure 6. Comparison on Cityscapes-coarse (cont.).

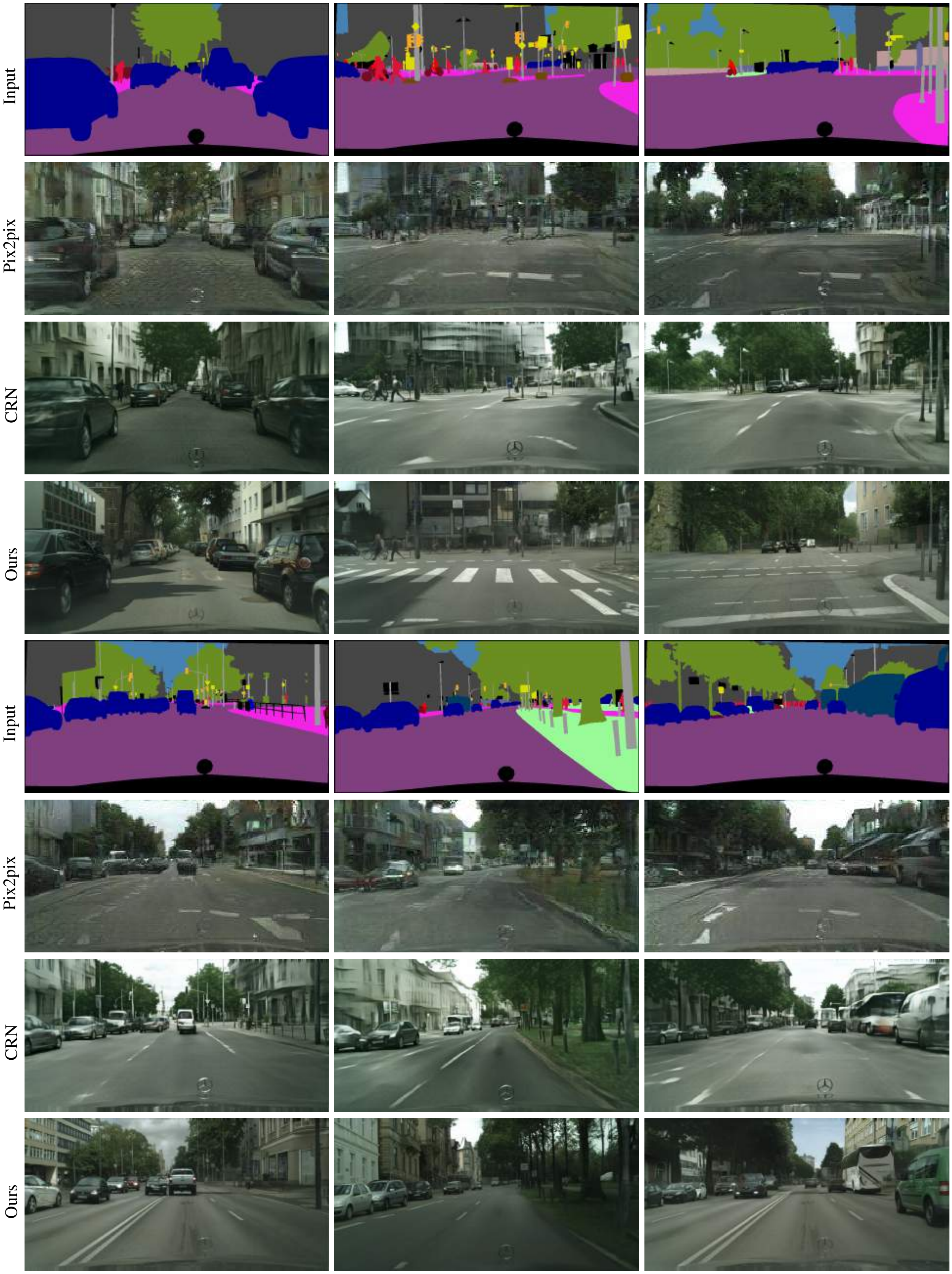


Figure 7. Comparison on Cityscapes-fine.

