Pano3D: A Holistic Benchmark and a Solid Baseline for 360^o Depth Estimation – Supplementary Material

Georgios Albanis[*1] Nikolaos Zioulis[*1,2] Petros Drakoulis[1] Vasileios Gkitsas[1] Vladimiros Strezentsenko[1]
Federico Alvarez[2] Dimitrios Zarpalas[1] Petros Daras[1]
[1] Centre for Research and Technology Hellas, Thessaloniki, Greece
[2] Universidad Politécnica de Madrid, Madrid, Spain
{galbanis,nzioulis,petros.drakoulis,gkitsasv,vladster}@iti.gr
fag@gatv.ssr.upm.es {zarpalas,daras}@iti.gr
vcl3d.github.io/Pano3D

1. Introduction

This supplementary material complements our original manuscript with additional results, supporting further ablation experiments, providing qualitative results on real data and comparisons between the different architectures.

2. Supplementary Results

Table 1 complements Table 1 of the main document, presenting the performance of all remaining metrics, namely the spherical direct depth metrics, the boundary preservation metrics, and the smoothness metrics. In addition, Figure 1 presents the different models' performance in terms of three indicators, one for each trait. These indicators combine an error and an accuracy metric:

$$i_d = \frac{1}{(1-\delta_{1.25}) \times RMSE}, \tag{1}$$

$$i_b = \frac{1}{(1-(rec_{0.25}+rec_{0.5}+rec_{1.0})/3) \times dbe^{acc}}, \tag{2}$$

$$i_s = \frac{1}{(1-(\alpha_{11.25^o}+\alpha_{22.5^o}+\alpha_{30^o})/3) \times RMSE^o}, \tag{3}$$

with i_d, i_b, and i_s the depth, boundary and smoothness performance indicators. Evidently, UNet performs significantly better than the other models, especially in the boundary consistency metrics, while all models benefit of the addition of extra losses. The addition, of skip connections in a common ResNet architecture offers better performance. While $\mathcal{L}_{grad}$ offers better depth performance for ResNet$_{\text{skip}}$, the variant trained with $\mathcal{L}_{comb}$ offers higher performance across the two secondary traits.

[*]Indicates equal contribution.

In addition, we complement the main's paper spherical metrics Table 2 by collating the traditional ones for a straightforward comparison.

Finally, Table 3 reproduces the grounds upon our methodology was designed, namely the efficacy of pre-trained models [6] and the L1 loss [2]. We use the DenseNet and Pnas models with the encoders initialized using weights pre-trained on ImageNet. Both claims stand, with all pre-trained models achieving better performance than the model trained from scratch. In addition, the L1 loss outperforms both berHu [5] and log loss. Interestingly, the perfomance drops significantly in DenseNet when trained with other losses, while for Pnas the performance gap is smaller. Therefore, when benchmarking different models, this needs to be taken into account as well. Only through consistent experimentation across different aspects measurable and explainable progress will be possible.

3. Qualitative Results

Finally we present additional qualitative results for different models. Apart from the collation of the predicted depth maps between the different models, we provide an advantage visualisation technique similar to that presented in HoHoNet [7]. The visualisation is the MAE difference between two comparable models.

To that end, Figure 2 demonstrates the comparison of ResNet and ResNet$_{\text{skip}}$ architectures, Figure 3 that of the UNet and Pnas architectures, and, finally Figure 4 presents the differences between the Unet and ResNet$_{\text{skip}}$ architectures.

Additionally, Figure 5 presents comparative results regarding the boundary preservation performance across models. Once again, UNet is able to capture finer-grained details while the Pnas model produces smoother results.

Table 1: Three axis depth metrics performance across models and supervision schemes. Best three performers are denoted with bold faced **light green** (1st), light blue (2nd) and light purple (3rd) respectively following the ranking order. Same scheme applies to all tables.

