Appendix for OmniFusion: 360 Monocular Depth Estimation via Geometry-Aware Fusion

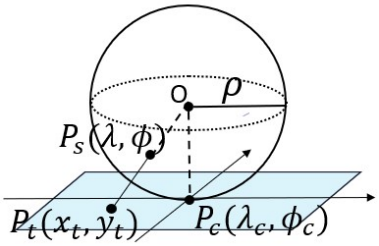

Figure 1. The illustration of the gnomonic projection. A point $P_s(\lambda, \phi)$ located on the spherical sphere is projected onto a point $P_t(x_t, y_t)$ on the flat plane which is tangent to a point $P_c(\lambda_c, \phi_c)$.

A. Gnomonic projection

We use the distortion-free tangent image representation to address the irregular 360 image distortion. Tangent image is the *gnomonic projection* of a sphere surface onto a flat, rectangular plane surface. The gnomonic projection [3] (Figure 1) is a map projection obtained by projecting points P_s on the surface of sphere from a sphere's center O to point P_t in a plane that is tangent to a point P_c.

For a pixel on the ERP image $P_e(x_e, y_e)$, we first find its corresponding point $P_s(\lambda, \phi)$ locating on the unit sphere.

$$\lambda = \frac{2\pi x_e}{W}, \ \phi = \frac{\pi y_e}{H} \tag{1}$$

where H and W are height and width of the ERP image. The projection from $P_s(\lambda, \phi)$ to $P_t(x_t, y_t)$ is defined as:

$$\begin{aligned} x_t &= \frac{\mathbf{cos}(\phi)\mathbf{sin}(\lambda - \lambda_c)}{\mathbf{cos}(c)} \\ y_t &= \frac{\mathbf{cos}(\phi_c)\mathbf{sin}(\phi) - \mathbf{sin}(\phi_c)\mathbf{cos}(\phi)\mathbf{cos}(\lambda - \lambda_c)}{\mathbf{cos}(c)} \\ \mathbf{cos}(c) &= \mathbf{sin}(\phi_c)\mathbf{sin}(\phi) + \mathbf{cos}(\phi_c)\mathbf{cos}(\phi)\mathbf{cos}(\lambda - \lambda_c) \end{aligned} \tag{2}$$

where (λ_c, ϕ_c) are the spherical coordinates of the tangent plane center P_s.

The inverse gnomonic transformations are:

$$\begin{aligned} \lambda &= \lambda_c + \mathbf{tan}^{-1}\left(\frac{x_t \ \mathbf{sin}(c)}{\gamma \ \mathbf{cos}(\phi_1)\mathbf{cos}(c) - y_t \ \mathbf{sin}(\phi_c)\mathbf{sin}(c)}\right) \\ \phi &= \mathbf{sin}^{-1}\left(\mathbf{cos}(c)\mathbf{sin}(\phi_c) + \frac{1}{\gamma} y_t \mathbf{sin}(c)\mathbf{cos}(\phi_c)\right) \end{aligned} \tag{3}$$

where $\gamma = \sqrt{x_t^2 + y_t^2}$ and $c = \mathbf{tan}^{-1}\gamma$.

With Equation 2 and 3, we can build one-to-one forward and inverse mapping functions between pixels on the ERP image and pixels on the tangent image.

B. Geometry-aware feature fusion

As the geometry-aware feature fusion module is one of the major innovations of our paper, in this section we provide more detailed illustrations. As shown in Figure 2, more intermediate representations involved in the module is visualized. Specifically, the patch-wise 2D image features and the patch-wise geometric features are visualized separately, along with the feature maps after fusion, in which the mean value of each feature is shown. For visual comparison, the patch-wise features before Figure 2 (b) and after fusion (c) are projected and merged into two ERP feature maps. As observed, the fused feature maps inherit more locally consistent structures, which is expected to lead to more locally consistent depth results. It is worth mentioning that patch-wise geometric features are fixed once learned when the inputs are just based on the spherical coordinates with fixed ρ, and independent from the image. This means no extra computation in inference is needed for the first iteration. While for the second iteration, since ρ depends on the input image, new geometric features need to be re-computed, but the MLPs are super light-weight compared to the original CNNs.

