Tell Me Where I Am: Object-level Scene Context Prediction Supplementary Material

Xiaotian Qiao Quanlong Zheng Ying Cao Rynson W.H. Lau
City University of Hong Kong

In this supplementary material, we provide the details of our network, the details of our user studies, and additional qualitative results of scene layout generation.

1. Details of Our Network

Figure 1 shows the module details of the proposed network architecture.

Encoder. The input to the encoder is an object layout $X_o \in \{0,1\}^{128\times128\times72}$, which is fed into five 3×3 stride-2 convolutions. All the convolutional layers are followed by batch normalization [2] and Leaky-ReLU [4], except for the first and last layers where only Leaky-ReLU is applied. The output size of the encoder is $4 \times 4 \times 512$.

Shape Generator. The input to the shape generator is a feature map obtained by concatenating the encoder output with a noise vector. The shape generator has two 4×4 stride-2 deconvolutions followed by batch normalization and ReLU, and a 1×1 convolution followed by sigmoid nonlinearities. The output size of the shape generator is $16 \times 16 \times 72$.

Region Generator. The input to the region generator is the same as that of the shape generator. The region generator has two 4×4 stride-2 deconvolutions, several residual blocks and 3×3 convolutions, and two fully connected layers. Each residual block consists of a 1×1 convolution, a 3×3 convolution and a skip connection. The last fully connected layer has 1080 units that are reshaped to $3 \times 5 \times 72$, which encodes the parameters of the bounding boxes of all the 72 categories (i.e., 5 parameters per bounding box $\times$ 3 bounding boxes per category).

Compositor. The outputs from the region generator and the shape generator are warped to form a coarse scene layout, which is refined in the compositor. The compositor contains two 3×3 stride-2 convolutions, and two cascaded refinement modules [1]. Each cascaded refinement module takes as input the feature map from a previous layer and a rescaled coarse scene layout. It is composed of a bilinear upsampling and two 3×3 convolutions. All the convolutional layers in the compositor are followed by batch normalization and Leaky-ReLU. The final output of the compositor is the refined scene layout of size $128 \times 128 \times 72$.

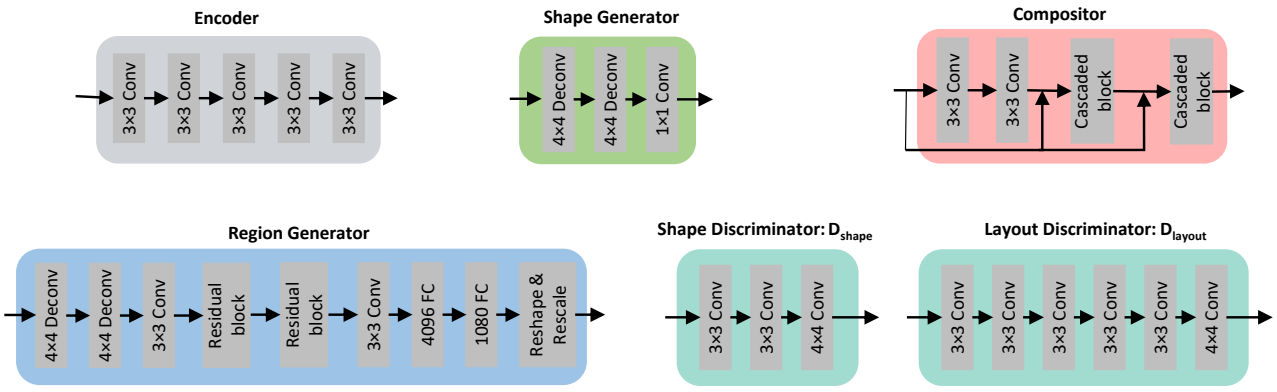

Figure 1. Details of each module in our architecture.

Shape Discriminator and Layout Discriminator. The input to the shape discriminator is the generated shape or the real shape of size $16 \times 16 \times 72$. The shape discriminator first processes the input with a series of 3×3 stride-2 convolutions. All the convolutional layers are followed by batch normalization and Leaky-ReLU, except for the first layer where only Leaky-ReLU is applied. Then, a 4×4 convolution and a fully connected layer are added to output the probability of the input being real or fake. Both layers are followed by Leaky-ReLU. The layout discriminator shares a similar architecture with the shape discriminator, but has a different number of convolutional layers.