	Model	Direct Depth									Depth Discontinuity								Depth Smoothness				
		Error ↓				Accuracy ↑					Error ↓		Accuracy ↑						Error ↓	Accuracy ↑			
		wRMSE	*wRMSLE*	*wAbsRel*	*wSqRel*	$\delta^{ico^6}_{1.05}$	$\delta^{ico^6}_{1.1}$	$\delta^{ico^6}_{1.25}$	$\delta^{ico^6}_{1.25^2}$	$\delta^{ico^6}_{1.25^3}$	dbe^{acc}	dbe^{comp}	$prec_{0.25}$	$prec_{0.5}$	$prec_1$	$rec_{0.25}$	$rec_{0.5}$	rec_1	*RMSE*^{v}	$\alpha_{11.25^\circ}$	$\alpha_{22.5^\circ}$	α_{30°
Pnas	$\mathcal{L}_1$	0.5606	0.0854	0.1328	0.1196	32.69%	56.94%	85.12%	95.38%	97.95%	2.6542	5.7303	38.73%	30.26%	23.58%	18.74%	10.48%	8.48%	20.12	53.88%	69.81%	75.65%
	$\mathcal{L}_{cosine}$	0.5622	0.0858	0.1338	0.1317	34.49%	58.06%	85.52%	95.44%	97.88%	2.7194	5.4964	36.16%	27.76%	22.22%	21.48%	13.55%	10.02%	15.70	67.14%	77.37%	81.05%
	$\mathcal{L}_{grad}$	0.5374	0.0822	0.1276	**0.1146**	35.68%	59.54%	86.41%	95.72%	98.04%	**2.5008**	5.4548	**40.91%**	**32.05%**	25.27%	22.49%	12.17%	9.07%	18.12	59.56%	73.24%	78.30%
	$\mathcal{L}_{comb}$	**0.5367**	**0.0811**	**0.1259**	0.1153	**36.44%**	**60.52%**	**86.80%**	**95.83%**	**98.11%**	2.5119	**5.3501**	39.83%	31.59%	**27.01%**	**23.53%**	**14.42%**	**10.98%**	**15.26**	**67.73%**	**77.99%**	**81.67%**
	$\mathcal{L}_{vnl}$	0.5403	0.0815	0.1280	0.1183	35.43%	59.72%	86.58%	95.79%	98.11%	2.5141	5.3893	40.14%	31.77%	24.47%	22.14%	12.69%	8.74%	15.57	66.61%	77.34%	81.23%
Unet	$\mathcal{L}_1$	0.4834	0.2361	0.1211	0.0913	35.18%	58.24%	86.80%	96.45%	98.43%	1.4011	4.3152	57.59%	58.00%	53.85%	38.74%	31.57%	24.31%	24.66	36.80%	60.60%	69.73%
	$\mathcal{L}_{cosine}$	0.4736	**0.0906**	0.1217	0.0891	32.65%	58.04%	87.40%	96.68%	98.61%	1.4513	5.0455	55.35%	52.16%	46.01%	39.36%	30.01%	21.69%	**15.80**	**63.10%**	**77.60%**	**82.38%**
	$\mathcal{L}_{grad}$	0.4659	0.5186	0.1209	0.0833	35.25%	58.79%	87.33%	96.56%	98.45%	1.3305	4.0582	63.27%	**63.13%**	**56.54%**	40.39%	32.47%	23.37%	19.52	52.23%	70.40%	76.91%
	$\mathcal{L}_{comb}$	0.4630	0.1690	0.1222	0.0847	34.79%	58.21%	87.08%	96.63%	98.66%	1.3077	4.2080	**63.31%**	61.74%	54.96%	39.38%	30.27%	22.00%	16.19	61.01%	76.18%	81.45%
	$\mathcal{L}_{vnl}$	**0.4520**	0.1300	**0.1147**	**0.0811**	**36.68%**	**60.59%**	**88.31%**	**96.96%**	**98.73%**	**1.2699**	**3.8876**	58.97%	57.54%	51.85%	**43.96%**	**36.69%**	**28.59%**	16.02	61.80%	76.58%	81.70%
DenseNet	$\mathcal{L}_1$	0.5441	0.6872	0.1348	0.1144	34.34%	57.10%	84.73%	95.28%	97.69%	2.3690	5.5135	40.40%	36.07%	28.78%	20.45%	11.54%	8.05%	21.08	49.98%	68.29%	74.78%
	$\mathcal{L}_{cosine}$	0.5361	**0.0822**	0.1239	0.1034	34.98%	59.34%	86.36%	95.94%	**98.13%**	2.3486	5.3702	41.01%	35.45%	29.10%	22.80%	14.19%	9.39%	**15.97**	**64.92%**	**76.91%**	81.15%
	$\mathcal{L}_{grad}$	**0.5202**	0.4655	0.1304	0.1045	32.68%	57.59%	85.69%	95.85%	98.06%	2.0789	5.2159	47.01%	40.61%	33.32%	23.68%	13.71%	9.35%	18.90	56.86%	71.79%	77.23%
	$\mathcal{L}_{comb}$	0.5209	0.1982	**0.1209**	**0.1013**	35.97%	**60.41%**	**87.02%**	**95.96%**	98.09%	2.0628	5.0977	**47.16%**	**40.77%**	**35.20%**	**26.09%**	**16.87%**	**12.21%**	15.98	64.58%	76.86%	**81.20%**
	$\mathcal{L}_{vnl}$	0.5232	0.7560	0.1258	0.1030	**36.28%**	60.04%	86.61%	95.66%	97.74%	**2.0525**	**5.0931**	44.81%	40.14%	32.30%	25.21%	15.71%	10.33%	16.51	63.43%	76.02%	80.53%
ResNet	$\mathcal{L}_1$	0.5500	0.1922	0.1394	0.1186	30.59%	54.17%	84.07%	95.47%	98.03%	2.4386	5.7688	39.10%	31.69%	23.28%	20.92%	10.24%	6.32%	22.83	44.68%	64.51%	72.02%
	$\mathcal{L}_{cosine}$	0.5435	**0.0864**	0.1364	0.1194	**34.77%**	56.32%	84.29%	95.64%	98.11%	2.6918	5.7928	38.35%	32.13%	26.82%	21.88%	12.61%	8.71%	**16.37**	**64.24%**	**76.30%**	**80.63%**
	$\mathcal{L}_{grad}$	0.5475	0.2976	0.1387	0.1151	32.43%	54.46%	83.76%	95.37%	97.97%	2.4112	5.7959	41.87%	33.23%	21.60%	21.31%	9.27%	4.95%	20.50	52.77%	68.97%	75.00%
	$\mathcal{L}_{comb}$	**0.5294**	0.1365	0.1374	0.1127	32.03%	55.31%	84.74%	95.81%	**98.21%**	2.2393	5.3796	44.10%	36.70%	27.44%	22.91%	12.23%	7.20%	16.63	63.09%	75.70%	80.20%
	$\mathcal{L}_{vnl}$	0.5324	0.3320	**0.1301**	**0.1070**	33.60%	**57.50%**	**85.20%**	**95.83%**	98.07%	**2.1335**	**5.1866**	**45.00%**	**38.70%**	**30.85%**	**24.88%**	**14.43%**	**9.28%**	17.07	61.99%	75.22%	79.91%
ResNet$_{skip}$	$\mathcal{L}_1$	0.5041	0.2924	0.1259	0.0977	34.10%	57.64%	86.05%	96.13%	98.30%	1.5462	4.7640	49.48%	47.23%	43.31%	32.86%	23.57%	16.63%	22.30	44.07%	65.82%	73.55%
	$\mathcal{L}_{cosine}$	0.5024	0.1207	0.1208	0.0958	37.15%	59.61%	87.03%	96.34%	98.35%	1.6012	4.7078	52.83%	49.23%	41.05%	32.03%	23.82%	16.75%	15.76	63.32%	77.05%	81.83%
	$\mathcal{L}_{grad}$	**0.4754**	0.3274	0.1183	0.0905	36.23%	60.44%	87.96%	**96.62%**	98.45%	1.5013	4.4831	56.27%	**54.26%**	**47.88%**	33.96%	23.52%	16.07%	18.72	55.00%	71.76%	77.82%
	$\mathcal{L}_{comb}$	0.4788	**0.0927**	**0.1166**	**0.0893**	36.20%	**60.64%**	**87.99%**	96.62%	**98.49%**	1.4883	4.5346	**57.34%**	54.11%	47.57%	33.99%	24.30%	16.37%	**15.27**	**64.18%**	**77.57%**	**82.27%**
	$\mathcal{L}_{vnl}$	0.4923	0.1095	0.1197	0.0941	**37.55%**	60.43%	87.23%	96.42%	98.46%	**1.4629**	**4.1408**	54.99%	51.98%	45.40%	**35.29%**	**25.22%**	**17.68%**	15.67	63.28%	77.05%	81.94%

Table 2: Direct depth performance using spherical and conventional metrics. Bottom part results are the same as those presented in Table 2 of the original document. Top part are the corresponding results from Table 1 of the original manuscript.