The intuition behind the geometry-aware fusion design can be visualized in high-dimensional feature space, see Figure 3. Based on the Equation 2, a single point from the ERP space, $P_s^i(\lambda^i, \phi^i, \rho^i)$, is projected to two tangent images centered at (λ_c^j, ϕ_c^j) and (λ_c^k, ϕ_c^k), and appear at (x_t^j, y_t^j) and (x_t^k, y_t^k), respectively. As observed, different appearances at the two points can lead to different image features encoded from the shared CNN kernel, which

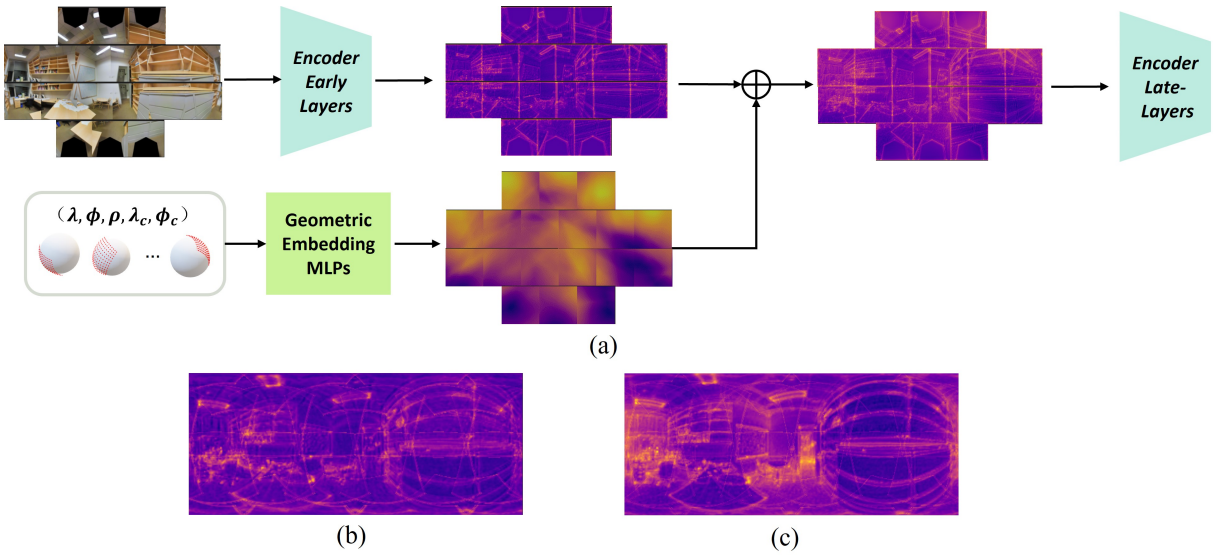

Figure 2. (a) Detailed pipeline of geometry-aware feature fusion. A set of tangent images are encoded into a set of image feature maps, while the 3D coordinates are encoded and converted into a set of geometric feature maps. The patch-wise 2D image features are fused with the patch-wise geometric feature. (b) The merged ERP feature map of patch features without the geometric fusion. (c) The merged ERP feature map of patch features with the geometric fusion. Comparing to the merged ERP feature maps without geometric fusion in (b), the geometry-aware fused ERP feature map in (c) appears to be more locally consistent.

