UniDAC: Universal Metric Depth Estimation for Any Camera

Supplementary Material

7. Data

7.1. Training Data

Tab. 5 provides an overview of the training datasets. In addition to the training datasets utilized in DAC [19], we add Argoverse2 and A2D2 to balance the indoor and outdoor distribution in the training set.

We observe that out of seven cameras in Argoverse2, the front camera's aspect ratio is different than the rest of the six cameras. Specifically, the resolution (H×W) is 1550×2048 while the rest of the cameras have a resolution of 2048×1550. Since the aspect ratio of the front camera is less than one, we omit images from the front camera for our training. Furthermore, Argoverse2 contains 1.5M samples, and to prevent introducing bias to outdoor data, we randomly sample 300K image-depth pairs to complete our training set.

A2D2 consists of six cameras, namely, front-center, front-right, front-left, side-right, side-left, and rear-center, with corresponding LiDAR acquisitions. We exclude the images from the rear-center cameras and add 350K images obtained from the remaining five cameras to the training set.

We note that UniK3D's [41] training set consists of more than 8M samples, whereas our training set consists only 1.45M samples. Moreover, out of the 8M images in the training set of UniK3D, 72.4% are perspective images, 27.27% are fisheye images, and 0.33% are ERP images. In contrast, 100% of UniDAC's training set consists of perspective images as seen in Tab. 5

Table 5. **Training Datasets.** List of training datasets with the following attributes: number of images, scene type, and acquisition method. [Key: Syn=Synthetic, Rec=Mesh reconstruction]

Dataset	#Images	Scene	Acquisition
HM3D-tiny [43]	310K	Indoor	Rec
Taskonomy-tiny [68]	300K	Indoor	RGB-D
Hypersim [47]	54K	Indoor	Syn
DDAD [17]	80K	Outdoor	LiDAR
LYFT [22]	50K	Outdoor	LiDAR
Argoverse2 [59]	300K	Outdoor	LiDAR
A2D2 [16]	350K	Outdoor	LiDAR

7.2. Testing Data

Tab. 6 details the testing data used to evaluate UniDAC. We follow the testing data setup of DAC and evaluate all our baselines on them. While ScanNet++ and KITTI-360 are both fisheye datasets, they differ in their respective distortion models. ScanNet++ follows the KB [25] model while KITTI-360 follows the MEI [35] model. We adopt the lookup table provided by DAC to perform fast fisheye to ERP conversion. Pano3D-GV2 and Matterport3D datasets consist of 360° images, which are provided in the ERP images by default, and therefore we use them as is.

Table 6. **Testing Datasets.** List of testing datasets with the following attributes: camera type, scene type, and acquisition method. [Key: Rec=Mesh reconstruction]

Dataset	Cam.Type	Scene	Acquisition
ScanNet++ [63]	Fisheye	Indoor	Rec
Matterport3D [5]	ERP	Indoor	Rec
Pano3D-GV2 [2]	ERP	Indoor	Rec
KITTI-360 [31]	Fisheye	Outdoor	LiDAR
KITTI [14]	Persp.	Outdoor	LiDAR
NYU-v2 [52]	Persp.	Indoor	RGB-D
nuScenes [4]	Persp.	Outdoor	LiDAR
IBims-1 [28]	Persp.	Indoor	RGB-D

Table 7. **Effect of camera intrinsics on depth performance.** We study the impact of utilizing predicted and ground-truth {P,G} camera intrinsics during inference. [Key: **Best**, Second Best]

Method	ScanNet++			KITTI-360		
	$\delta_1 \uparrow$	A.Rel ↓	RMSE ↓	$\delta_1 \uparrow$	A.Rel ↓	RMSE ↓
UniK3D	0.651	0.253	**0.285**	0.817	0.244	**2.400**
UniDAC- P	0.894	0.110	0.274	0.706	0.198	4.397
+A2D2	0.917	0.104	0.279	0.815	0.154	4.091
UniDAC- G	0.905	0.104	0.274	0.757	0.169	4.470
+A2D2	**0.918**	**0.097**	0.277	**0.836**	**0.141**	3.977

8. Comparison with UniK3D

As mentioned in Sec. 5.2, the comparison with UniK3D [41] is not fair to UniDAC, since [41] is trained on large-FoV images. However, we note that the comparison is also unfair towards [41] since UniDAC requires ground-truth camera parameters while [41] doesn't.

For a fairer comparison, we employ AnyCalib [56], an off-the-shelf camera intrinsics estimation model, and utilize the predicted intrinsics for ERP transformations. [56] predicts intrinsics for KB [25] and UCM [25] camera models. ScanNet++ [63] follows the KB model, whereas KITTI-360 [31] follows the MEI [35] model, which is not handled by [56]. We approximate the MEI [35] model from UCM [15] by setting the distortion parameters to zero.