Model	Depth Error ↓				Depth Accuracy ↑				
	RMSE	*RMSLE*	*AbsRel*	*SqRel*	$\delta_{1.05}$	$\delta_{1.1}$	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
Pnas^{comb}	0.4613	**0.0740**	0.1143	0.0892	38.56%	**63.31%**	88.70%	96.68%	98.62%
Unet^{vnl}	**0.3967**	0.1182	**0.1095**	**0.0672**	**38.62%**	62.16%	89.08%	**97.35%**	**99.03%**
DenseNet^{comb}	0.4490	0.2565	0.1129	0.0806	38.30%	63.02%	88.56%	96.66%	98.54%
ResNet^{comb}	0.4573	0.1200	0.1272	0.0894	34.53%	57.97%	86.26%	96.56%	98.71%
ResNet$^{comb}_{skip}$	0.4165	0.0843	0.1102	0.0722	36.71%	61.92%	**89.17%**	97.24%	98.90%
	wRMSE	*wRMSLE*	*wAbsRel*	*wSqRel*	$\delta^{ico^6}_{1.05}$	$\delta^{ico^6}_{1.1}$	$\delta^{ico^6}_{1.25}$	$\delta^{ico^6}_{1.25^2}$	$\delta^{ico^6}_{1.25^3}$
Pnas^{comb}	0.5367	**0.0811**	0.1259	0.1153	36.44%	60.52%	86.80%	95.83%	98.11%
Unet^{vnl}	**0.4520**	0.1300	**0.1147**	**0.0811**	**36.68%**	60.59%	**88.31%**	**96.96%**	**98.73%**
DenseNet^{comb}	0.5209	0.1982	0.1209	0.1013	35.97%	60.41%	87.02%	95.96%	98.09%
ResNet^{comb}	0.5294	0.1365	0.1374	0.1127	32.03%	55.31%	84.74%	95.81%	98.21%
ResNet$^{comb}_{skip}$	0.4788	0.0927	0.1166	0.0893	36.20%	**60.64%**	87.99%	96.62%	98.49%

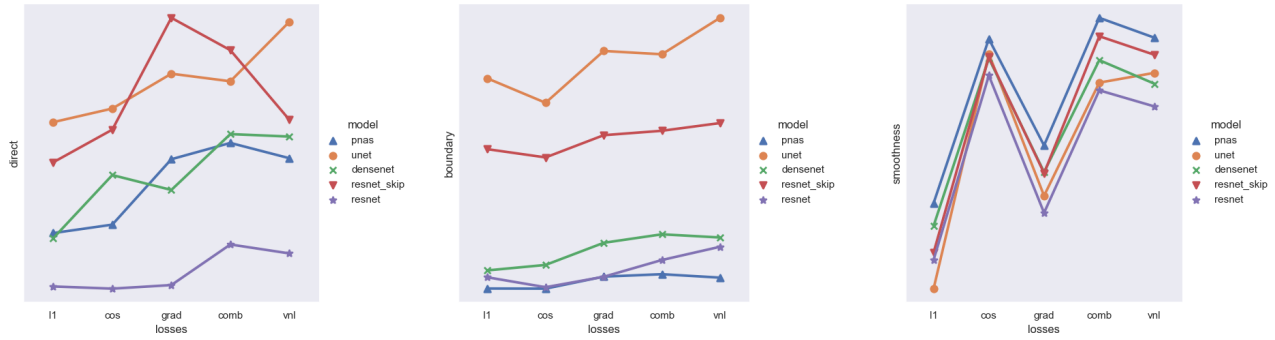


Figure 1: Performance indicators (higher is better) of different loss functions per model in three different axis. From left to right: depth indicator i_d, boundary indicator i_b and, smoothness indicator i_s.

Similarly, the differences between ResNet and ResNet$_{skip}$, attributed to the addition of the skip connections are apparent across all samples.

Nonetheless, Pnas better captures the global context as seen in Figure 6, where the scene's dominant planar surfaces are better preserved by it than UNet.

Table 3: Direct depth performance metrics across different variations of DenseNet and Pnas.

			Depth Error ↓				Depth Accuracy ↑
model	pretrained	$\mathcal{L}$	*RMSE*	*RMSLE*	*AbsRel*	*SqRel*	$\delta_{1.25}$
DenseNet	✗	$\mathcal{L}_1$	0.4672	0.5580	0.1223	0.0896	86.72%
	✓	$\mathcal{L}_1$	**0.4072**	**0.3194**	**0.1140**	**0.0694**	**88.91%**
	✓	$\mathcal{L}_{log}$	0.5597	0.5720	0.1528	0.4475	80.48%
	✓	$\mathcal{L}_{berHu}$	0.4532	0.3754	0.1228	0.0852	86.68%
Pnas	✗	$\mathcal{L}_1$	0.4817	0.0780	0.1213	0.0933	87.25%
	✓	$\mathcal{L}_1$	**0.3998**	**0.0634**	**0.0975**	**0.0661**	**91.91%**
	✓	$\mathcal{L}_{log}$	0.4135	0.0656	0.0999	0.0697	91.09%
	✓	$\mathcal{L}_{berHu}$	0.4059	0.0643	0.0992	0.0666	91.56%

Figures 7, 8, 9 demonstrate qualitative results in GV2 *tiny* split for the UNet, Pnas, and $\text{ResNet}_{\text{skip}}$ architectures respectively. Apart from the predicted point cloud we visualise the *c2c* error on the ground truth point cloud, with a blue-green-red colormap denoting the error's magnitude.