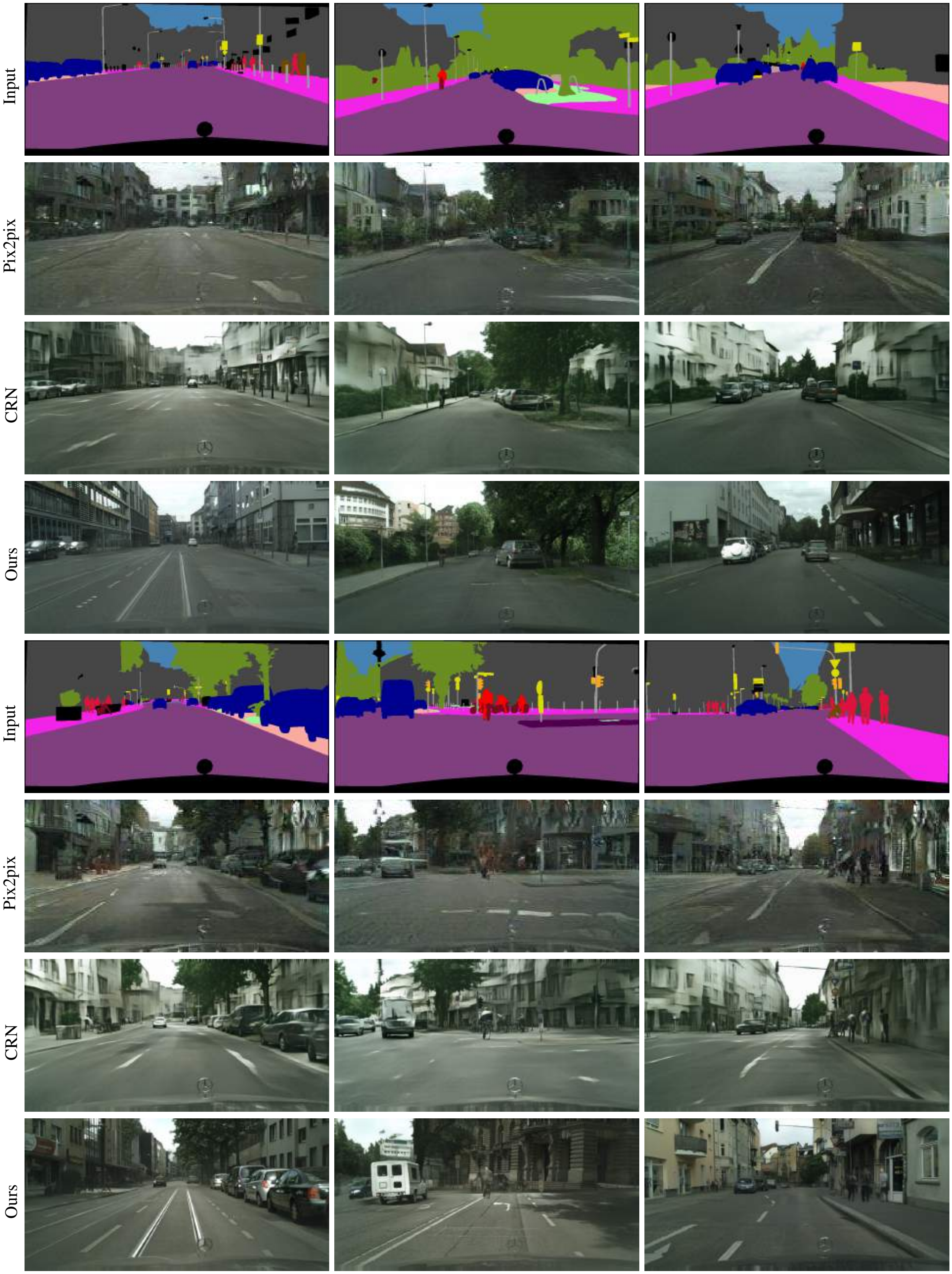


Figure 8. Comparison on Cityscapes-fine (cont.).

