Improving Depth Gradient Continuity in Transformers: A Comparative Study on Monocular Depth Estimation with CNN – Supplementary Material –

Jiawei Yao[1]
jwyao@uw.edu
Tong Wu[1]
tw96@uw.edu
Xiaofeng Zhang[2]
framebreak@sjtu.edu.cn

[1] School of Engineering and Technology,
University of Washington,
Tacoma, USA

[2] School of Electronic Information and
Electrical Engineering,
Shanghai Jiao Tong University,
Shanghai, China

1 Transformer Estimation Depth from Sparse Pixels

As shown in Table 1 and Figure 1, we progressively increase the sparsity level of the input image. Our observations are as follows: (1) The depth reconstruction accuracy of both networks decreases with increasing mask sparsity. At the same sparsity level, the performance of the Transformer is notably superior to that of the CNN. Remarkably, the Transformer, when recognizing only 25% of the image regions, exhibits comparable performance to the CNN recognizing 60% of the regions. (2) The Transformer demonstrates better robustness compared to the CNN. As evident from the table, with increasing sparsity, even when retaining only 36% of the input image information (the sparsity level used in subsequent experiments), the RMSE drops by only 0.123 for the Transformer, while it drops by 0.203 for the CNN.

From the above data analysis, it is evident that the Transformer exhibits better robustness than the CNN. However, the exact nature of this disparity remains elusive. To delve deeper, we conduct a visual analysis at a sparsity level of 36% for both models. To further investigate the differences between the two networks, we analyze their differences through subjective visual results. Figure 2 displays the input image, masks trained by Transformer and CNN, and depth maps obtained from sparse pixels by Transformer and CNN. From the images, we observe that: (1) Both Transformer and CNN show interest in similar regions when selecting depth cues. However, the

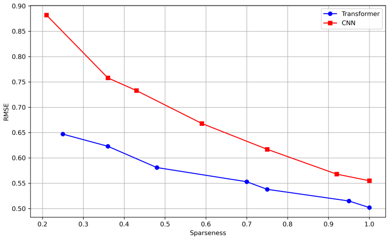

Figure 1: Performance of the two networks at different sparsity levels. A larger RMSE indicates poorer performance.

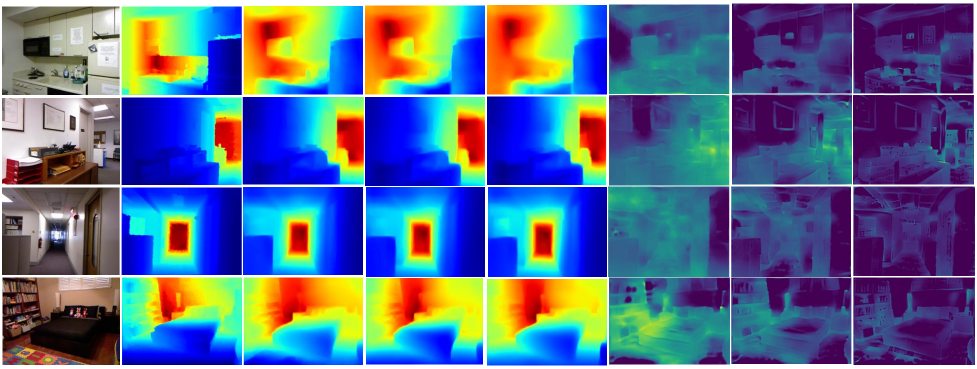

Figure 2: From left to right: RGB, Ground Truth, depth maps predicted by the network, and areas of interest in the image when λ = 1, 3, 5.

λ	RMSE	Sparseness	RMSE	Sparseness
original	0.502	1.00	0.555	1.00
$\lambda = 1$	0.515	0.95	0.568	0.92
$\lambda = 2$	0.538	0.75	0.617	0.75
$\lambda = 3$	0.553	0.70	0.668	0.59
$\lambda = 4$	0.581	0.48	0.733	0.43
$\lambda = 5$	0.623	0.36	0.758	0.36
$\lambda = 6$	0.647	0.25	0.882	0.21

Table 1: Pilot study results on NYU-Depth-V2 [1]. Depth reconstruction performance of two networks at different sparsity levels. From left to right, the columns represent the hyperparameter controlling sparsity, depth estimation accuracy of Transformer and CNN, sparsity level, and RMSE based on ground truth predictions.

Transformer is more sensitive to image boundaries and object contours than the CNN, resulting in clearer and more accurate depth estimation at the boundaries. Moreover, the Transformer has a stronger ability to distinguish between the foreground and background of an image, whereas the CNN might not differentiate them well in certain scenarios. (2) Due to its global attention mechanism, the Transformer captures the contextual relationships of the entire image better, especially in distant areas. This makes its depth estimation between objects and the background more accurate. In contrast, the CNN, with its convolution operation, excels in capturing local textures and shape information, producing depth maps with smooth depth gradients. However, for complex textures or color patterns, the CNN might misinterpret. (3) While the Transformer captures global information, its generated depth map might show unnatural depth jumps in some smooth areas, affecting continuity. The CNN excels in producing depth maps with clear depth gradients, especially at object edges and in texture-rich areas, offering a more vivid and three-dimensional visual effect.

References

[1] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th*

European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12, pages 746–760. Springer, 2012.