1. Network details

1.1. Semantic segmentation network

Our pre-trained semantic segmentation network used for semantic visual conditions in the diffusion denoising process consists of an encoder-decoder structure that follows a design similar to the Deeplabv3+ [3]. Here we leverage a ResNet101 [4] encoder architecture. The encoder module produces multi-scale semantic feature maps with resolution down-sampled by $1/4, 1/8, 1/16$, and $1/32$ of the original image size (512×1024). These multi-scale features are then embedded by the Atrous Spatial Pyramid Pooling (ASPP) module. The ASPP module utilizes parallel convolution layers with multiple rates of $1, 6, 12$, and 18 capturing an effective field of view. The low-level encoded feature and the up-sampled $128 \times 256\times$ channel-dimension ASPP-embedded features are then concatenated and provided to the decoder module with three stages of upsampling, 3×3 kernel convolution followed by batch normalization and ReLU activation layers to produce the semantic segmentation output $512\times1024\times C$ where C is the number of semantic category channels.

For semantic segmentation network training, we use per-pixel cross-entropy loss given by Eq. (1):

$$\mathcal{L}_{semantic} = -\sum_{i} p_i log(\hat{p}_i) \tag{1}$$

where p_i and $\hat{p}_i$ are the ground truth and predicted pixel-wise semantic labels, respectively. The ground truth pixel-wise semantic labels are utilized from the dataset. For Stanford2D3D [1] we utilize the provided ground truth semantic segmentation labels which consist of 14 categories namely unknown, beam, board, bookcase, ceiling, chair, clutter, column, door, floor, sofa, table, wall, and window. For Matterport3D [2] dataset we use the provided ground truth semantic segmentation consisting of 40 semantic labels namely void, wall, floor, chair, door, table, picture, cabinet, cushion, window, sofa, bed, curtain, chest of drawers, plant, sink, stairs, ceiling, toilet, stool, towel, mirror, tv monitor, shower, column, bathtub, counter, fireplace, lighting, beam, railing, shelving, blinds, gym equipment, seating, board panel, furniture, appliances, clothes, and objects.

1.2. Surface normal estimation network

In addition to semantics, we also utilize predicted surface normal information for the geometric planar guidance to the depth denoising process. For this, we use a surface normal estimation network. The network consists of a pre-trained ResNet101 feature extractor, a spatial pooling module, and a decoder architecture similar to a semantic segmentation network. However, the last layer of the decoder produces $512 \times 1024 \times 3$ shaped surface normal maps at the output.

For surface normal estimation network training, we utilize negative cosine loss given by Eq. (2):

$$\mathcal{L}_{normal} = 1 - \sum_{i} \hat{N}_i N_i^T \tag{2}$$

where $\hat{N}_i$ and $N_i{}^T$ are the predicted and ground truth pixel-wise normal values, respectively. The ground truth pixel-wise normal values are generated from the ground truth depth maps. For this, we first convert the panorama 360 ground truth depths to the 3D point clouds. Then we calculate each pixel's normal vector by performing the cross-product between its 3D position and neighbors and finally normalize the vector. This process generates a 3-channel (due to 3D normal vector) ground truth surface normal map from a 1-channel ground truth depth map.

1.3. Details of the conditional latent diffusion model

In this section, we discuss the details of the OmniDiffusion architecture as shown in Fig. 1. As mentioned in the main paper we propose utilizing RGB (channel = 3), semantic segmentation (channel = S: number of semantic categories), and surface normal maps (channels = 3) as the visual conditions for the Denoising Diffusion Probabilistic Model (DDPM) network. Our framework utilizes a latent space diffusion model for which we encode the visual conditions of shape $512\times1024\times(3+S+3)$ to $256\times512\times C''$ where C'' is the channel dimension. First, we divide the input feature maps with shape $512\times1024\times(3+S+3)$ to multiple patches of size = 4×4 and then concatenate these patch features with their positional embedding to be given to the Swin transformer [9] that uses window-based self-attention to produce effective hierarchical feature maps with shapes $128\times256\times C$, $64\times128\times2C$, $32\times64\times4C$ and $16\times32\times8C$. These hierarchical feature maps are then aggregated using multi-scale self and cross-deformable attention mechanism presented in DepthFormer [8] to produce an encoded feature map of shape $128 \times 256 \times C'$. We then up-sample these encoded features to $256 \times 512 \times C''$ to match it to the latent ground truth depth features. We utilize the depth encoder discussed in the main paper to encode the ground truth depth features to produce $256 \times 512 \times 16$ shape latent depth feature map. These latent depth features are then concatenated with the latent visual condition features to be utilized for the conditional DDPM process.

