Learning Object Context for Novel-view Scene Layout Generation Supplementary Material

Anonymous CVPR submission

Paper ID 2722

In this supplementary material, we first provide the details of our model in Section 1. We then show more qualitative comparison results on novel-view scene layout generation in Section 2.

1. Model Details

Given a single scene layout map at the source view as well as the camera pose transformation as inputs, our model can generate a scene layout for the target view. We first extract the initial object representation from the inputs as described in the paper, and then use the following three key modules to output the novel-view scene layout.

OCT Module. Given the initial object representation and the camera transformation vector as inputs, we use the OCT module to produce the contextualized object representation for each object. Our OCT module uses the multi-head attention, which computes the view-aware attention blocks (VAB) 4 times in parallel. The details of the VAB architecture are illustrated in the paper. Note that the parameters of the four VABs are not shared. The VAB splits its Query, Key, and Value parameters and passes each split independently through a separate Head. All of these Attention calculations are then combined together to produce a final Attention score. We apply the attention score on the initial object representation and add a residual connection to obtain the contextualized object representation as the output.

OLG Module. The OLG module takes the reshaped contextualized object representation as input to predict an object layout for each object. The OLG module has an object bounding box branch and an object mask branch. The object bounding box branch aims to predict the bounding box of the object at the target view from the corresponding contextualized object representation. It has two 3×3 stride- 2 deconvolutions, two 3×3 convolutions, two residual blocks, and two fully connected layers. Each residual block consists of a 1×1 convolution, a 3×3 convolution and a skip connection. The object mask branch aims to predict the shape of the object at the target view from the corresponding contextualized object representation. It has three 4×4 stride-2 de-convolutions, two 3×3 convolutions with batch normalization and ReLU, and a 1×1 convolution followed by Sigmoid nonlinearity.

OLC Module. Given the set of predicted object layouts, the OLC module generates a scene layout as the final output of our model. Our OLC module contains two sub-modules. One is the OrderNet, and the other is the refinement block. The architecture of the OrderNet is based on VGG-19 [4]. The inputs to the ordering network are two object layouts. The output is a binary label that indicates if an object is on top of another when they are composed together. We then use four refinement blocks in total to further remove the noise in the scene layout. Each refinement block takes as input the coarse scene layout to generate a refined one. It contains a bilinear upsampling layer and two 3×3 convolutions with batch normalization and Leaky-ReLU. We use the refined scene layout from the last refinement block as the final output of our model.

2. More Qualitative Results

Figure 1 and Figure 2 present more qualitative comparisons on indoor and outdoor scenes, respectively. Each column shows an input scene layout followed by four generated scene layouts for the target view. In particular, we compare our novel-view scene layout generation results with those from three baselines, *i.e.*, UNet [2], LayoutGAN [3], and GVSNet [1]. We can see that our model is able to generate more geometrically and semantically consistent novel-view scene layouts than those from the baselines.

References

[1] Tewodros Habtegebrial, Varun Jampani, Orazio Gallo, and Didier Stricker. Generative view synthesis: From single-view semantics to novel-view images. In *NeurIPS*, 2020. 1