2. User Studies

We describe how we conducted the user studies on Amazon Mechanical Turk, as discussed in Section 4.4 in the main paper. We asses the quality of the generated scene layouts by conducting two user studies: (a) plausibility evaluation and (b) fitness evaluation. In (a), our goal is to evaluate whether the objects in generated scene layouts have plausible spatial relations. In (b), we aim to evaluate whether the generated scene layouts provide convincing context for the respective input object(s). Details of the experiments are described below.

(a) Plausibility evaluation: Figure 2 is an example of our MTurk experiment for plausibility evaluation. AMT workers were given a sequence of scene layouts selected randomly from three sources (Ours, Baseline and GT), and asked to evaluate whether the objects in the scene layouts have plausible spatial relations. For the scene layouts judged to be implausible, they were asked to label at least one pair of objects that are in incorrect spatial relation. Each Human Intelligence Task (HIT) contains 50 scene layouts, along with 10 duplicate layouts for consistency check. We discarded the responses from the worker who has less than 80% consistency on the duplicate questions. We end up with 30 workers in our experiments and have each scene layout evaluated by 10 workers.

(b) Fitness evaluation: Figure 3 is an example of our MTurk experiment for fitness evaluation. The AMT workers were presented with an input object layout, along with two scene layouts generated from the input, and asked to select which scene layout illustrates a better context for the input objects. In each comparison, we displayed two scene layouts chosen randomly from three sources (Ours, Baseline and GT) side by side in randomized order. We used 150 pairwise comparisons (50 for Ours vs. Baseline, 50 for Ours vs. GT and 50 for Baseline vs. GT). We randomly divided the 150 comparisons into three HITs uniformly. In each HIT, we added 5 duplicate comparisons for consistency check. We discarded the responses from the worker who has less than 80% consistency on the duplicate comparisons. We have a total of 9 workers in the experiment, and each comparison was evaluated by 3 workers.

3. Qualitative Results of Scene Layout Generation

Figure 4 shows more qualitative results of our method, compared with a baseline method [3]. Overall, our method can predict plausible scene layouts that fit the input object(s) well given large variations of input object(s) in terms of category, shape and position..

Figure 5 shows more results of our method and the baseline when varying the categories and shapes of input objects and the spatial relation between input objects. Note that, when inputs are changed, our method can generate the scene contexts that better adapt to the inputs. For instance, when changing an input object from car to elephant (the third column to the fourth column at the top part of Figure 5), our predicted context change the region supporting the input object from road to grass. Moreover, when changing the spatial relation between a person and a surfboard from $\langle person, above, surfboard \rangle$ to $\langle person, right, surfboard \rangle$ (the first column to the second column at the bottom part of Figure 5), the region below the person changes from sea to sand accordingly.

Figure 6 visualizes the object bounding boxes predicted by our model. As can be seen, our predictions are able to recall most of the important object regions, even though they are not exactly the same as the ground truth (due to the multi-modal nature of scene context prediction).

References

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

[2] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 1

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

[4] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013. 1

Figure 2. An example of the plausibility evaluation in our user study.

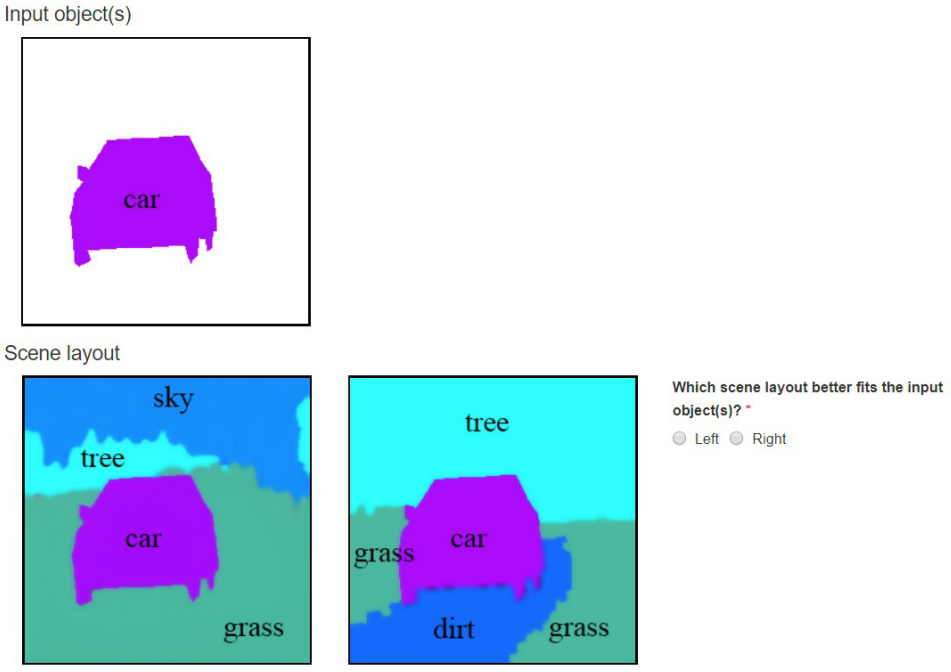

Figure 3. An example of the fitness evaluation in our user study.