1.4. Dataset and metric details

We perform experiments on the widely known benchmark datasets called Stanford2D3D [1] and Matterport3D [2] to evaluate the depth estimation performance. Stanford2D3D [1] is a large-scale real-world indoor scene dataset. In total, it consists of 1413 panorama images out of which we use 1040 for training and 373 for testing according to the official split. Matterport3D [2] consists of 10800

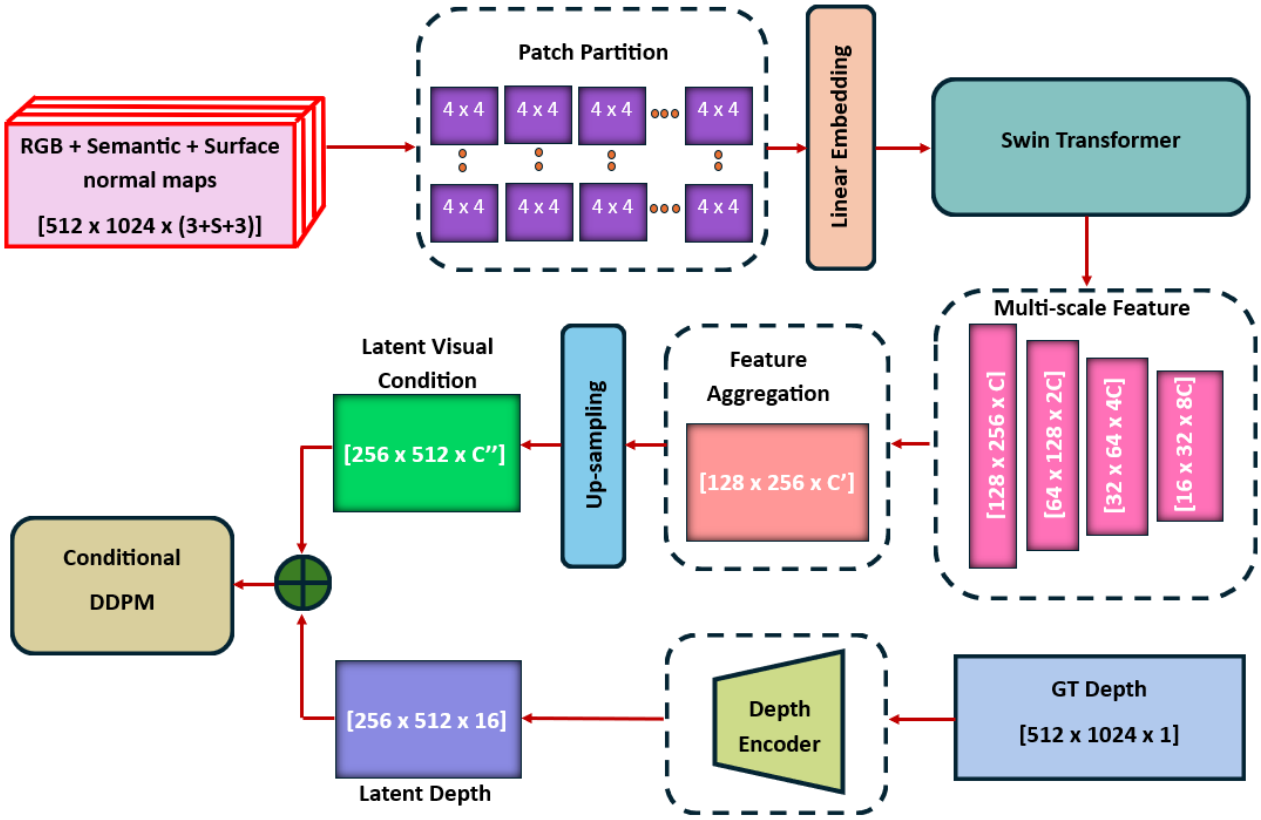

Figure 1. Details of the *OmniDiffusion* architecture. Here S denotes the number of semantic category channels.

RGBD images, of which we use 8786 images for training and the rest for testing. Our network takes an ERP image of 512×1024 resolution as input.

To evaluate the depth estimation performance we utilize the widely used metrics, mentioned in the literature work [5, 6, 12], called Absolute Relative Error (Abs Rel), Root Mean Squared Error (RMSE), and threshold-based accuracy δ_t, where $t \in 1.25, 1.25^2, 1.25^3$. The definitions of the mentioned metrics are shown in Eq. (3), Eq. (4), and Eq. (5) respectively.

$$AbsRel = \frac{1}{N}\sum_{i=1}^{N}\frac{|d_i - d_i^{'}|}{d_i^{'}} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|d_i - d_i^{'}|^2} \tag{4}$$

$$\delta_n = max(\frac{d_i}{d_i^{'}}, \frac{d_i^{'}}{d_i}) < (1.25^n) \tag{5}$$

where, $d^{'}$ is the ground truth depth value, d is the predicted depth value and N is the total number of valid image pixels.

2. Qualitative comparison using Matterport3D [2] dataset

We present the qualitative depth estimation comparison with the existing methods listed in the main paper using the larger Matterport3D [2] dataset as shown in Fig. 2 and Fig. 3. Similar to the Stanford2D3D [1] performance our model can recover better structure details with sharper object boundaries/edges and globally consistent geometric structure for Matterport3D [2] dataset as well.

3. Qualitative comparison with the existing diffusion-based monocular depth estimation models

In Fig. 4 and Fig. 5 we present the qualitative comparative depth performance of our OmniDiffusion method with the existing SOTA diffusion-based methods called VPD [13] and Ecodepth [10]. As observed, compared to our method the existing stable-diffusion-based methods produce low quality, blurred edges, poor geometric details, and inconsistent depths.