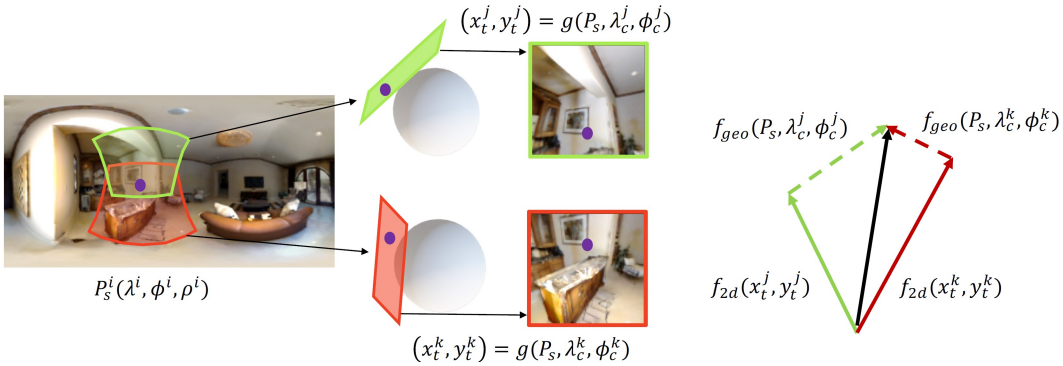

Figure 3. A more intuitive view of geometry-aware feature fusion. Based on the gnomonic geometry, a single point from the ERP space, $P_s^i(\lambda^i, \phi^i, \rho^i)$ is projected to two tangent images centered at (λ_c^j, ϕ_c^j) and (λ_c^k, ϕ_c^k), and appear at two different pixels (x_t^j, y_t^j) and (x_t^k, y_t^k), respectively. Image features located at the two pixels can be visualized in high-dimensional vectors (solid green and red arrows in the right panel, respectively). Since the discrepancy is caused by the gnomonic transformation from (P_s, λ_c, ϕ_c), we utilize geometric features encoded from (P_s, λ_c, ϕ_c) to compensate for the discrepancy (dashed arrows).

can be visualized as high-dimensional vectors (solid green and red arrows on the right panel). Such difference in the 2D features will make the merged results appear to be locally inconsistent. Since the discrepancy is caused by the gnomonic transformation from (P_s, λ_c, ϕ_c), we believe a point-encoding model can learn a geometric embedding space out of (P_s, λ_c, ϕ_c) to mitigate the discrepancy (dashed arrows). While P_s makes the embedding to be aware of the global position, (λ_c, ϕ_c) differentiates between patches to enable the compensation.

C. Transformer Architecture and Ablation Study

The architecture of the multi-head attention transformer follows [8]:

$$
\begin{aligned}
z_0 &= [x^1E, x^2E, ..., x^NE] + E_{pos}, \\
z'_l &= Norm(MSA(z_{l-1}, z_0) + z_{l-1}), \\
z_l &= Norm(FFN(z'_l) + z'_l),
\end{aligned} \tag{4}
$$

Configurations	#Params	Abs Rel↓	Sq Rel ↓	RMSE↓	RMSE(log)↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
depth = 2, num of heads = 2	19M	0.1091	0.0614	0.3885	0.1782	0.8738	0.9670	0.9891
depth = 4, num of heads = 4	24M	**0.1016**	0.0583	**0.3796**	0.1774	0.8867	0.9688	0.9885
depth = 6, num of heads = 4	32M	0.1026	**0.0572**	0.3883	**0.1753**	**0.8893**	**0.9689**	**0.9892**
depth = 8, num of heads = 8	38M	0.1044	0.0596	0.3926	0.1819	0.8739	0.9650	0.9873

Table 1. The ablation study of the transformer configurations. We use ResNet18 as encoder for all experiments.

where $Norm$ represents layer normalization, $l = 1, ..., L$ is the index of the transformer block. The multi-headed self-attention (MSA) is computed as:

$$\begin{aligned} MSA(X) &= concat_{h=1}^{H}[Attn_h(X)]W \\ Attn_h(X) &= softmax(\frac{QK^T}{\sqrt{d_h}})V \\ Q = XW_Q,\ K &= XW_K,\ ,V = XW_V \end{aligned} \tag{5}$$

where Q, K, V correspond to query, key, value matrix, respectively. h denotes the number of heads. We reshape the transformer output, then use another 1×1 convolution layer to increase feature dimension, and add the encoder output as residual.

An ablation study on the transformer depth and the number of heads is shown in Table 1. The ablation study here is conducted based on ResNet18, not the ResNet34 used in our final pipeline, in order to conduct the experiments more efficiently. The number of parameters shown in the table considers the entire network rather than the transformer module alone. We chose 6 transformer blocks (depth=6) and a number of 4 heads (number of heads=4) as the default configuration, as this configuration tends to have fewer errors and higher inlier ratios.

