Photographic Image Synthesis with Cascaded Refinement Networks
Supplementary Material

Qifeng Chen[†,‡] Vladlen Koltun[†]

A. Training

We use Adam with a learning rate of 10^{-4} [4]. The weights are initialized as proposed by Glorot and Bengio [2]. The CRN is trained in stages. We first train a network to synthesize 256p images for 200 epochs. We then add one refinement module and fine-tune the full 512p model for 20 epochs. Then we add the final refinement module and fine-tune the full 1024p model for 5 epochs.

B. Experimental Procedure

In this appendix, we provide a more detailed description of each type of experiment performed using the Mechanical Turk platform. There are two types of experiments: unlimited-time pairwise comparisons and limited-time pairwise comparisons. These two types of experiments are also shown in videos that are enclosed with this supplement: `experiment-unlimited-time.mp4` and `experiment-limited-time.mp4`.

Unlimited-time comparisons. All models are trained on the training set, and are used to synthesize images for the respective test set (500 images for Cityscapes, 249 images for NYU). Images from these synthesized sets are sampled and paired at random. Pairwise comparisons are grouped into Human Intelligence Tasks (HITs) that are deployed on the Amazon Mechanical Turk (MTurk) platform. We construct four batches of HITs. Each batch is for comparisons between our model and a baseline. In each batch, we have 5 HITs each containing 110 comparisons. In each HIT, there are 100 "ours vs baseline" comparisons. The other 10 comparisons are sentinel comparisons where a real Cityscapes image is compared to a synthesized image. For each comparison, two images are displayed side-by-side and the user is asked to pick the most realistic among the two. The left-right order is randomized, as is the order of pairs within a HIT, as is the grouping of pairs into HITs.

For each HIT, the worker is given 15 minutes to complete it. Each HIT is completed by 10 distinct workers. For each batch, the HITs are completed by 50 unique workers in total. If a worker gives an incorrect answer on two or more out of the 10 sentinel comparisons, the HIT is discarded. (2% of the HITs were discarded.) The compensation for one completed HIT is $1.

Please see `experiment-unlimited-time.mp4`.

Limited-time comparisons. These are fine-grained experiments that evaluate the relative realism of images synthesized by our approach, images synthesized by the approach of Isola et al. [3], and real Cityscapes images. The results are reported in Figure 4 in the paper.

The images from the different conditions are paired at random. Each pair is shown for a timespan that is chosen at random from the following: $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, 4, and 8 seconds. Since we have three types of pairwise comparisons ("CRN vs Cityscapes", "Pix2pix vs Cityscapes", and "CRN vs Pix2pix"), there are $3 \times 7 = 21$ modes in this experiment.

The pairs are grouped into 10 HITs. Each HIT contains 5 comparisons from each of the 21 modes, for a total of 105 genuine comparisons. In addition, 5 sentinel pairs are inserted at random. The total number of pairs in a HIT is thus 110. The left-right order for each pair is randomized, as is the order of pairs within a HIT, as is the assignment of pairs to HITs. Each HIT is duplicated 20 times and performed by 20 unique workers. In total, the time-limited HITs are performed by 200 unique workers. If one or more of the 5 sentinel pairs was ranked incorrectly, the HIT is discarded. (2% of the HITs were discarded.) The compensation for one HIT of this type is $2.

Please see `experiment-limited-time.mp4`.

C. Videos

For additional illustration of the performance of different models, we have synthesized videos using the GTA5 dataset, which provides label maps for images from the game Grand Theft Auto V [5]. The set of labels used in this dataset is consistent with Cityscapes, thus models trained on Cityscapes can be applied out of the box to synthesize images for GTA5 label maps. We follow the standard 12K/6K train/val split for GTA5. The images in the validation set are in sequence and can be used to synthesize a continuous video. We use this dataset to synthesize four videos, which are enclosed:

[†]Intel Labs
[‡]Stanford University

- `Cityscapes-to-GTA5-Ours.mp4`: These are images synthesized by our model (CRN), trained on the Cityscapes dataset. This is the same model that is evaluated throughout the paper, applied to label maps from the GTA5 dataset. Each frame is synthesized separately, but the results are nevertheless quite coherent.
- `Cityscapes-to-GTA5-Isola.mp4`: These are images synthesized by the model of Isola et al. [3], trained on the Cityscapes dataset.
- `GTA5.mp4`: In addition, we trained a CRN from scratch on the GTA5 training set. We use exactly the same architecture as for the Cityscapes dataset. The video shows the images synthesized by this model on the GTA5 validation set.
- `GTA5-diverse.mp4`: We also trained a CRN on the GTA5 training set using the diversity loss presented in Section 3.4 of the paper (Equation 3). We use $k = 9$. The video shows the 9 synthesized images, arranged in a 3×3 grid. As can be seen in the video, the different output channels specialize in different appearance: the hood of the driven car is consistently red, black, white, or grey; other vehicles have different colors; the sky has a different tint, etc.