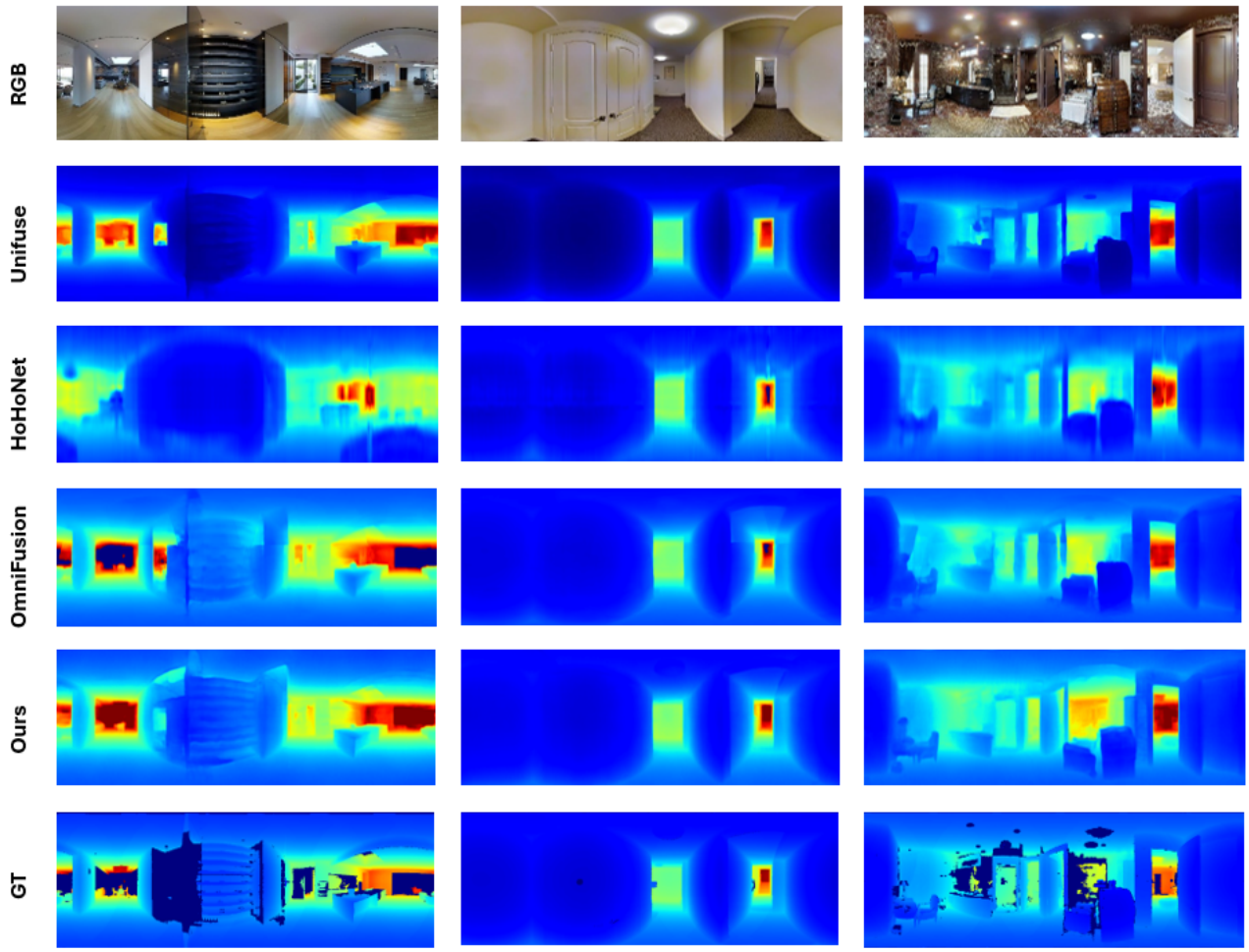

Figure 2. Comparative qualitative depth estimation results on Matterport3D [2] benchmark dataset. We show the performance of UniFuse [5] (second row), HoHoNet [11] (third row), OmniFusion [7] (fourth row) and our model (fifth row) with RGB ERP input and Ground Truth (GT) depth map shown in the first and last row respectively.

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 1, 2, 5, 6

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1, 2, 3, 4

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[5] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360 panorama depth estimation. *IEEE Robotics and Automation Letters*, 6(2):1519–1526, 2021. 2, 3, 4

[6] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 2

[7] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2801–2810, 2022. 3, 4

[8] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *Machine Intelligence Research*, 20(6):837–854, 2023. 1