D. Loss Function

Our network is trained in an end-to-end fashion. We adopt BerHu loss [6] for optimizing depth predictions of all iterations.

$$\mathcal{L}_{depth} = \begin{cases} |\Delta D|, |\Delta D| \leq c \\ \frac{\Delta D^2 + c^2}{2c}, |\Delta D| > c \end{cases} \tag{6}$$

where $\Delta D = |D_{gt} - D_e| * M$ is the absolute difference of ground truth depth D_{gt} and the predicted depth D_e. M is a binary mask that mask out invalid depth pixels. c is a border value defined as the 20% of the maximum per batch residual $c = 0.2max(\Delta D)$.

The final loss term is the combination of losses from all iterations:

$$\mathcal{L}_{total} = \sum_i \mathcal{L}_{depth} \tag{7}$$

E. Generalization

We conducted a cross-dataset evaluation and summarized the results in Table 2. All methods in the table are trained on Matterport3D [2] training set and evaluated on Stanford2D3D [1] test set. We used the official pre-trained models and the evaluation code provided by UniFuse [5] and HoHoNet [7] for a fair comparison. As observed, our method showed superior generalization ability compared to these state-of-the-arts methods.

Methods	Abs Rel↓	Sq Rel↓	RMSE↓
UniFuse [5]	0.1192	0.0813	0.4291
HoHoNet [7]	0.1083	0.0755	0.4166
OmniFusion, Ours	**0.1044**	**0.0620**	**0.3781**

Table 2. Cross-dataset evaluation.

F. Additional qualitative comparisons

Besides the qualitative comparison between our method and the baseline method tailored from [4], we also extend to qualitatively compare our method with current state-of-the-art methods, HoHoNet [7] and UniFuse [5] on three datasets: Stanford2D3D [1], Matterport3D [2], and 360D [9]. The results are shown in Figure 4, 5, 6, respectively. We use the pretrained models downloaded from their official GitHub repositories, respectively. [1] [2] Note that the results from HoHoNet [7] are not included in Figure 6 because they have not reported results or releases code on 360D [9] dataset. Figure 7 shows additional qualitative results of our OmniFusion on Matterport3D [2] besides what have been provided on Stanford2D3D [1] in the main paper. All of these comparisons clearly show that our method recovers more structural details in the final depth maps, maintains sharp edges, smooth surfaces, and exhibits fewer errors.

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. 3, 4

[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. *International Conference on 3D Vision (3DV)*, 2017. 3, 4, 5