Finally, Figures 10 and 11 offer qualitative results of our best performing method in real world, in-the-wild, data captures. We also qualitatively compare our predictions with a state-of-the-art 360^o depth estimation model (i.e. BiFuse [8]). It is worth highlighting that even the two of the three 360^o images are captured by a panorama camera, the last two images are captured by a smartphone camera, and as such there are artifacts. Yet, it seems that this does not greatly affect the performance of models. The UNet model produces higher quality depth estimates than BiFuse, albeit trained only on the train split of M3D, while the publicly available BiFuse model, as reported in UniFuse [3], has been trained on the *entire* M3D dataset.

References

[1] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986. 5

[2] Marcela Carvalho, Bertrand Le Saux, Pauline Trouvé-Peloux, Andrés Almansa, and Frédéric Champagnat. On regression losses for deep depth estimation. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2915–2919. IEEE, 2018. 1

[3] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360 panorama depth estimation. *arXiv preprint arXiv:2102.03550*, 2021. 3

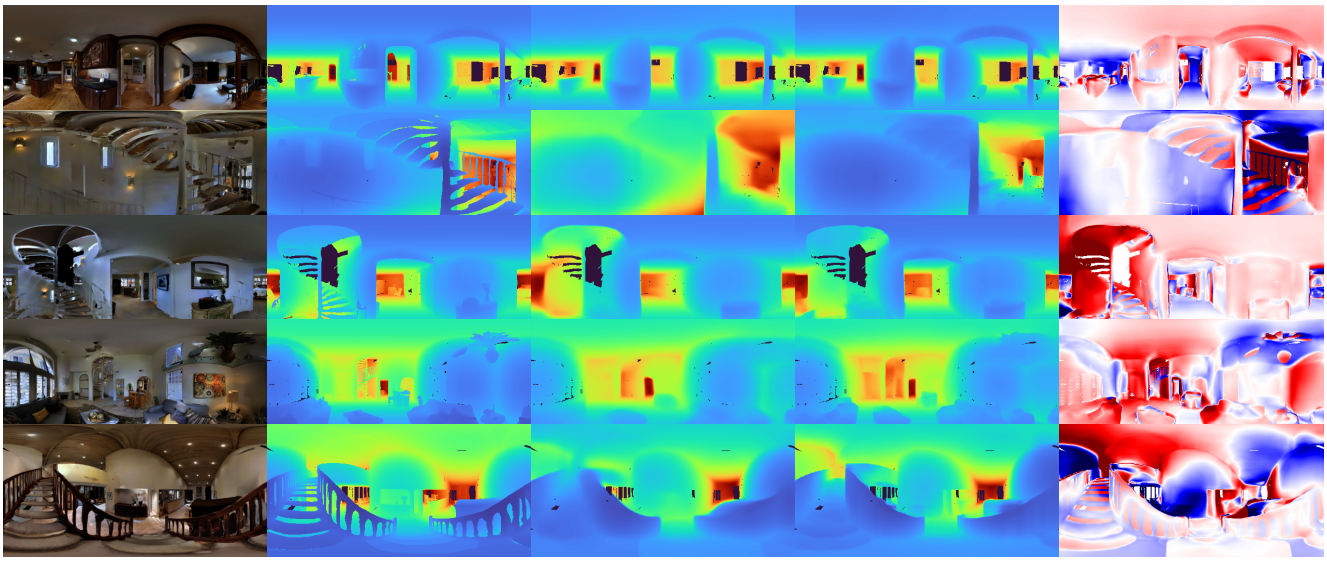
[4] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 6

[5] 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. 1

[6] Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

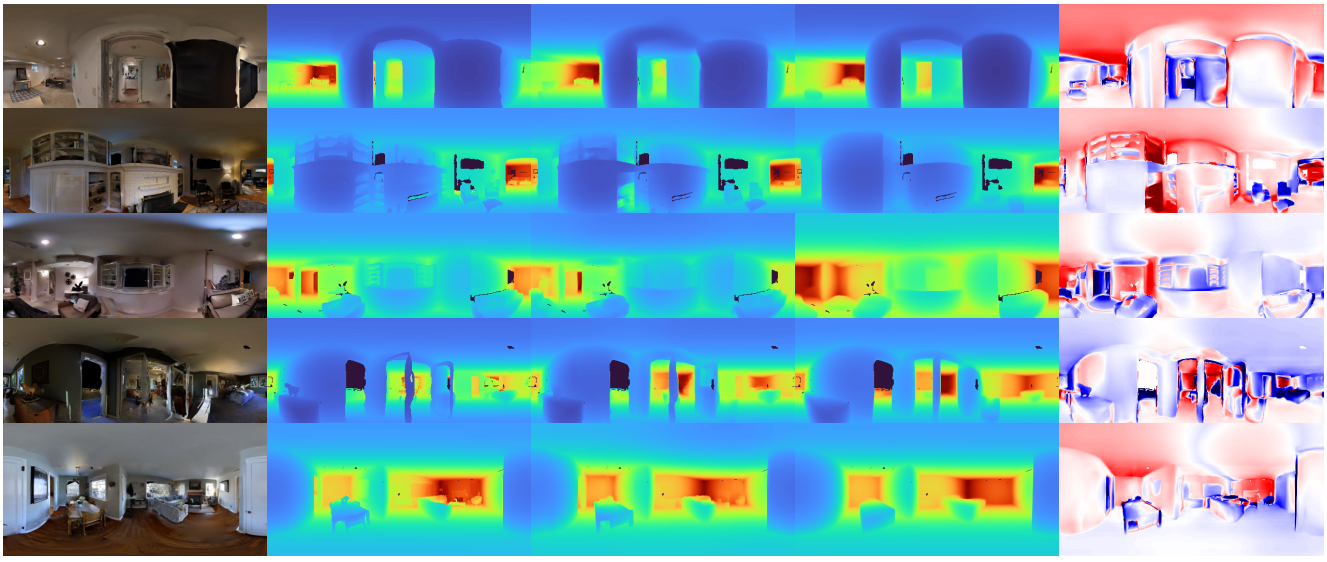
[7] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. *arXiv preprint arXiv:2011.11498*, 2020. 1