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng

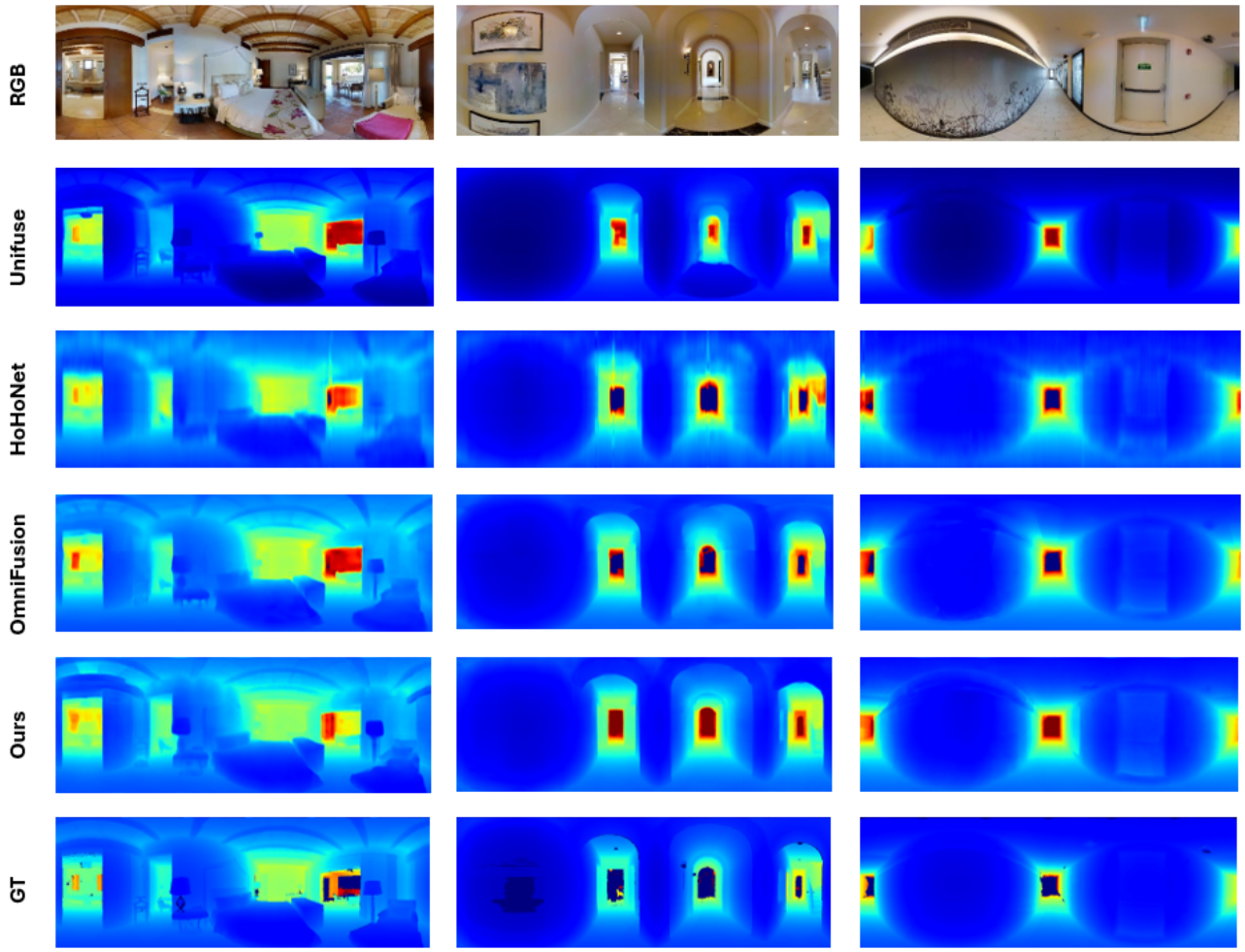

Figure 3. More comparative qualitative depth estimation results on Matterport3D [2] benchmark dataset. We show the performance of UniFuse [5] (second row), HoHoNet [11] (third row), OmniFusion [7] (fourth row) and our model (fifth row) with RGB ERP input and Ground Truth (GT) depth map shown in the first and last row respectively.

Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[10] Suraj Patni, Aradhye Agarwal, and Chetan Arora. Ecodepth: Effective conditioning of diffusion models for monocular depth estimation. *arXiv preprint arXiv:2403.18807*, 2024. 2, 5, 6