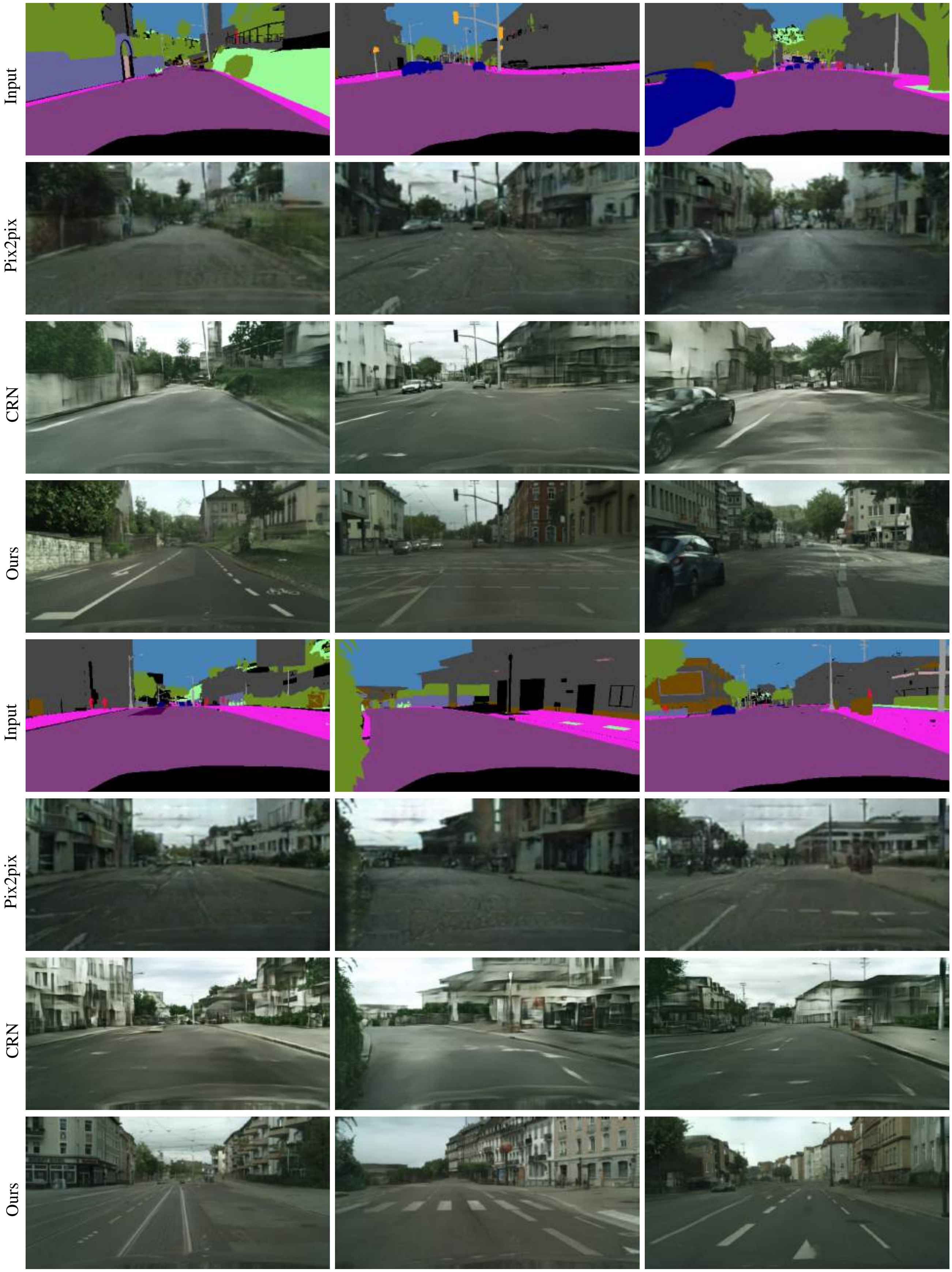


Figure 9. Comparison on Cityscapes→GTA5.

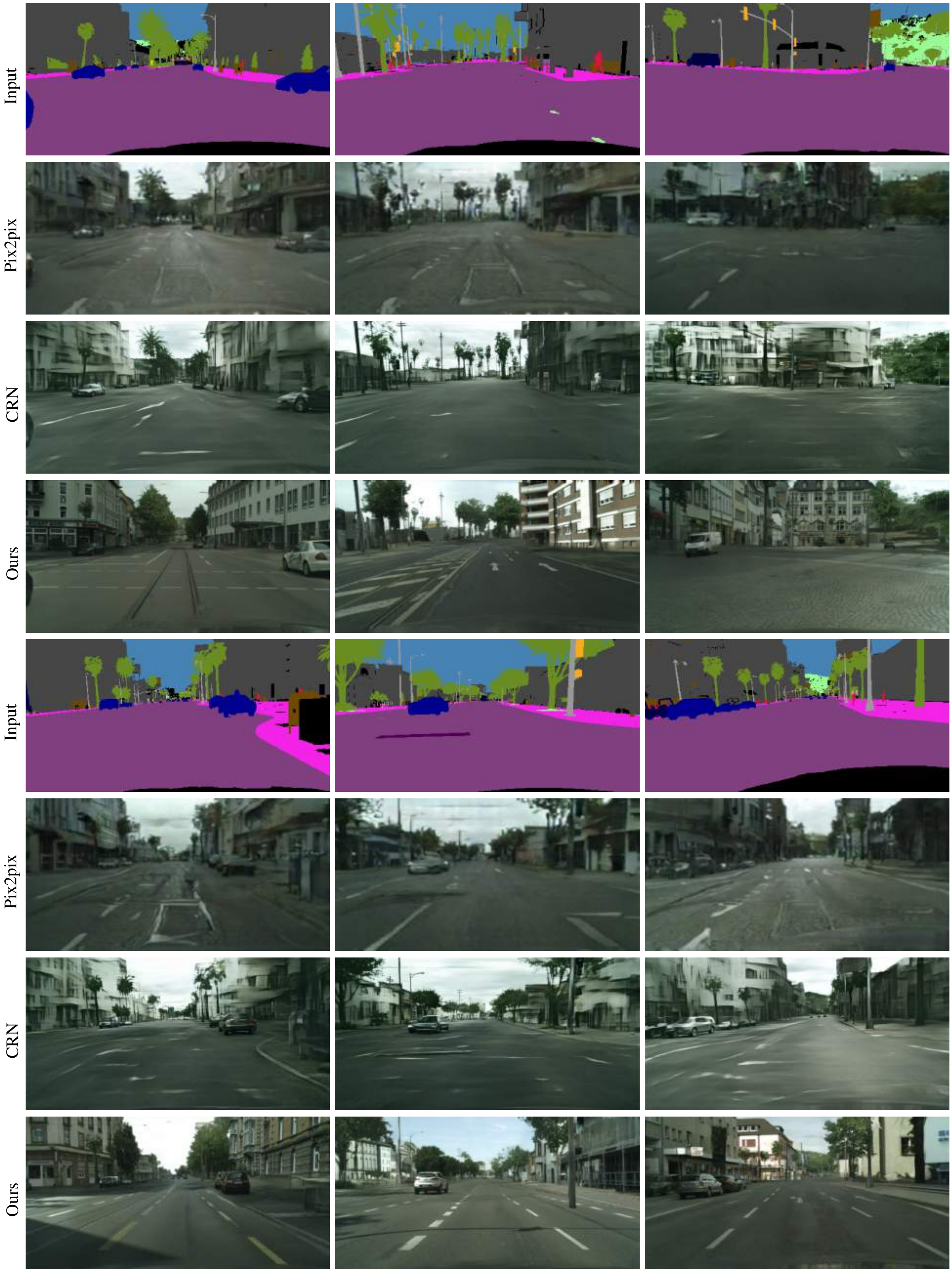


Figure 10. Comparison on Cityscapes→GTA5 (cont.).

