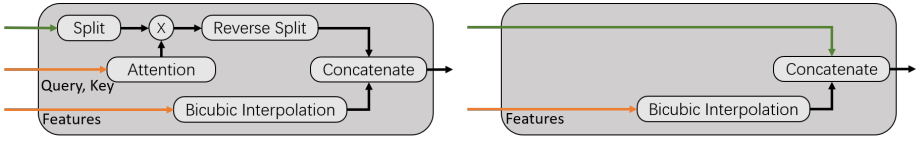
Transformer-based Monocular Depth Estimation with Attention Supervision - Supplementary Material

Wenjie Chang
changwj@mail.ustc.edu.cn
Yueyi Zhang*
zhyuey@ustc.edu.cn
Zhiwei Xiong
zwxiong@ustc.edu.cn

Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China

1 DUB Architecture

In this section, we show the difference of Attention-based Up-sample Block (AUB) and Direct Up-sample Block (DUB). Fig. 1(a) and Fig. 1(b) show the architectures of AUB and DUB, respectively. Compared with AUB, DUB directly concatenates feature maps from Image Branching and the outputs from Transformer Layer at the same scale.



(a) AUB (b) DUB

Figure 1: Architecture comparison between AUB and DUB. The orange route denotes elements from Transformer Layer. The green route represents the feature maps from Image Branching.

2 Influence of the factor λ

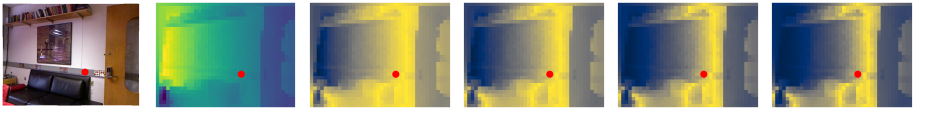
We calculate attention maps used in AS as follows:

$$\hat{A}_{i,j} = SOFTMAX(-\lambda|\hat{D} - \hat{d}_{i,j}|), \tag{1}$$

where $\hat{D}$ represents the depth map which is $16\times$ down-sampled from the ground-truth depth map and normalized to $[0,1]$, $\hat{d}_{i,j}$ represents the depth value at position (i,j), λ is a hyper-parameter. Fig. 2 shows attention maps with different λ values. It can be seen that with the

* Corresponding Author

increase of the λ value, small regions are be paid attention to. In our experiment, we set λ to 8.

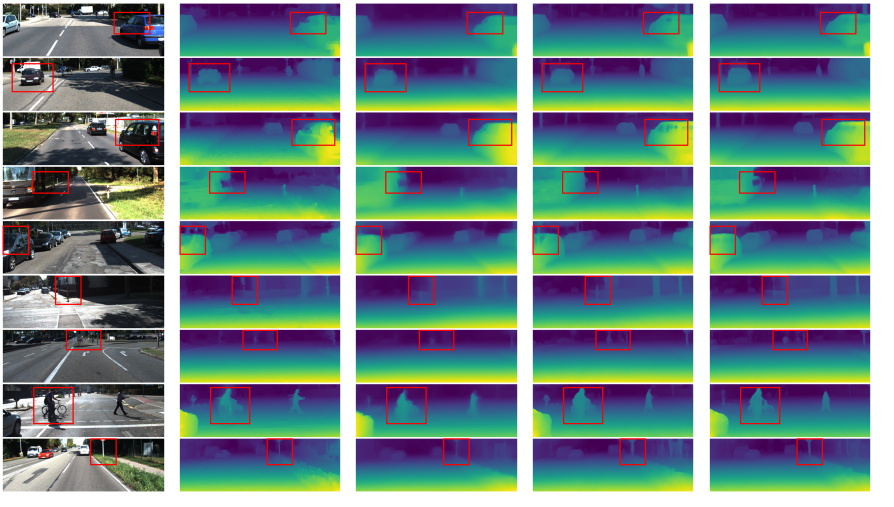


(a) RGB (b) Depth Map (c) $\lambda = 1$ (d) $\lambda = 4$ (e) $\lambda = 8$ (f) $\lambda = 10$

Figure 2: Attention maps (c) (d) (e) (f) shows the correlations between the red point and other pixels.