[11] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2573–2582, 2021. 3, 4

[12] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 462–471, 2020. 2

[13] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *ICCV*, 2023. 2, 5, 6

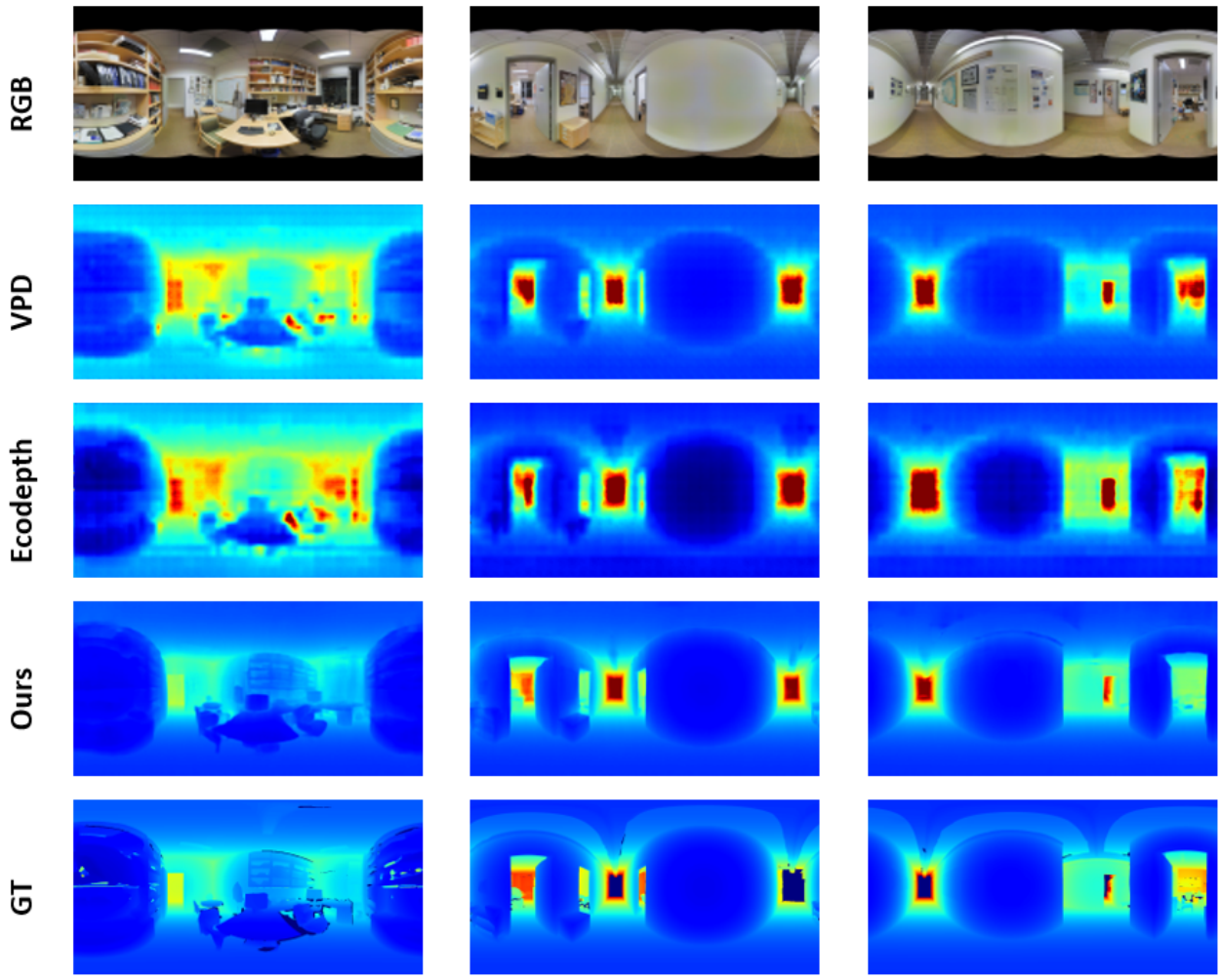

Figure 4. Comparative qualitative depth estimation results on Stanford2D3D [1] benchmark dataset. We show the performance of VPD [13] (second row), Ecodepth [10] (third row), and our model (fourth row) with RGB ERP input and Ground Truth (GT) depth map shown in the first and last row respectively.