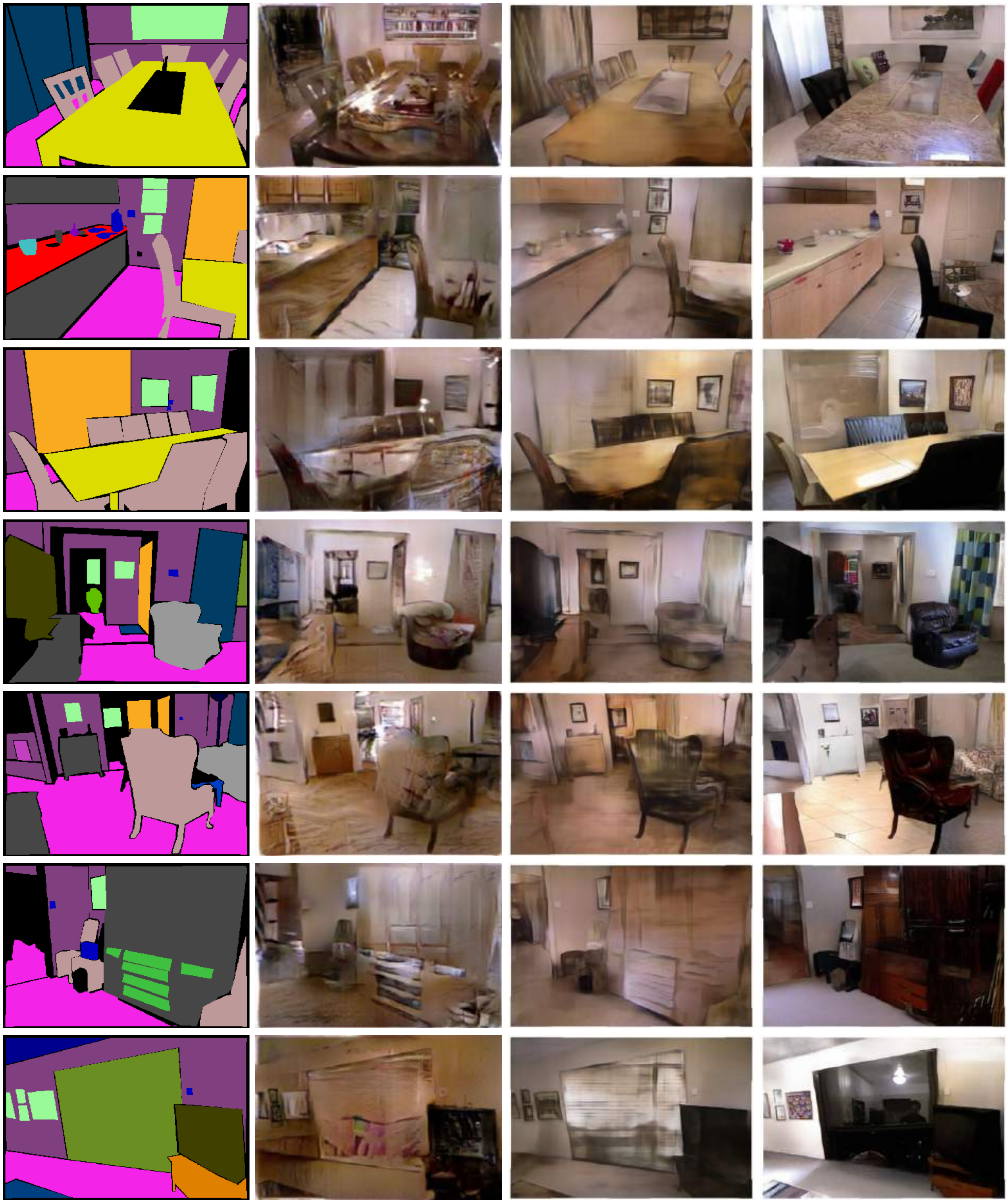


Figure 11. Comparison on the NYU dataset.