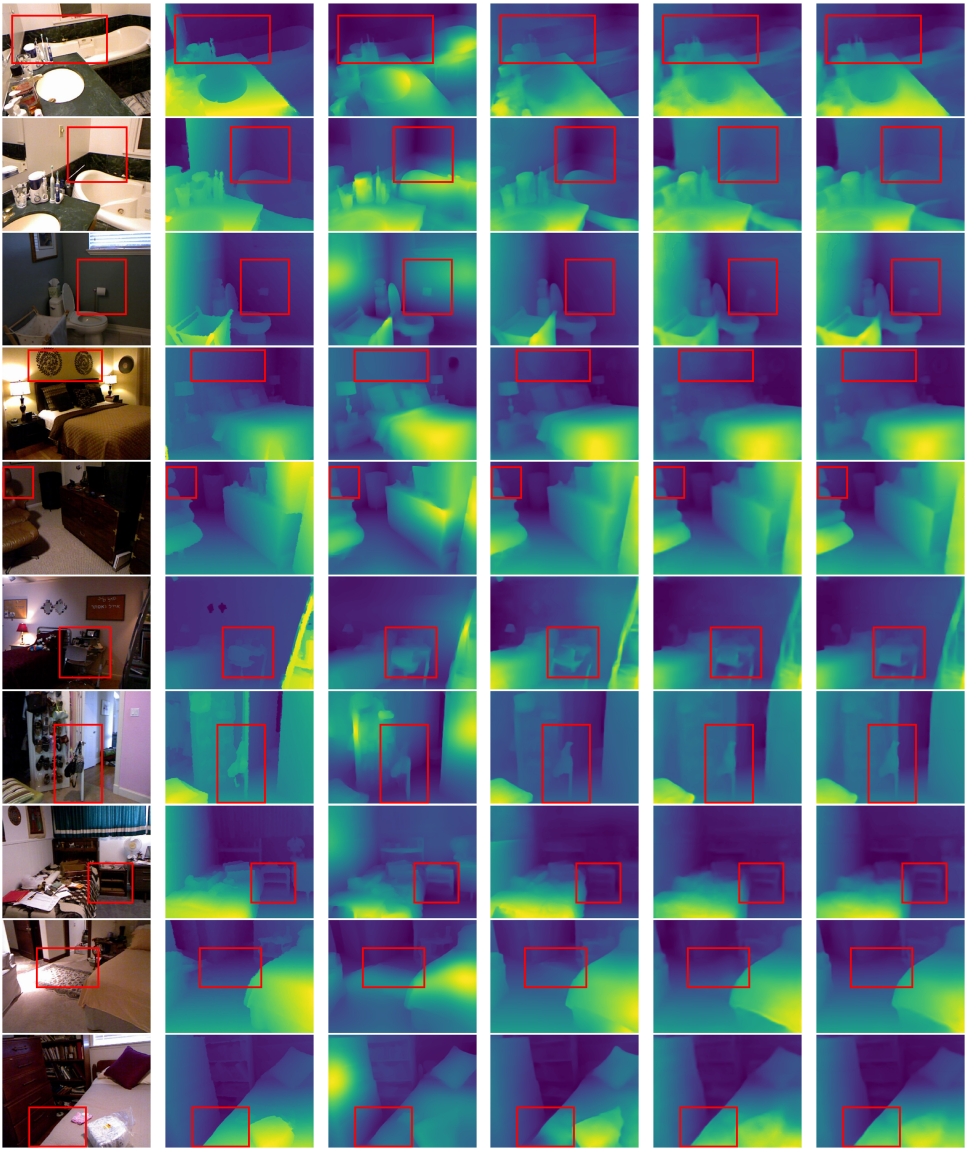
3 Additional Qualitative Results

We present the qualitative comparison results with the state-of-the-art methods in Fig. 3 and Fig. 4. Fig. 3 shows the depth results on the KITTI dataset. Our proposed method is compared with VNL [3] and LapDepth [2]. It can be observed that in the areas with complex lighting situations, such as car windows (Row 1, 3, 5 in Fig. 3) and shadow regions (Row 4, 8 in Fig. 3), our method always provides better depth results. Fig. 4 shows the depth results on the NYU Depth V2 dataset. Comparing with SARPN [1], LapDepth [2] and our network without Attention Supervision (AS), our method demonstrates more accurate depth results around the planes (Row 4, 9 in Fig. 4) and overcomes the artifacts coming from shadows (Row 5, 6 in Fig. 4). By comparing Fig. 4(e) and Fig. 4(f), it is obvious that after adding the Attention-based Up-Sample Block (AUB), our method predicts depth maps with sharper boundaries.



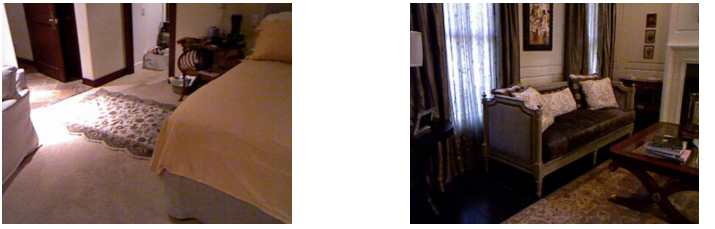
(a) RGB (b) Ground Truth (c) VNL [3] (d) LapDepth [2] (e) Ours

Figure 3: Qualitative comparisons with other methods on the KITTI dataset.



(a) RGB (b) Ground Truth (c) SARPN [8] (d) LapDepth [20] (e) w/o AUB (f) Ours

Figure 4: Qualitative comparisons with other methods on the NYU Depth V2 dataset.



(a) Scene 1 (b) Scene 2

(c) Ground Truth (d) LapDepth [] (e) Ours w/o AS (f) Ours

(g) Ground Truth (h) LapDepth [] (i) Ours w/o AS (j) Ours

Figure 5: Point clouds comparison with other methods on the NYU Depth V2 dataset.

4 Point Cloud Reconstructions

The point clouds shown in Fig. 5 are rendered from depth maps in the same way as [4]. Point clouds in same row are captured from the same viewpoint. We compare our method with LapDepth [2] and our method without AS. It can be observed that point clouds generated from our proposed method is closer to that of ground-truth, especially around the wall and floor.

References

[1] Xiaotian Chen, Xuejin Chen, and Zheng-Jun Zha. Structure-aware residual pyramid network for monocular depth estimation. In *International Joint Conferences on Artificial Intelligence*, page 694–700, 2019.

[2] M. Song, S. Lim, and W. Kim. Monocular depth estimation using laplacian pyramid-based depth residuals. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[3] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5684–5693, 2019.

[4] Wei Yin, Yifan Liu, and Chunhua Shen. Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.