Tab. 7 provides the performance comparison between

Table 8. **Zero-shot evaluation on perspective datasets.** We evaluate all unified models on perspective datasets. All models are trained on a mix of indoor and outdoor datasets. [Key: **Best**]

Method	Training Size	KITTI			NYU-v2			nuScenes			IBims-1		
		$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$
Metric3Dv2	16.20M	**0.974**	**0.053**	**2.493**	0.972	0.067	0.262	0.841	0.236	9.400	0.684	0.207	0.700
UniDepth	3.83M	0.964	0.116	2.788	**0.988**	**0.052**	**0.194**	**0.846**	**0.127**	**4.560**	0.157	0.410	1.250
UniK3D	7.94M	0.833	0.159	4.323	0.899	0.133	0.400	0.840	0.189	10.830	**0.919**	**0.104**	**0.406**
DAC_U	0.79M	0.767	0.180	5.332	0.816	0.140	0.505	0.631	0.225	8.321	0.808	0.370	1.182
UniDAC	1.45M	0.872	0.122	3.784	0.934	0.093	0.354	0.801	0.151	6.335	0.845	0.129	0.577

[41] and UniDAC using predicted and ground-truth intrinsics. '*+A2D2*' denotes adding A2D2 [16] in the training data as detailed in Sec. 7.1. We observe that even under this fairer comparison, we still outperform [41] on ScanNet++ [63]. We attribute the decrease in the performance on KITTI-360 [31] to the approximation of predicted intrinsics from UCM [15] to the MEI [35] model. We believe that with better intrinsic estimation models that can handle the MEI [35] camera model, UniDAC will retain its performance even with predicted intrinsics.

Table 9. **Ablation on Encoder Weight**. D2, D3 indicate the ViT encoders have been initialized with DINOv2 and DINOv3 pre-trained weights, respectively. [Key: **Best**]

Method	ScanNet++			KITTI-360		
	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$
DAC_U- D2	0.368	0.305	0.939	0.334	0.318	6.872
DAC_U- D3	0.707	0.211	0.471	0.428	0.342	5.305
UniDAC- D2	0.533	0.242	0.703	0.445	0.287	6.924
UniDAC- D3	**0.792**	**0.140**	**0.396**	**0.622**	**0.239**	**5.057**

9. Evaluation on Perspective Datasets

We compare UniDAC against our baselines on four perspective datasets, KITTI [14], NYU-v2 [52], IBims-1 [28], and nuScenes [4]. While [14, 28, 52] provide artifact-free depthmaps in their official dataset, we utilize [69] to estimate artifact-free depthmaps for [4]. We observe from Tab. 8 that UniDAC outperforms UniK3D and DAC on two important perspective benchmarks, namely, KITTI and NYU-v2, demonstrating the generalization capability of UniDAC even in perspective datasets, beyond large-FoV datasets.

Table 10. **Ablation on Shift Estimation.** We study the impact of depth-guided shift map estimation. [Key: **Best**]

Shift	ScanNet++			KITTI-360		
	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$
$t \in \mathbb{R}$	0.792	0.140	0.396	0.622	0.239	5.057
$\mathbf{T} \in \mathbb{R}^{H \times W}$	**0.798**	**0.139**	**0.393**	**0.630**	**0.235**	**4.985**

10. Ablation on Encoder Weights

Tab. 9 evaluates the effect of initializing encoders $\mathcal{E}$ with different pre-trained weights on the model performance. We train DAC_U and UniDAC using DINOv2 and DINOv3 encoders on HM3D and DDAD datasets. While DAC's proposed framework is compatible with any depth estimation model, they use iDisc [39] for its simplicity and effectiveness. iDisc requires multi-scale features from the encoder for its pipeline. Since DINO features are at a downscaled resolution, we apply consecutive upsampling via transpose convolutions to obtain multi-resolution features.

One of the major differences between DINOv2 and DINOv3 is the positional embedding scheme. DINOv2 uses additive absolute positional embedding, whereas DINOv3 utilizes 2D-RoPE [21]. We do not modify the positional embedding scheme of the DINO encoders, and thus the overall performance of DAC and UniDAC is affected by the compatibility of the positional embeddings with the respective frameworks and the task of large-FoV depth estimation.