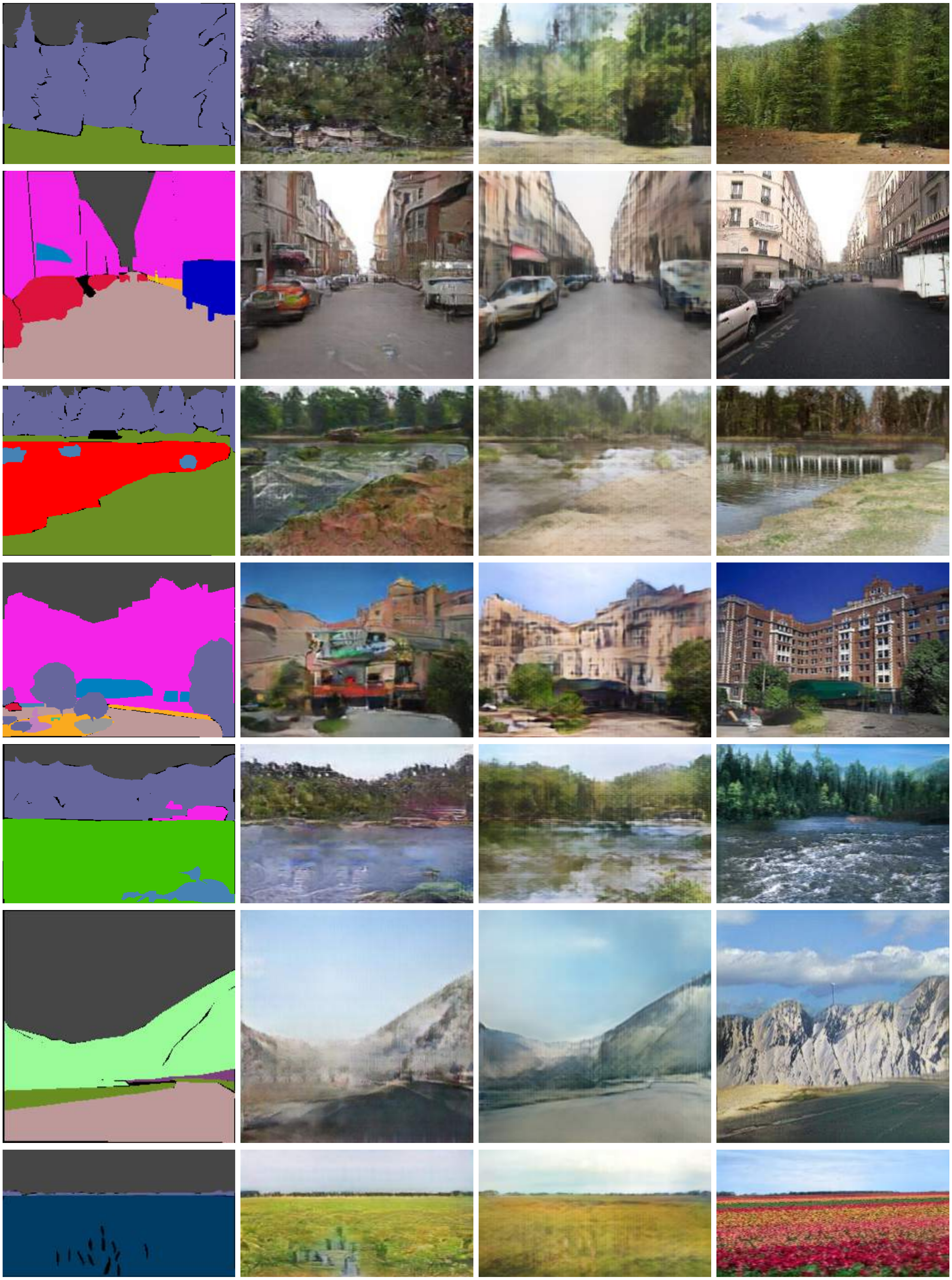


Figure 12. Comparison on the ADE20K dataset.

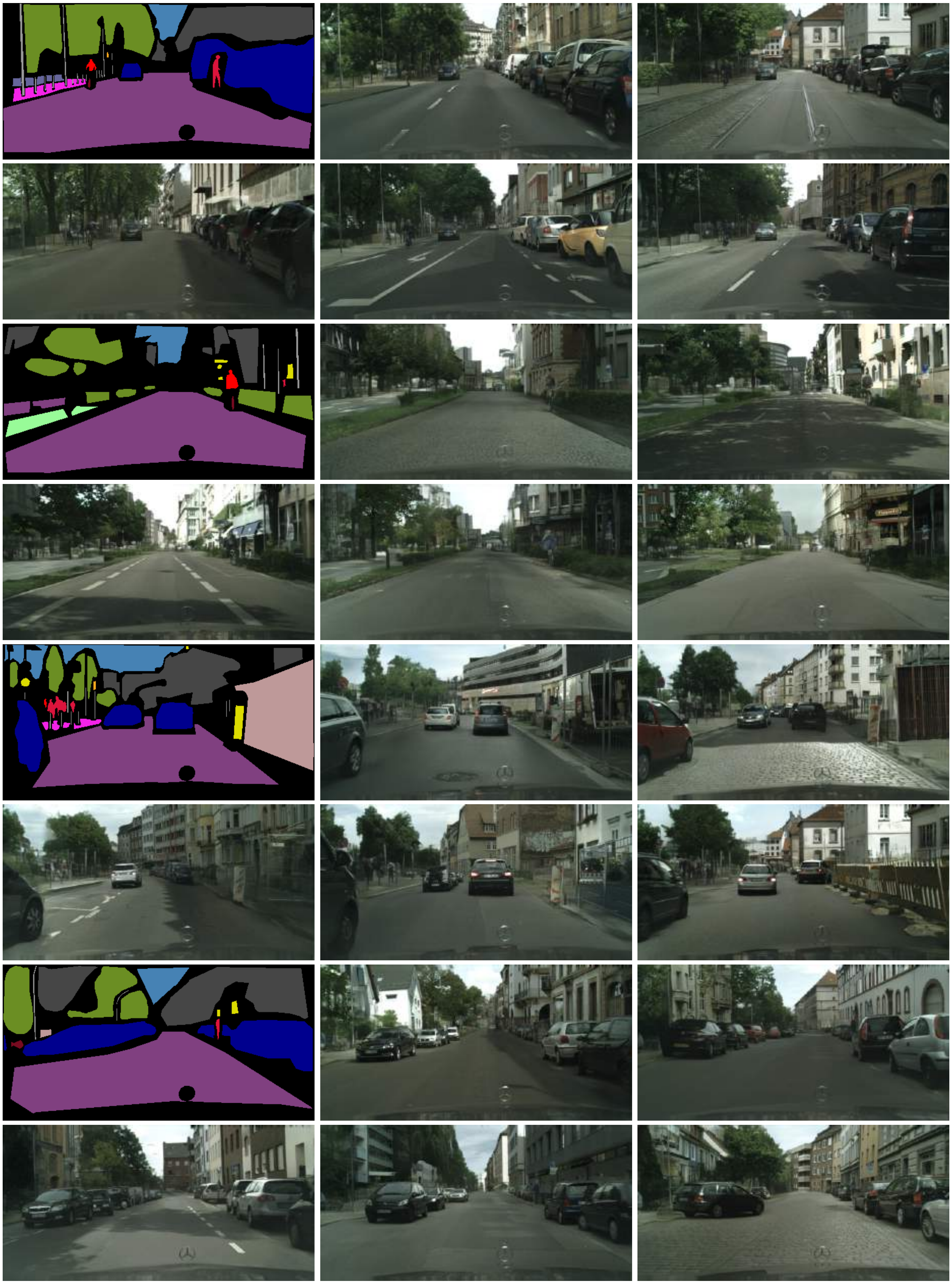


Figure 13. Diversified syntheses, illustrated on the Cityscapes dataset. Every six images form a group. The top-left image is the given semantic layout. The other 5 images are synthesized for this input layout by the presented approach.