[8] 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. 3, 11



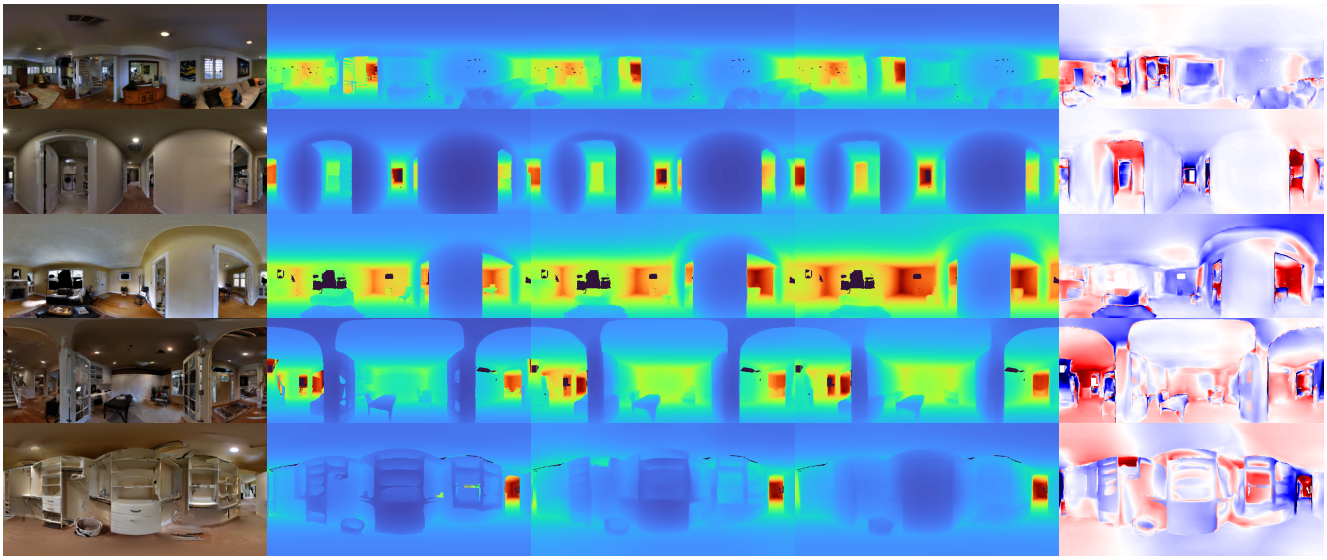
(a) Color (b) Ground Truth Depth (c) ResNet Depth (d) $\text{ResNet}_{\text{skip}}$ Depth (e) Depth Advantage

Figure 2: Qualitative comparison between the ResNet and $\text{ResNet}_{\text{skip}}$ architectures. On the right the advantage visualization shows with blue color the areas where the former performs better, and with red color the areas where the latter performs better. The color magnitude corresponds to the MAE difference between the two models, illustrating the performance deviation between the two models. The addition of skip connections allows $\text{ResNet}_{\text{skip}}$ to capture finer-grained details.



(a) Color (b) Ground Truth Depth (c) Unet Depth (d) Pnas Depth (e) Depth Advantage

Figure 3: Qualitative comparison between the UNet and Pnas architectures. On the right the advantage visualization shows with blue color the areas where the former performs better, and with red color the areas where the latter performs better. The color magnitude corresponds to the MAE difference between the two models, illustrating the performance deviation between the two models. Pnas provides smoother results while it is clear that UNet is able to capture finer-grained details.



(a) Color (b) Ground Truth Depth (c) Unet Depth (d) ResNet$_{skip}$ Depth (e) Depth Advantage

Figure 4: Qualitative comparison between the UNet and ResNet$_{skip}$ architectures. On the right the advantage visualization shows with blue color the areas where the former performs better, and with red color the areas where the latter performs better. The color magnitude corresponds to the MAE difference between the two models, illustrating the performance deviation between the two models.

