CompletionFormer: Depth Completion with Convolutions and Vision Transformers

Youmin Zhang[1] Xianda Guo[2] Matteo Poggi[1]
Zheng Zhu[2] Guan Huang[2] Stefano Mattoccia[1]

[1] University of Bologna [2] PhiGent Robotics

[1]{youmin.zhang2, m.poggi, stefano.mattoccia}@unibo.it

A. Appendix

A.1. Qualitative Results on NYUv2 Dataset

Qualitative results concerning the NYUv2 dataset [3] are provided in Fig. 1. In both visualized cases, we can notice the improved results yielded by our CompletionFormer compared to NLSPN [2]. Especially for the transparent regions near the windows in both cases, with local details of convolution and global cues of Transformer, our complete model (Ours) predicts clear object boundaries while NLSPN and Ours-ViT give blurry estimations.

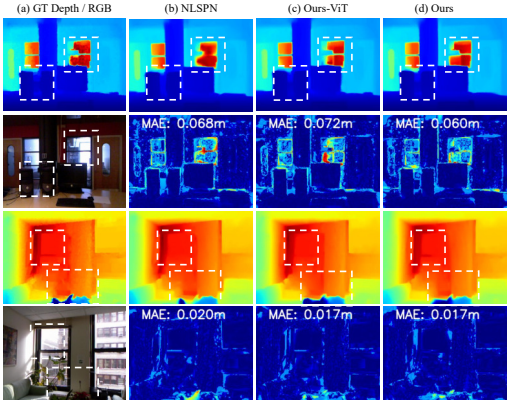

Figure 1. **Qualitative results on NYUv2 dataset.** Comparisons of our method against state-of-the-art method, *i.e*. NLSPN [2] are presented. We provide RGB images, and dense predictions. The colder the colors of the error map, the lower the errors. Ours-ViT denotes that only the Transformer layer is enabled in our proposed block.

A.2. Model Architecture Details

To better understand our architecture and to ease reproducibility, we present the network parameters of our CompletionFormer in Tab. 1.

Name	Layer setting	Output dimension
RGB and Depth Embedding		
input		RGB Image: $H \times W \times 3$ Sparse Depth: $H \times W \times 1$
conv_separate	Conv $3 \times 3, 48$ for RGB Image Conv $3 \times 3, 16$ for Sparse Depth	RGB Feature: $H \times W \times 48$ Sparse Depth Feature: $H \times W \times 16$
conv1	concat [RGB, Sparse Depth Feature] Conv $3 \times 3, 64$	$H \times W \times 64$
Joint Convolutional Attention and Transformer Encoder		
conv2	ResNet34 [1] BasicBlock $\times 3$	$H \times W \times 64$
conv3	ResNet34 [1] BasicBlock $\times 4$	$\frac{1}{2}H \times \frac{1}{2}W \times 128$
conv4	Joint Convolutional Attention and Transformer Block $\times 3$	$\frac{1}{4}H \times \frac{1}{4}W \times 64$
conv5	Joint Convolutional Attention and Transformer Block $\times 3$	$\frac{1}{8}H \times \frac{1}{8}W \times 128$
conv6	Joint Convolutional Attention and Transformer Block $\times 6$	$\frac{1}{16}H \times \frac{1}{16}W \times 320$
conv7	Joint Convolutional Attention and Transformer Block $\times 3$	$\frac{1}{32}H \times \frac{1}{32}W \times 512$
Decoder		
dec6	ConvTranspose 3×3, stride $= 2$, 256 Convolutional Attention Layer	$\frac{1}{16}H \times \frac{1}{16}W \times 256$
dec5	concat [dec6, conv6] ConvTranspose 3×3, stride $= 2$, 128 Convolutional Attention Layer	$\frac{1}{8}H \times \frac{1}{8}W \times 128$
dec4	concat [dec5, conv5] ConvTranspose 3×3, stride $= 2$, 64 Convolutional Attention Layer	$\frac{1}{4}H \times \frac{1}{4}W \times 64$
dec3	concat [dec4, conv4] ConvTranspose 3×3, stride $= 2$, 64 Convolutional Attention Layer	$\frac{1}{2}H \times \frac{1}{2}W \times 64$
dec2	concat [dec3, conv3] ConvTranspose 3×3, stride $= 2$, 64 Convolutional Attention Layer	$H \times W \times 64$
Initial Depth, Confidence, Non-local Neighbors, Affinity Prediction Head		
dec1	concat [dec2, conv2] Conv $3 \times 3, 64$	$H \times W \times 64$
dec0	concat [dec1, conv1] Conv $3 \times 3, \eta$	$H \times W \times \eta$
SPN Refinement		
refine	Spatial Propagation Network [2] with recurrent time $K = 6$	$H \times W \times 1$

Table 1. Network parameters of CompletionFormer. 'concat' means performing concatenate operation at Channel dimension. For each prediction head, it takes almost the same design, and only the output channel η is dependent on the output type, *e.g*. $\eta = 1$ for initial depth prediction.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–645. Springer, 2016.

[2] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *ECCV*, pages 120–136. Springer, 2020.

[3] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. Springer, 2012.