We observe that utilizing DINOv3 as the encoder gives the best performance for both DAC and UniDAC. Since we train on small-FoV perspective images and test on large-FoV images, the absolute positional embedding of DINOv2 is not suitable for the task. However, the RoPE in DINOv3 offers relative positional embedding, thus facilitating generalization.

iDisc, utilized by DAC, internally uses absolute positional embedding for their proposed Internal Discretization Module. Therefore, the proposed method in UniDAC is most compatible with the DINOv3 encoder, giving the best performance. We believe the mismatch between the RoPE in DINOv3 and the absolute positional embedding in iDisc architecture to be one of the reasons for the performance gap between DAC-D3 and UniDAC-D3.

We also observe that the performance of DAC-D3 on KITTI-360 is quite low compared to UniDAC-D3, underscoring the benefit of our proposed depth-guided scale estimation module.

11. Ablation on Shift Estimation

As mentioned in Sec. 4.2, we estimate a scale map $\mathbf{S}$ instead of a 1-D scalar s to adjust for irregularities. However, we still estimated shift t as a 1-D scalar. Tab. 10 provides an ablation on estimating a shift scalar and a shift map while keeping scale estimation in the form of a scale map. Formally, we modify the architecture slightly to output $\{\mathbf{S}, \mathbf{T}\} \in \mathbb{R}^{H \times W}$. As expected, we can observe from Tab. 10 that there is a slight increase in performance by incorporating a shift map. However, the improvement from adopting the shift map is still smaller than the improvement from adopting the scale map, as seen in Tab. 3.

12. Additional Qualitative results

We provide additional qualitative results on ScanNet++ [63], Pano3D-GV2 [2], and KITTI-360 [31] for visual comparison in Fig. 7, Fig. 8 and Fig. 9 respectively.