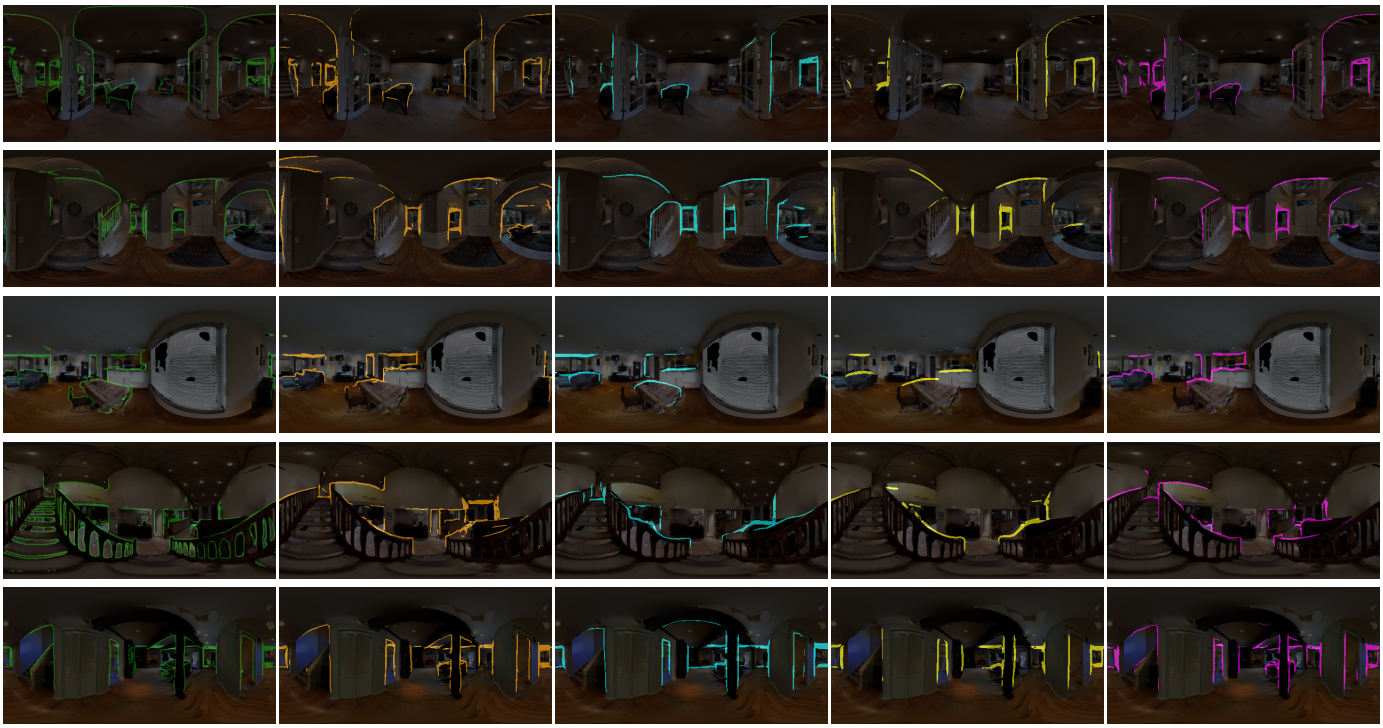


Figure 5: Boundary preservation qualitative comparison between the UNet, Pnas, ResNet and ResNet$_{skip}$ models. Boundaries are extracted by applying a Canny edge detector [1] with predefined thresholds on normalized predicted depth maps, and then are blended with the original color panorama. From left to right: **i)** GT depth (green), **ii)** UNet (orange), **iii)** Pnas (cyan), **iv)** ResNet (yellow), and **v)** ResNet$_{skip}$ (magenta).

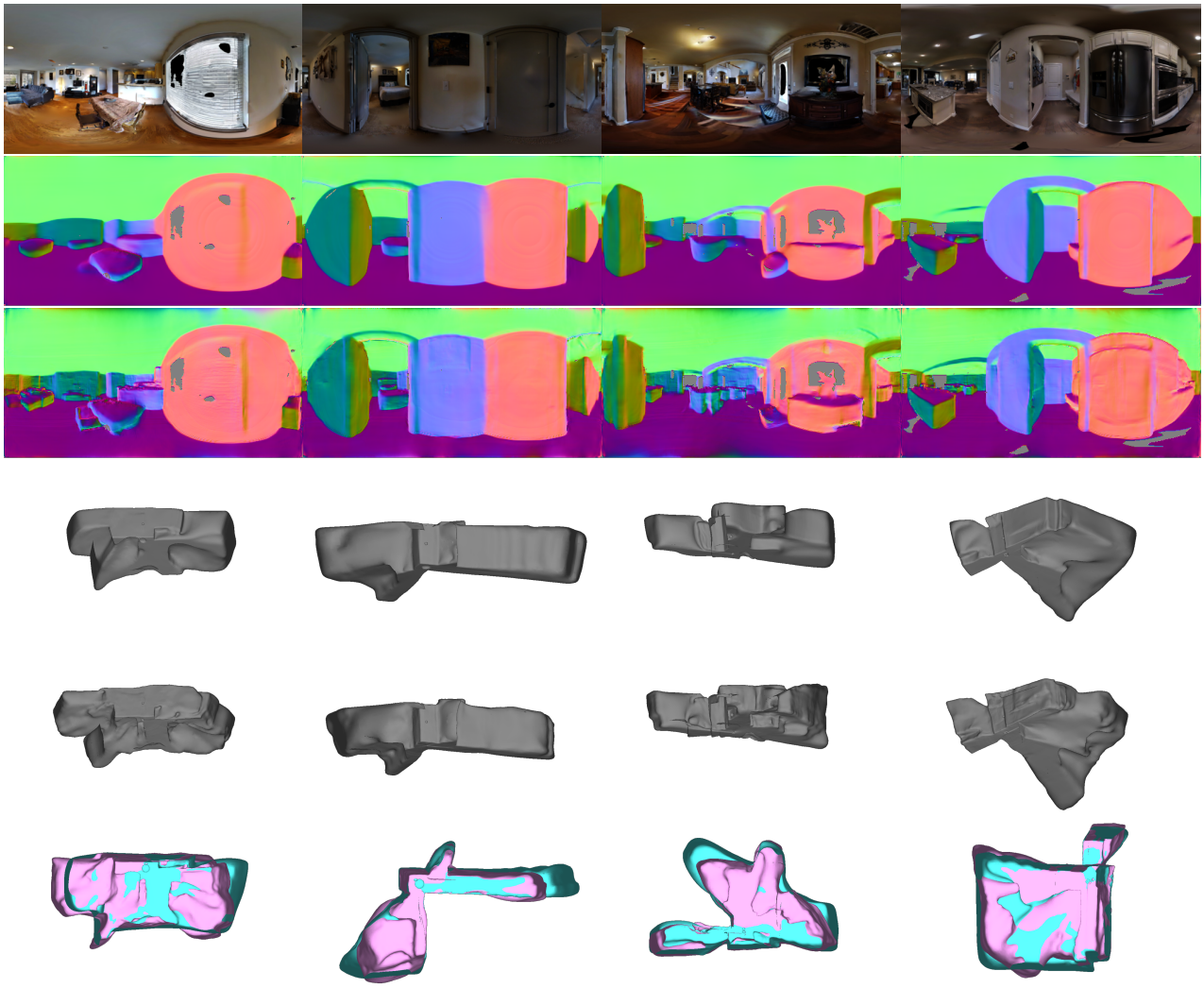


Figure 6: Qualitative comparison of the Pnas and UNet models in surface reconstruction. From top to bottom: **i)** input color panorama, **ii)** Pnas normal map from the estimated depth map, **iii)** UNet normal map, **iv)** Pnas Screened Poisson Surface Reconstruction [4] 3D surface reconstruction, **v)** UNet 3D surface reconstruction, **vi)** overlaid Pnas (cyan) and UNet (pink) 3D surface reconstructions from birds eye view.