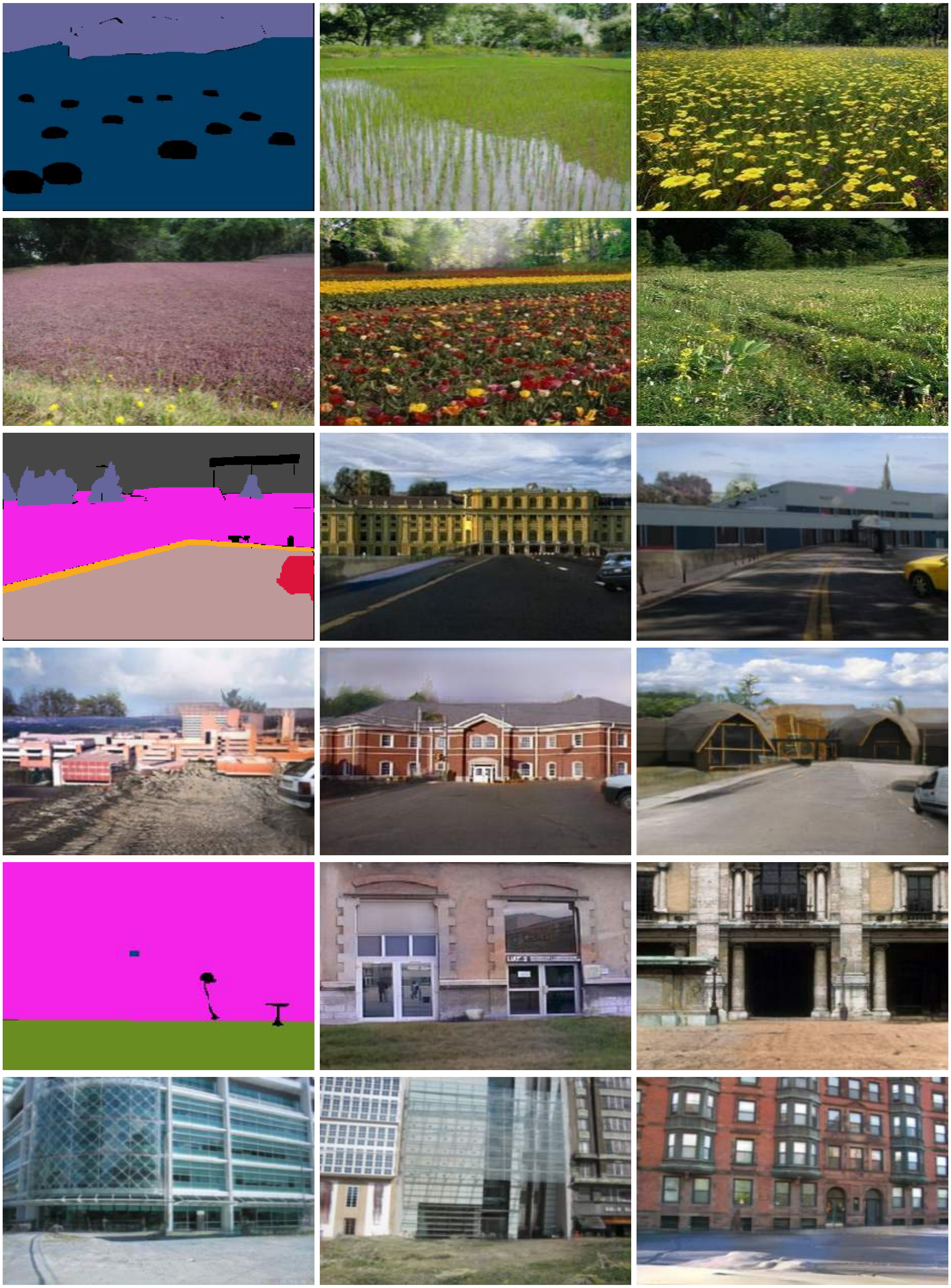


Figure 14. Diversified syntheses, illustrated on the ADE20K dataset. Every six images form a group. The top-left image is the given semantic layout. The other 5 images are synthesized for this input layout by the presented approach. Black color indicates invalid label.