[3] Harold Scott Macdonald Coxeter. Introduction to geometry. 1961. 1

[1] https://github.com/sunset1995/HoHoNet

[2] https://github.com/alibaba/UniFuse-Unidirectional-Fusion

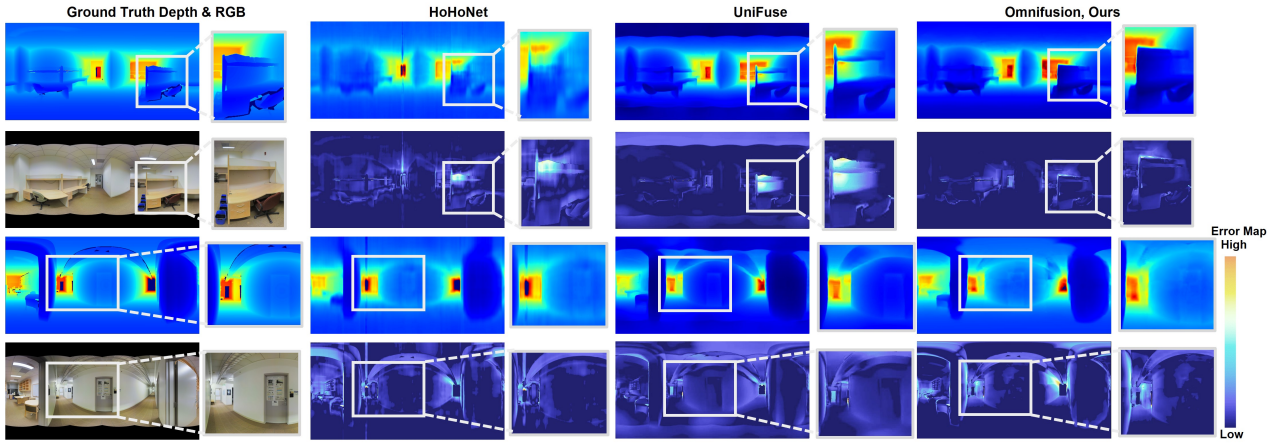

Figure 4. The qualitative comparisons with the current state-of-the-art works on the dataset Stanford2D3D [1]. We show the results of HoHoNet [7] (second column), UniFuse [5] (third column), and ours (last column). Both the depth maps and the error maps against the ground-truth are included for comparison. See the zoomed-in areas for detailed comparisons.

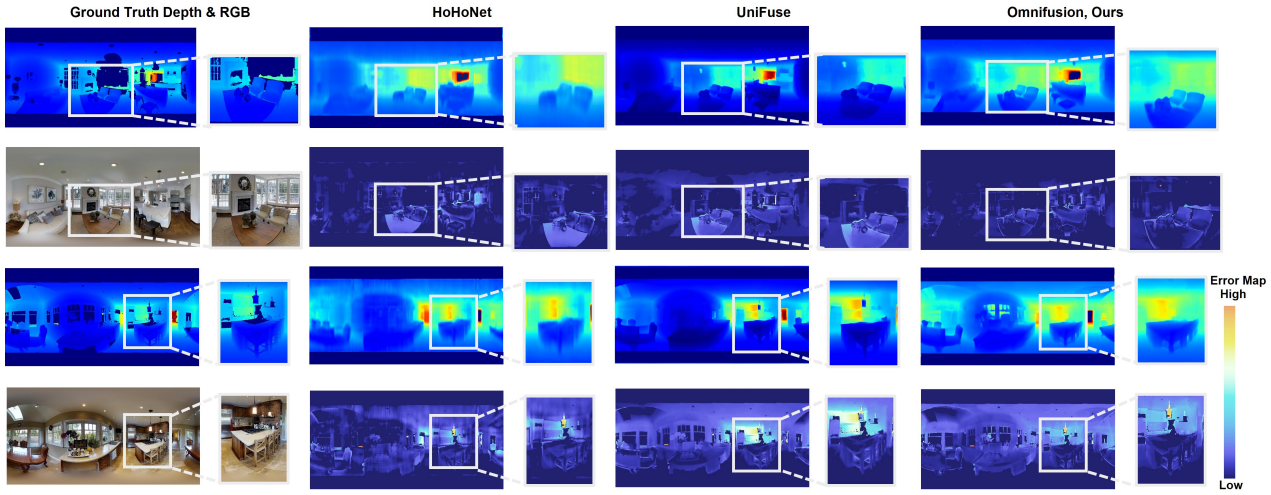

Figure 5. The qualitative comparisons with current state-of-the-art works on the dataset Matterport3D [2]. We show the results of HoHoNet [7] (second column), UniFuse [5] (third column), and ours (last column). Both the depth maps and the error maps against the ground-truth are included for comparison. See the zoomed-in areas for detailed comparisons.

[4] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12426–12434, 2020. 3

[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*, 2021. 3, 4, 5

[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. 3

[7] 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

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[9] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 448–465, 2018. 3, 5