D. Additional Qualitative Results

Figure 1 (3 pages) shows additional qualitative results on the Cityscapes dataset. This extends Figure 5 in the paper.

Figure 2 shows additional qualitative results on the NYU dataset. This extends Figure 6 in the paper.

Figure 3 shows collections of diverse images synthesized for different label maps on the NYU dataset. Recall that $k = 9$ images are synthesized for each label map, by a model that is trained using the loss given in Equation 3 in the paper. Figure 3 shows all 9 synthesized images for a number of label maps.

E. Coarse Input Layouts

As a stress test, we have also trained a model to synthesize images given coarse and incomplete input layouts. To this end, we use the set of coarsely labeled Cityscapes images [1]. We train on coarsely labeled images from the training set, and test on coarsely labeled images from the validation set. The results are shown in Figure 4.

References

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2

[2] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010. 1

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

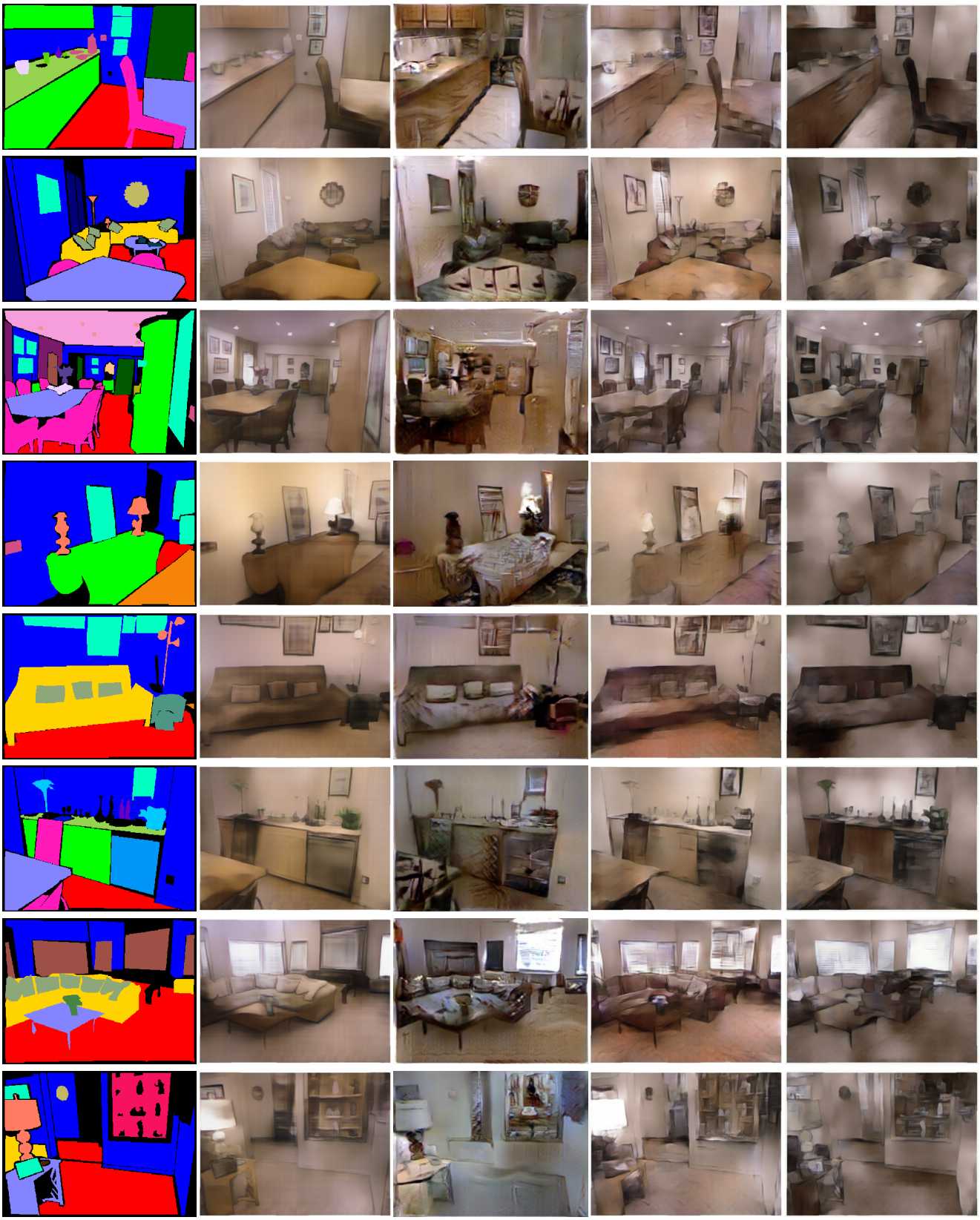
[4] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[5] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 1