Figure 4. Qualitative results from our model and the baseline. Given the input object layouts (left diagrams in each column), which contains one or two standalone objects, we generate the output scene layouts using our model (middle diagrams in each column) and the baseline (right diagrams in each column).

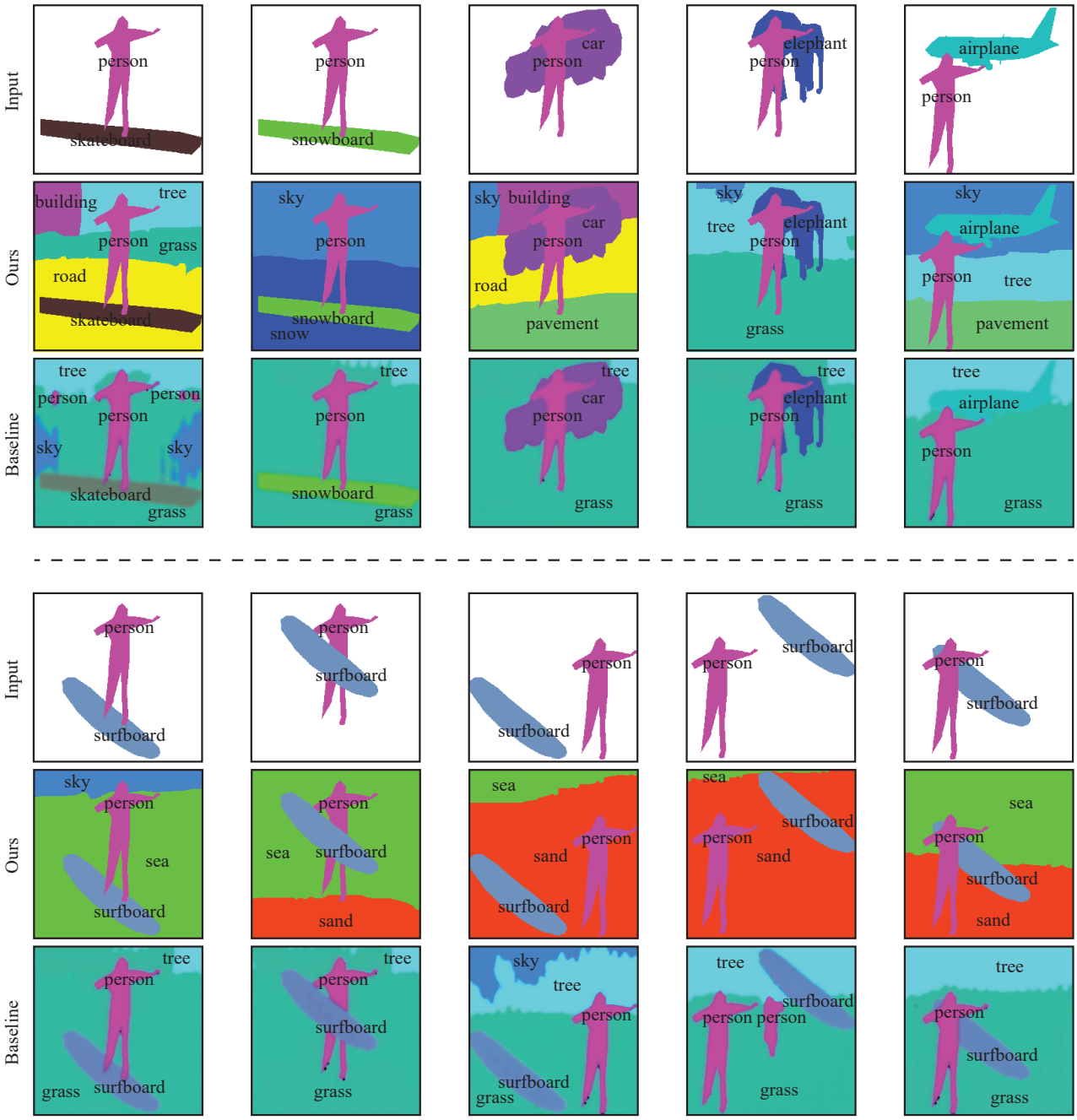

Figure 5. Qualitative results of our model and the baseline by varying the category and shape of the input objects (top) and the spatial relation between the input objects (bottom).