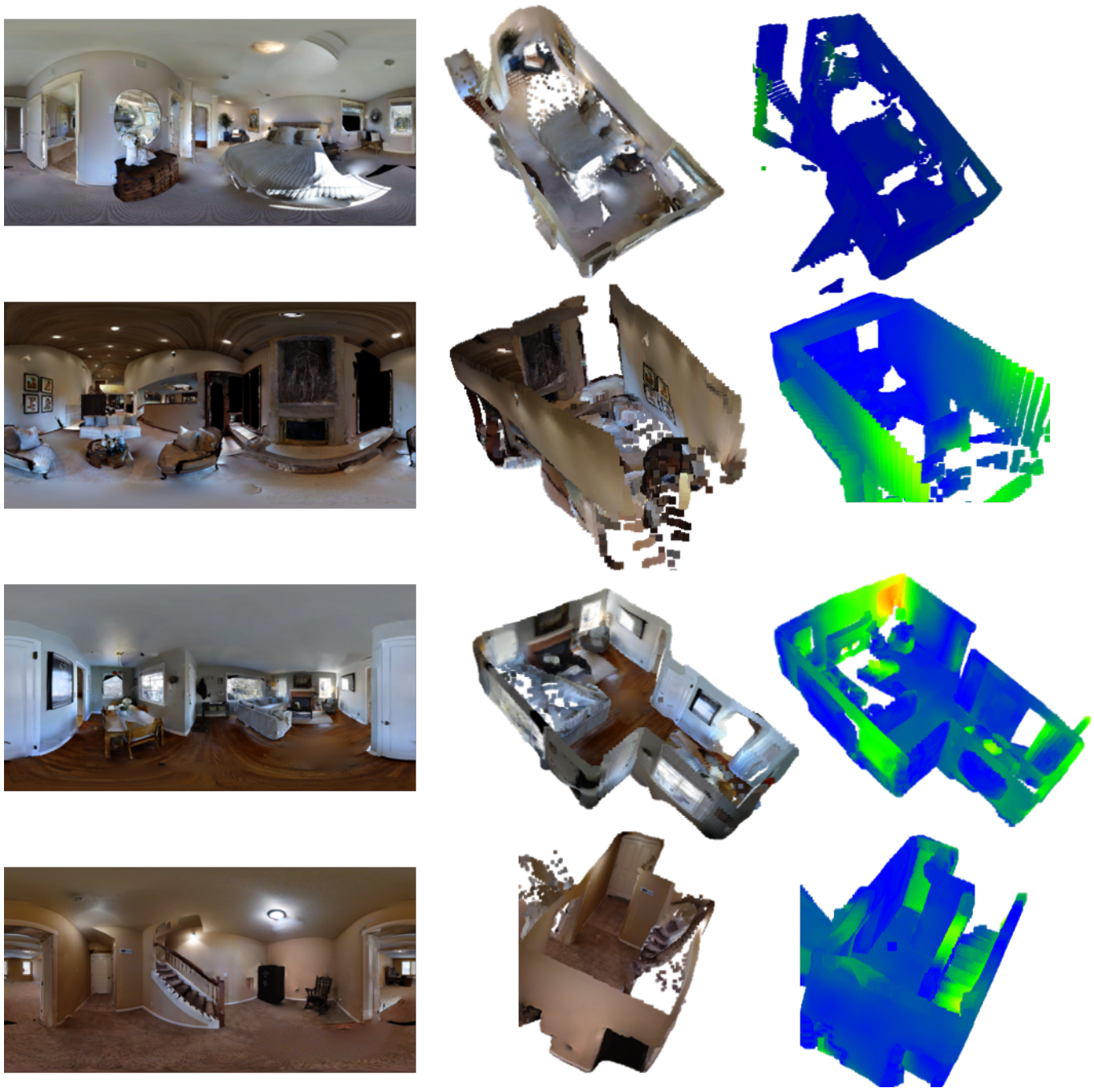


Figure 7: UNet qualitative results. From left to right: **i)** Input color panorama, **ii)** colored predicted point cloud, and **iii)** heatmap visualization of the *c2c* error on the ground truth point cloud.

Figure 8: Pnas qualitative results. From left to right: **i)** Input color panorama, **ii)** colored predicted point cloud, and **iii)** heatmap visualization of the *c2c* error on the ground truth point cloud.