Figure 1: Qualitative comparison on the Cityscapes dataset.

Figure 1 (cont.): Qualitative comparison on the Cityscapes dataset.

Figure 1 (cont.): Qualitative comparison on the Cityscapes dataset.



Semantic layout | Our result | Isola et al. | Full-resolution network | Encoder-decoder

Figure 2: Qualitative comparison on the NYU dataset.

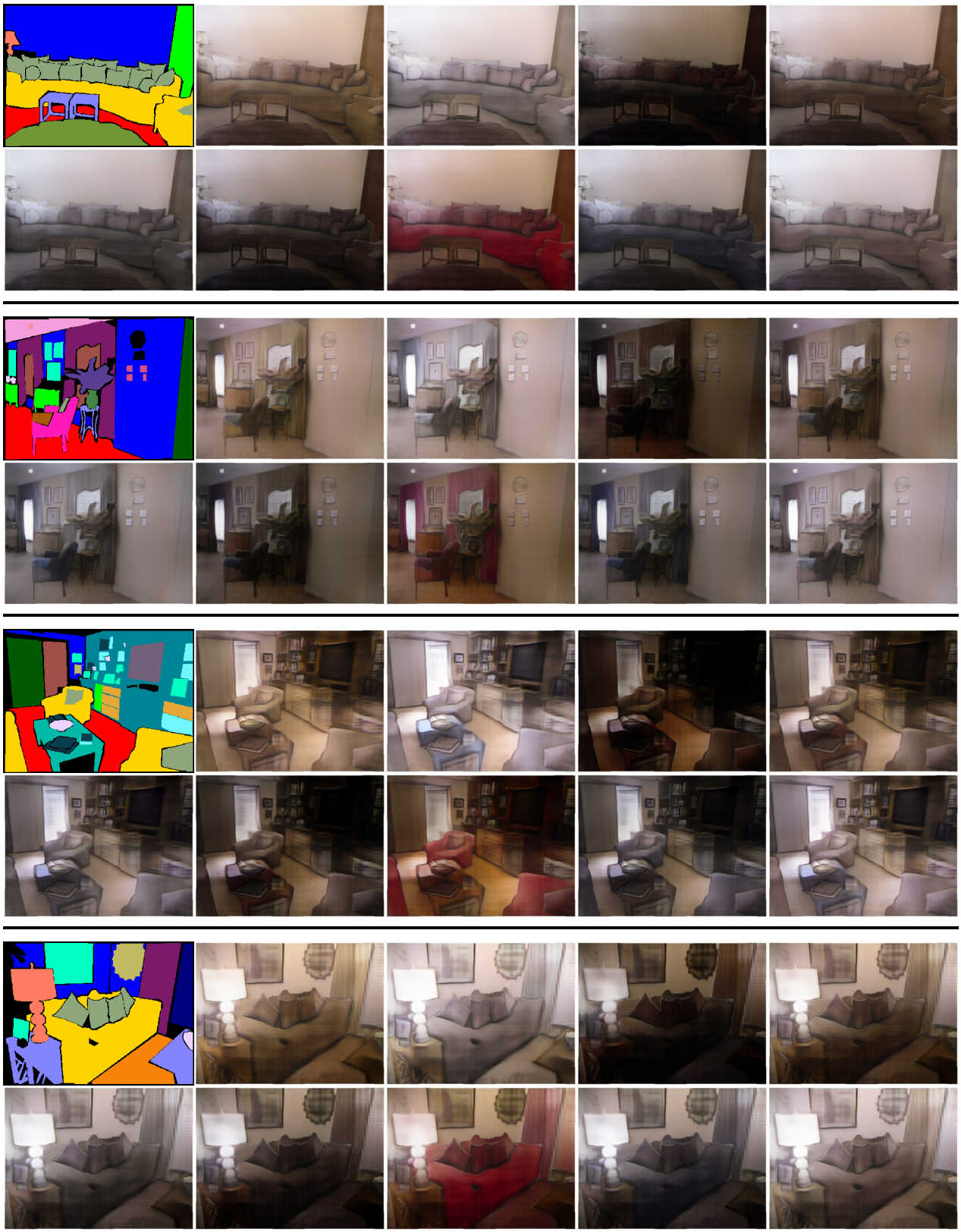