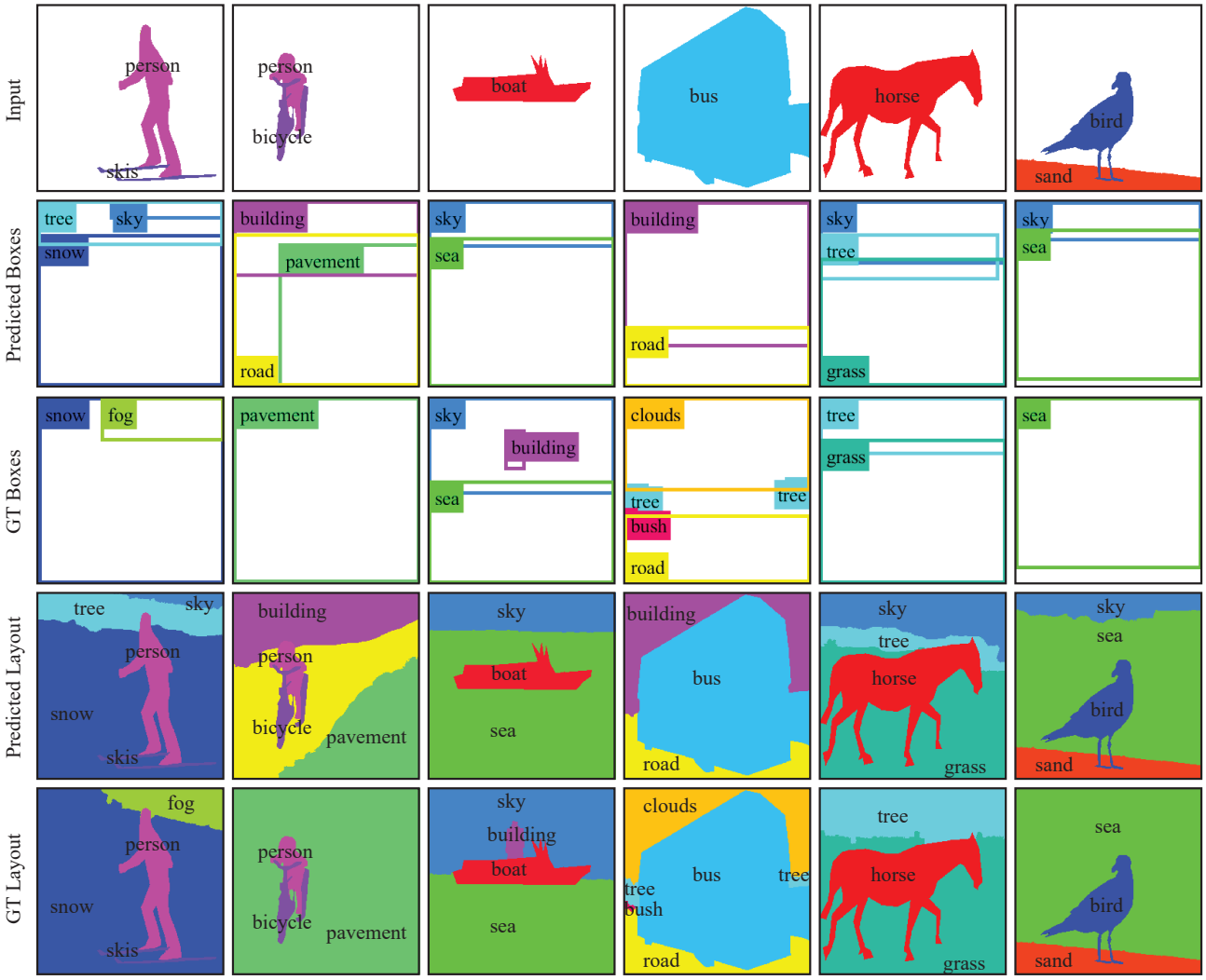

Figure 6. Object bounding boxes predicted by our model. For each input object layout (first row), we show our predicted object bounding boxes (second row) and the ground truth bounding boxes (third row). We also compare our scene layout prediction (fourth row) and the ground truth layout (fifth row).