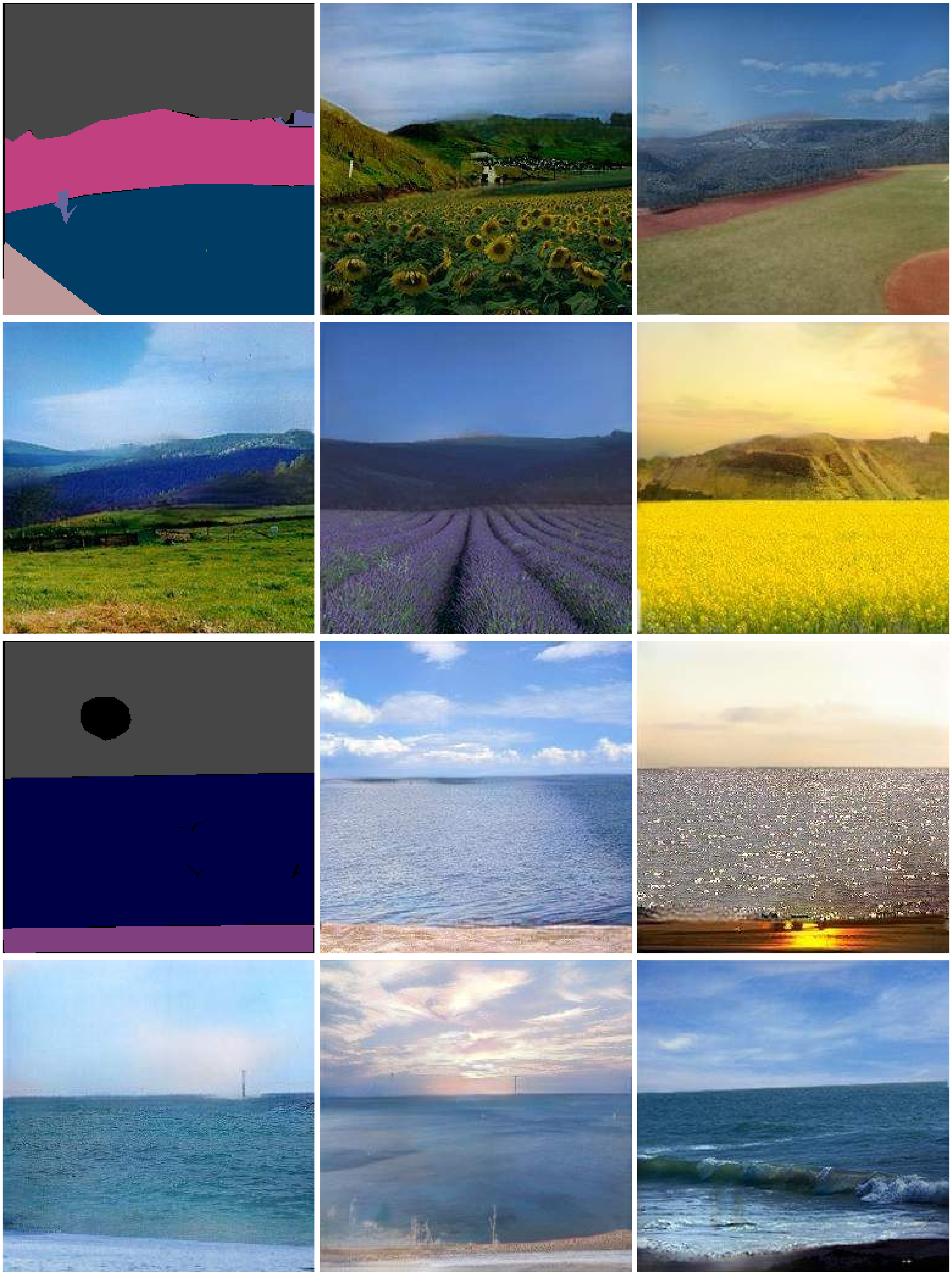
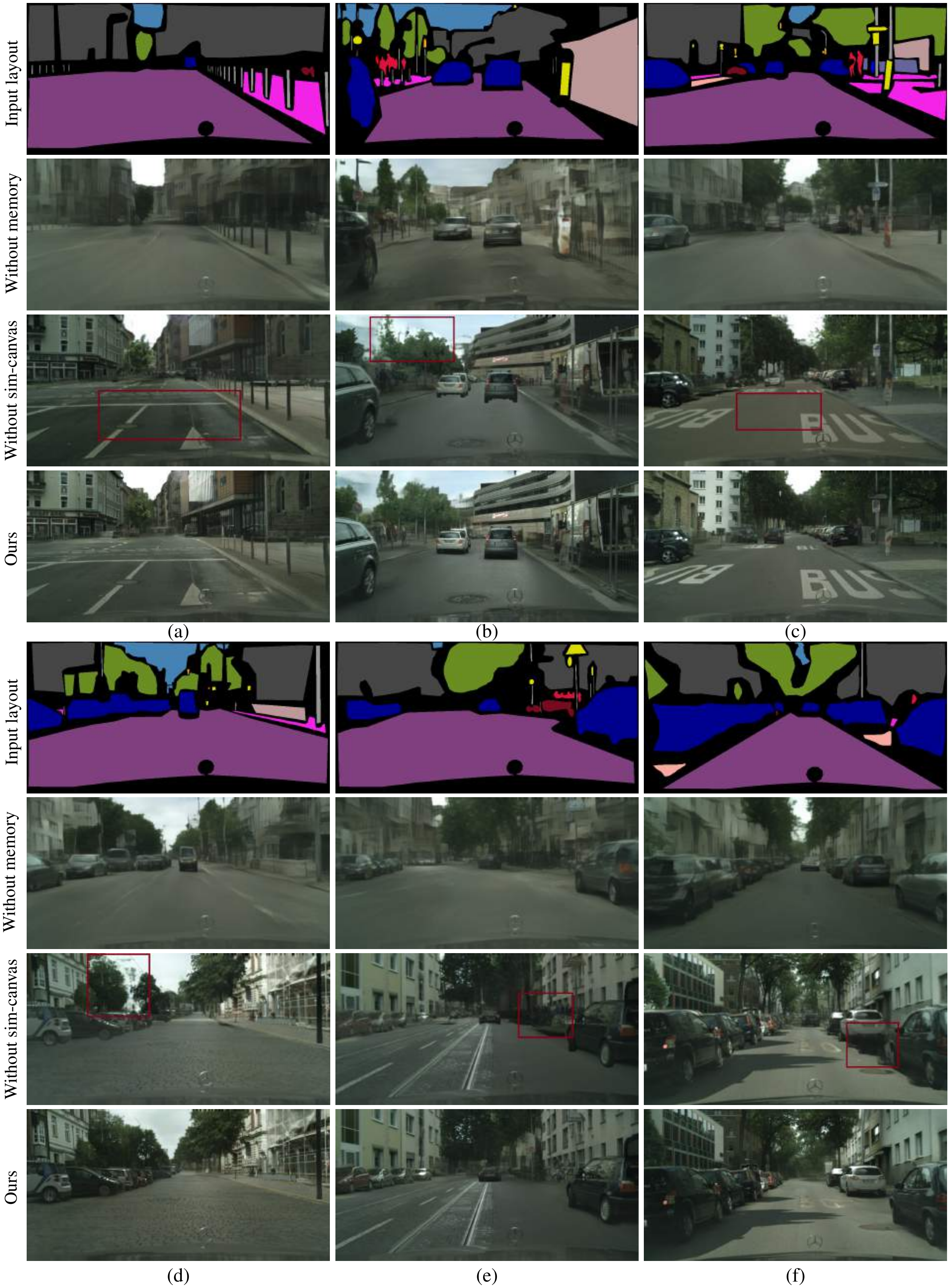