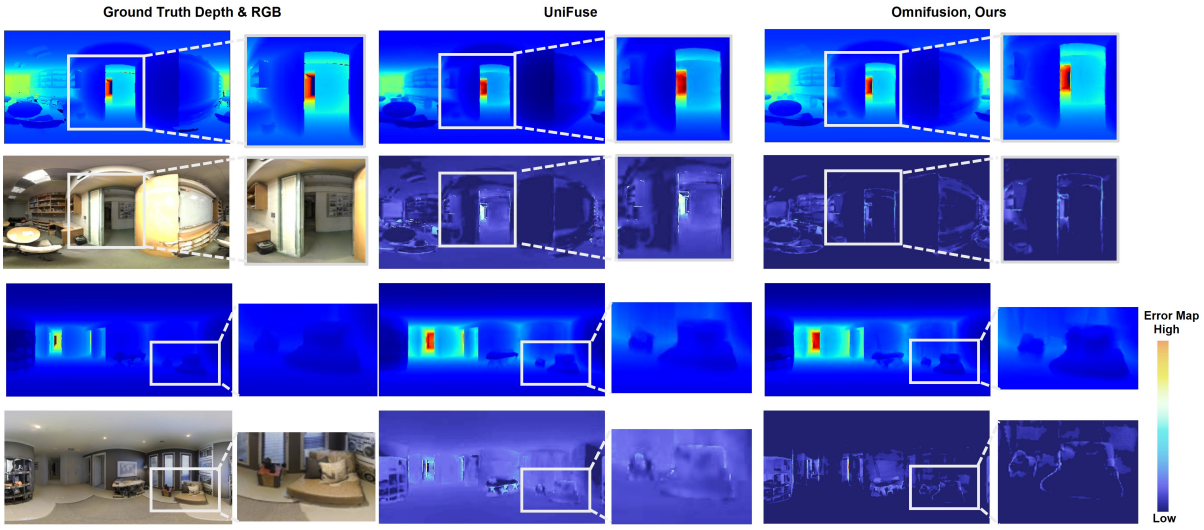

Figure 6. The qualitative comparisons with current state-of-the-art works on the dataset 360D [9], We show the results of UniFuse [5] (second column), and ours (last column). Both the depth maps and the error maps against the ground-truth are included for comparison. See the zoomed-in areas for detailed comparisons.

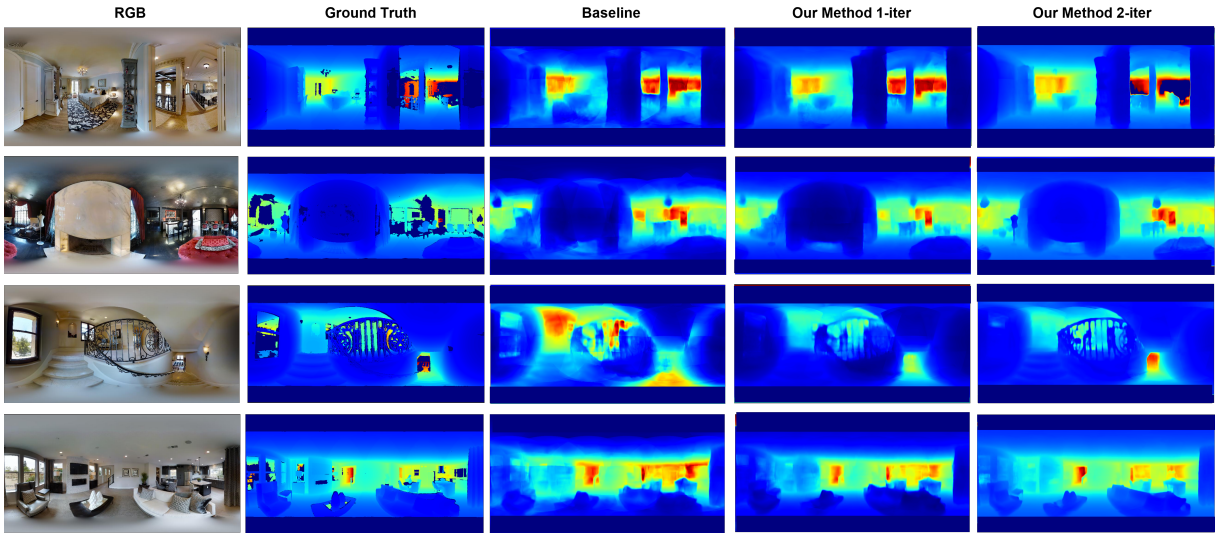

Figure 7. More qualitative results of OmniFusion on Matterport3D [2].