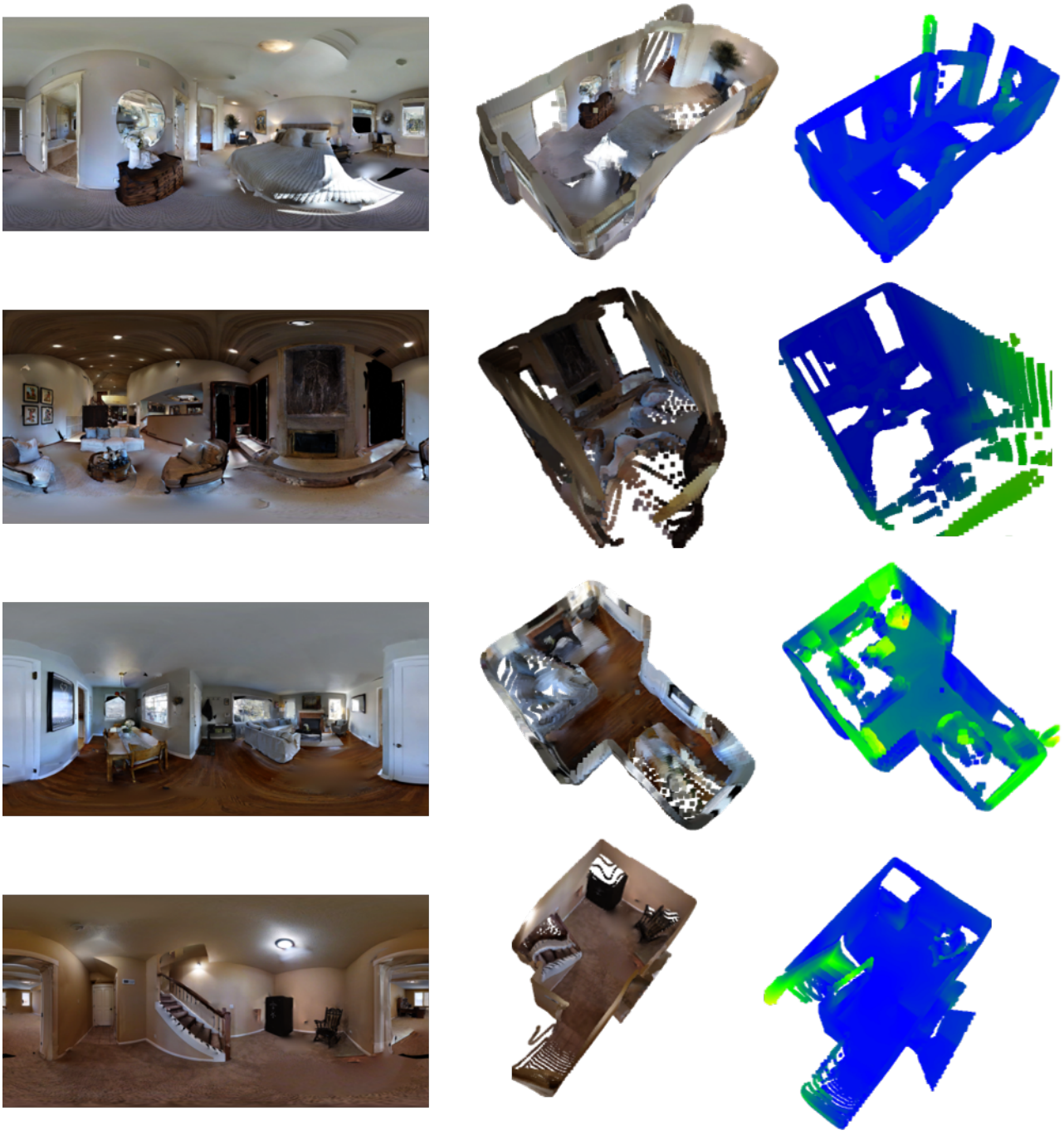


Figure 9: ResNet$_{skip}$ qualitative results. From left to right: **i)** Input color panorama, **ii)** colored predicted point cloud, and **iii)** heatmap visualization of the *c2c* error on the ground truth point cloud.

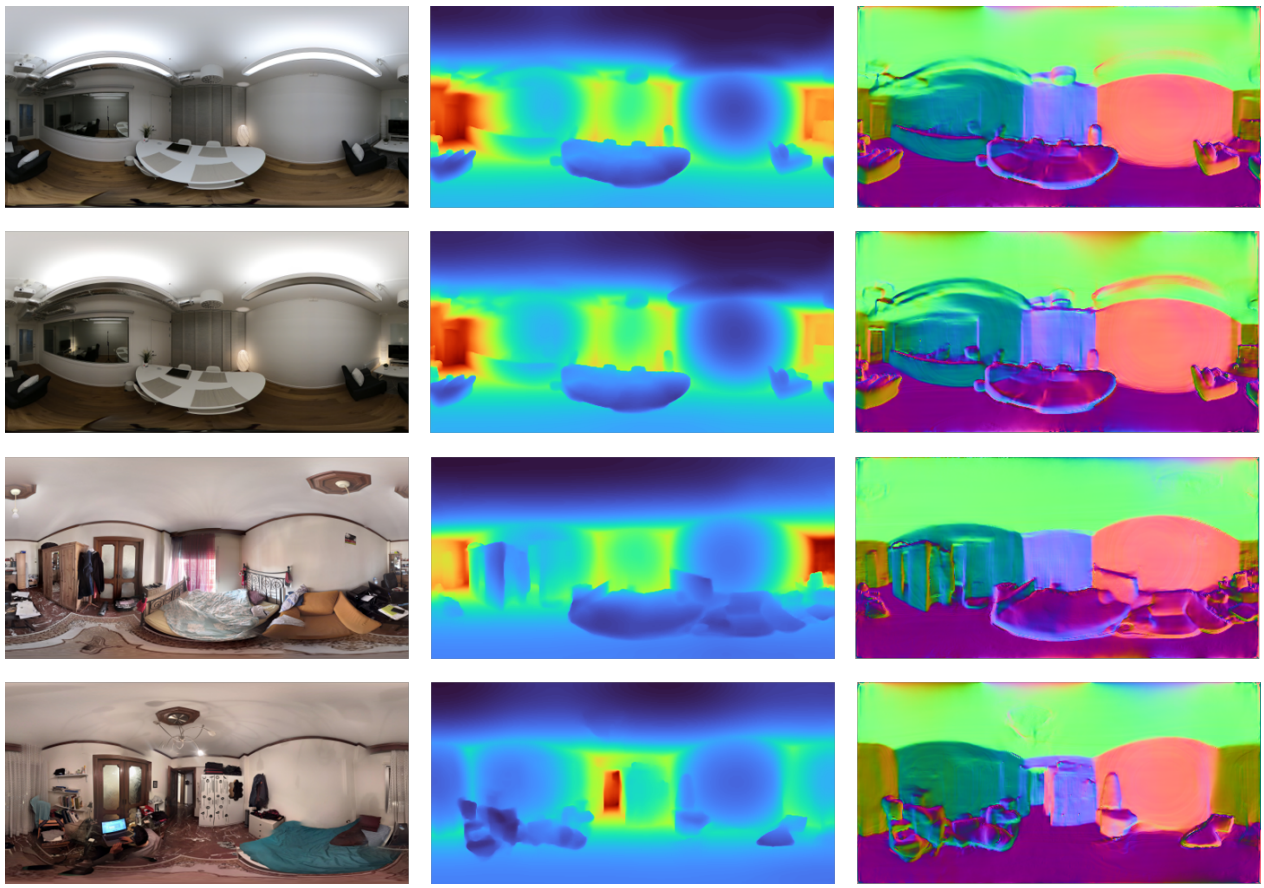


Figure 10: Qualitative results using the UNet model applied to in-the-wild real data captures. The top two rows are captures with a 360^o camera, while the bottom two rows are stitched panoramas from a mobile phone. From left to right: **i)** Input color panorama, **ii)** predicted depth, and **iii)** normals derived from the predicted depth.

Figure 11: Qualitative results using in-the-wild data. On the left the input color panoramas are depicted. The two top rows are captured with a 360^o camera, while the bottom two rows are stitched panoramas from a mobile phone. The colored point clouds of the predicted depths from our UNet model (middle) and BiFuse [8] (right). Ceilings have been removed for visualization purposes.