Figure 15. Diversified syntheses, illustrated on the ADE20K dataset. Every six images form a group. The top-left image is the given semantic layout. The other 5 images are synthesized for this input layout by the presented approach. Black color indicates invalid label.



(a) (b) (c)

(d) (e) (f)

Figure 16. Ablation study. Red boxes indicate the high-lighted regions.

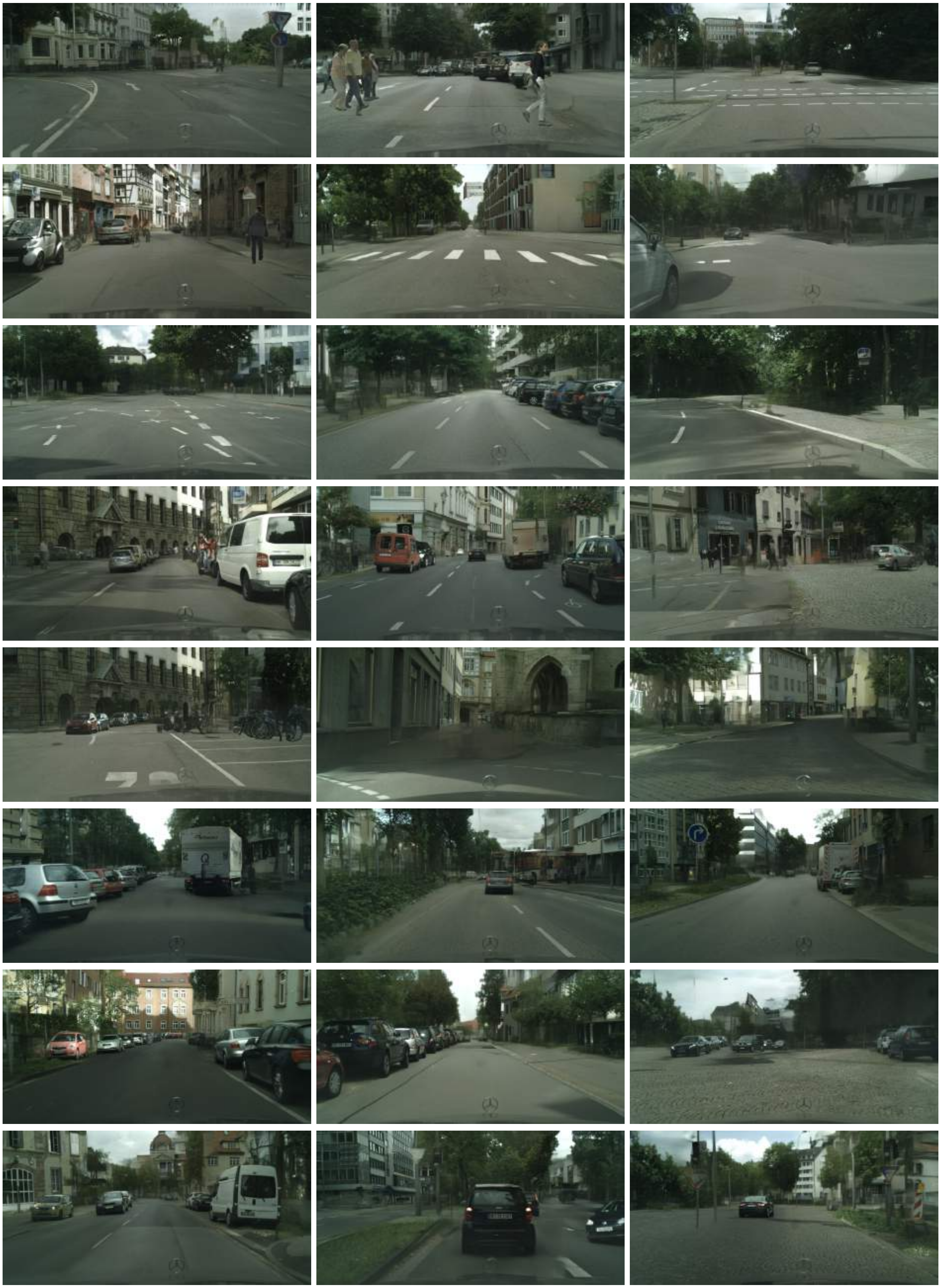


Figure 17. Randomly picked results on cityscape coarse set.