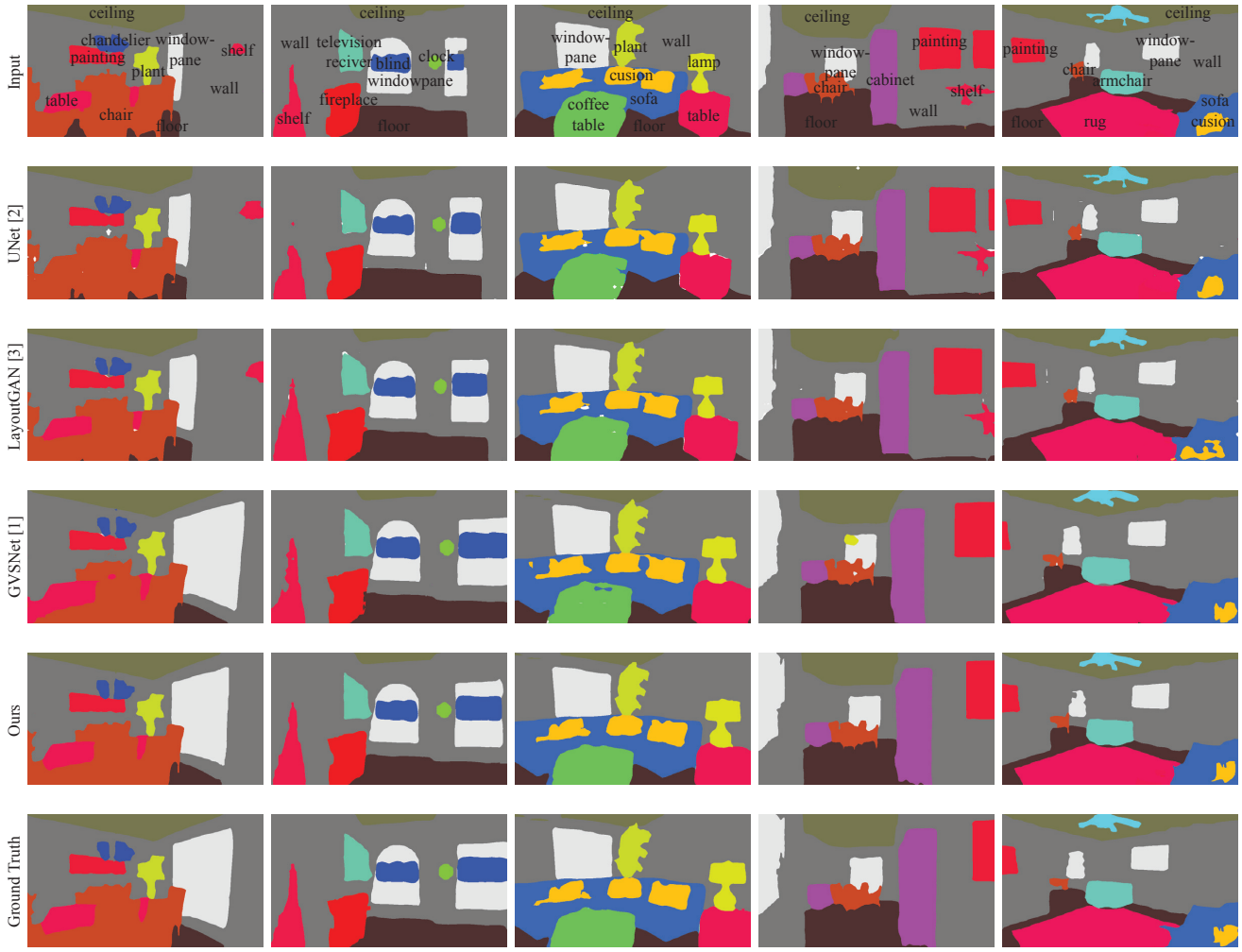

Figure 1. Visual comparison of our method against the baselines on indoor scenes. Given the input scene layout at the source view (left), we generate the output scene layout at the target view using our model and the baselines.

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1

[3] Jianan Li, Jimei Yang, Aaron Hertzmann, Jianming Zhang, and Tingfa Xu. Layoutgan: Synthesizing graphic layouts with vector-wireframe adversarial networks. *IEEE TPAMI*, 2020. 1

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 1

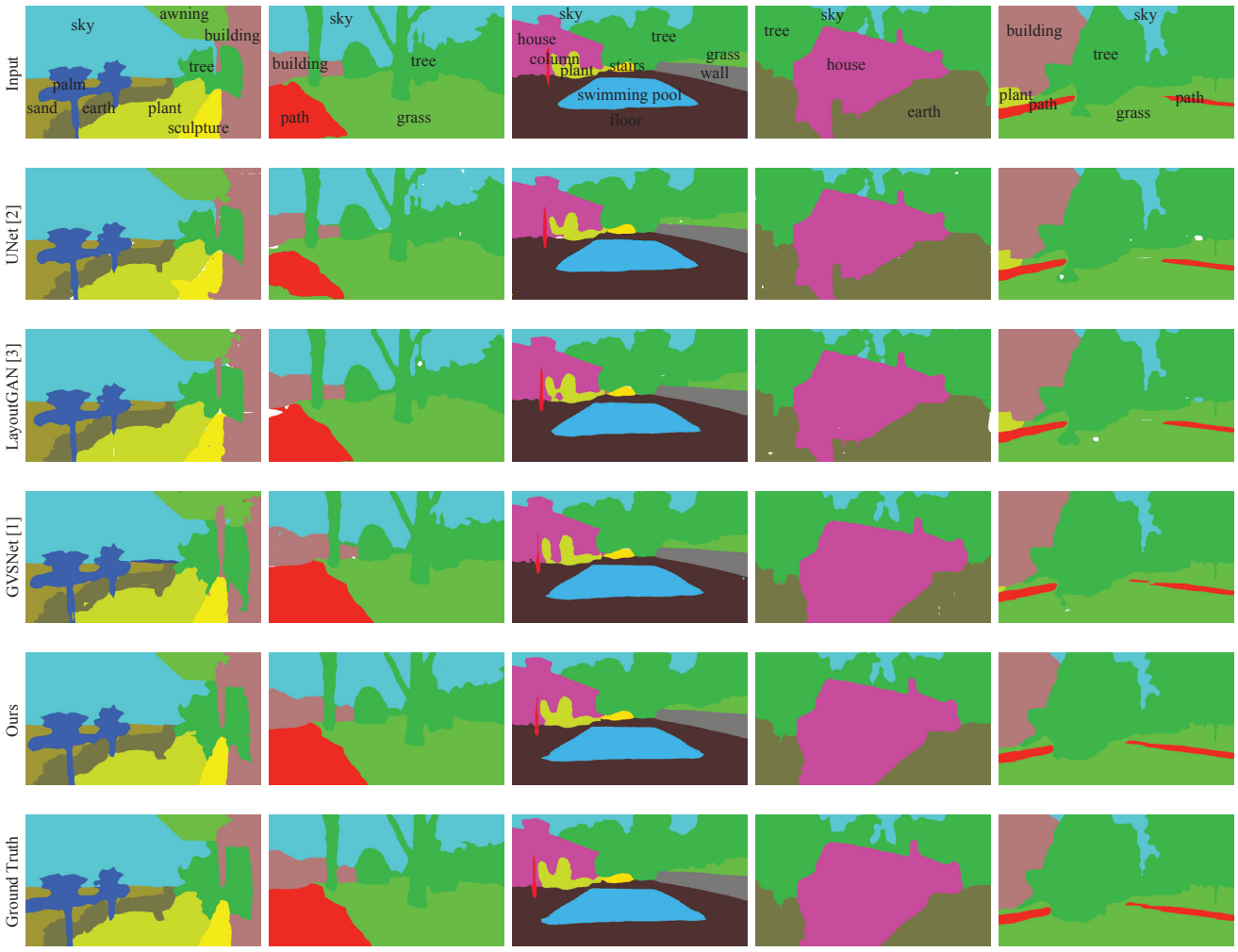

Figure 2. Visual comparison of our method against the baselines on outdoor scenes. Given the input scene layout at the source view (left), we generate the output scene layouts at the target view using our model and the baselines.