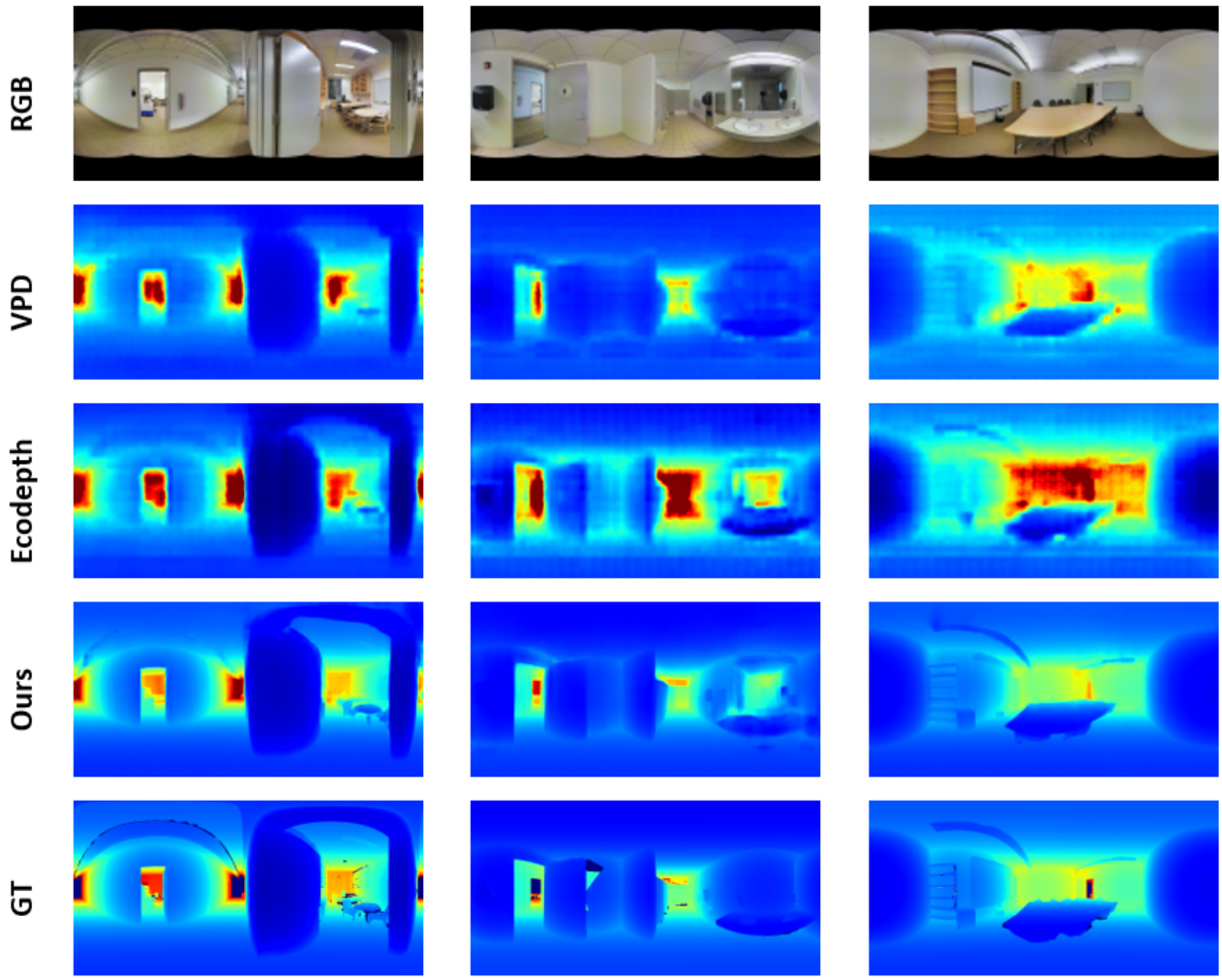

Figure 5. More comparative qualitative depth estimation results on Stanford2D3D [1] benchmark dataset. We show the performance of VPD [13] (second row), Ecodepth [10] (third row), and our model (fourth row) with RGB ERP input and Ground Truth (GT) depth map shown in the first and last row respectively.