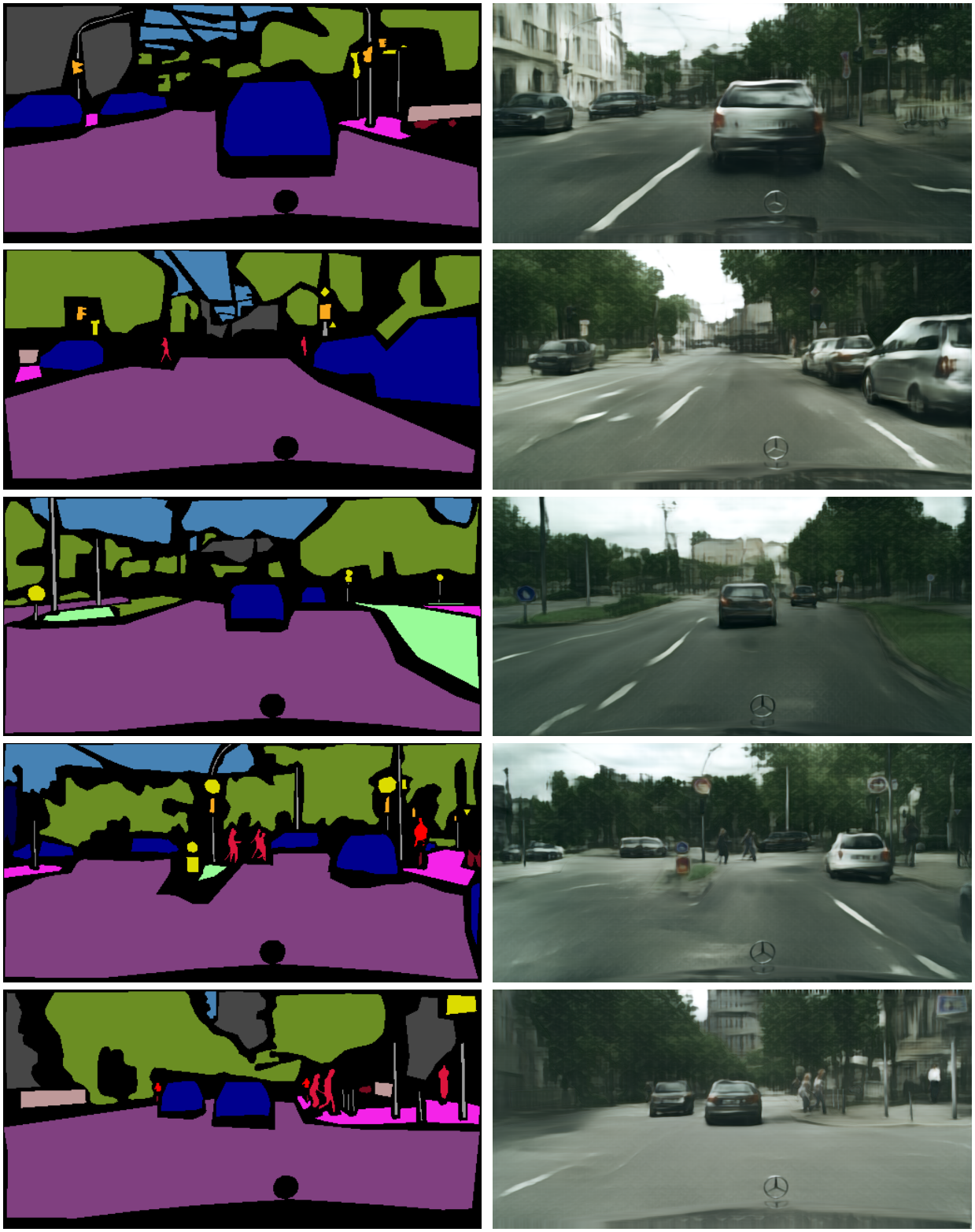


Figure 3: Synthesizing a diverse collection, illustrated on the NYU dataset. This figure shows four groups of images. Each group contains an input label map and $k = 9$ images synthesized by a model trained with the diversity loss.



Input semantic layout Our result

Figure 4: The presented approach can be used to synthesize images given coarse and incomplete input layouts.