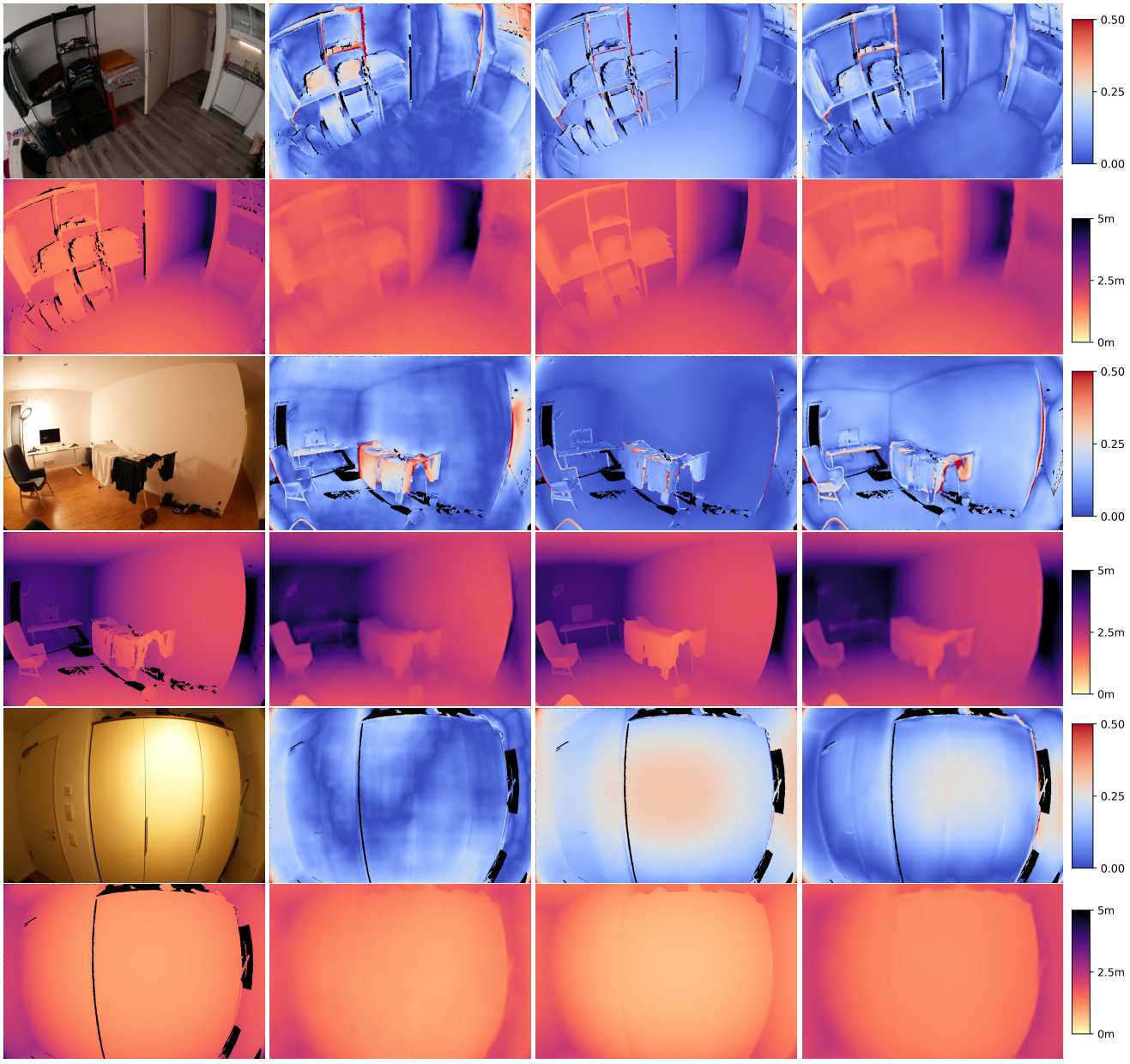
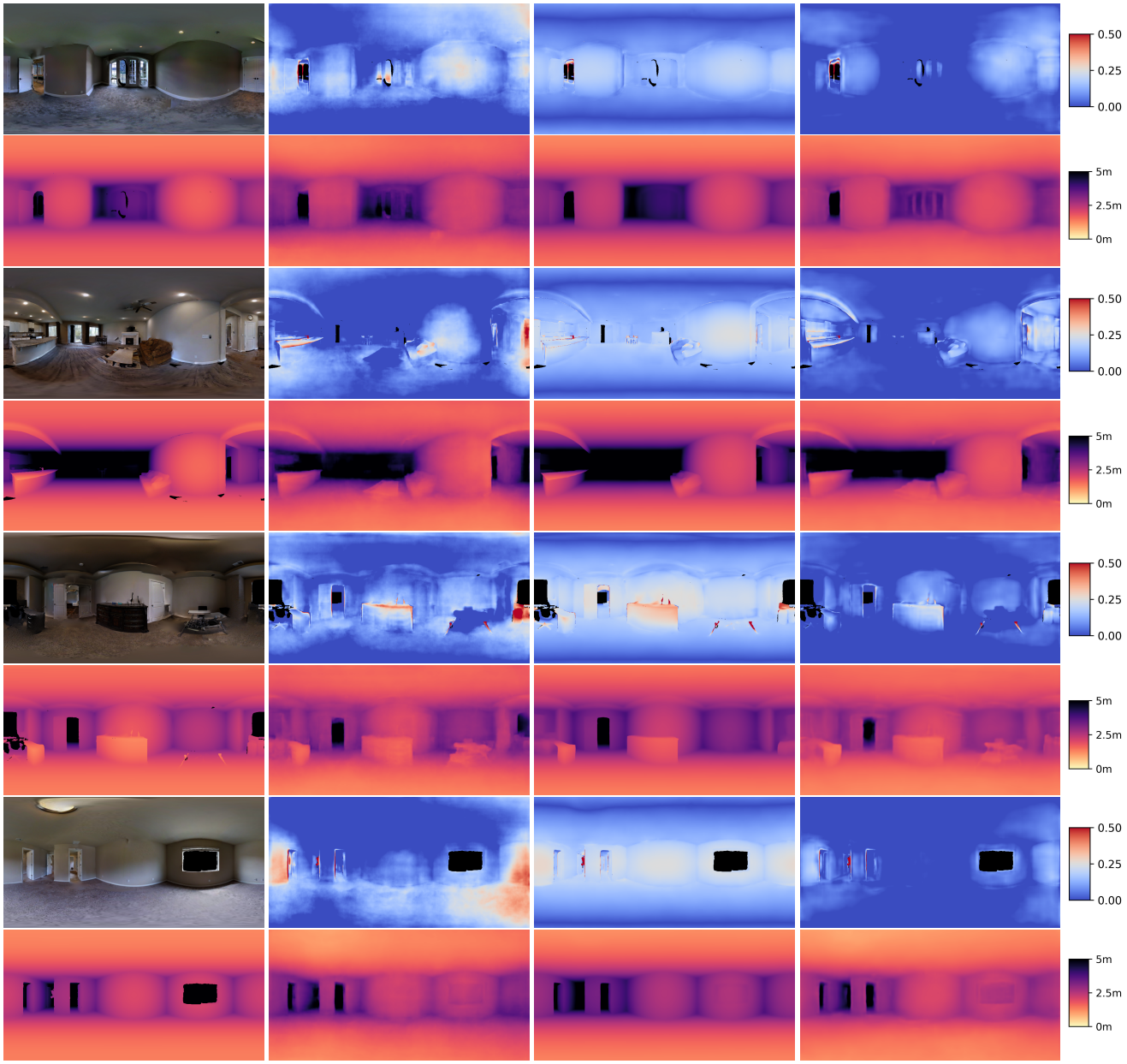
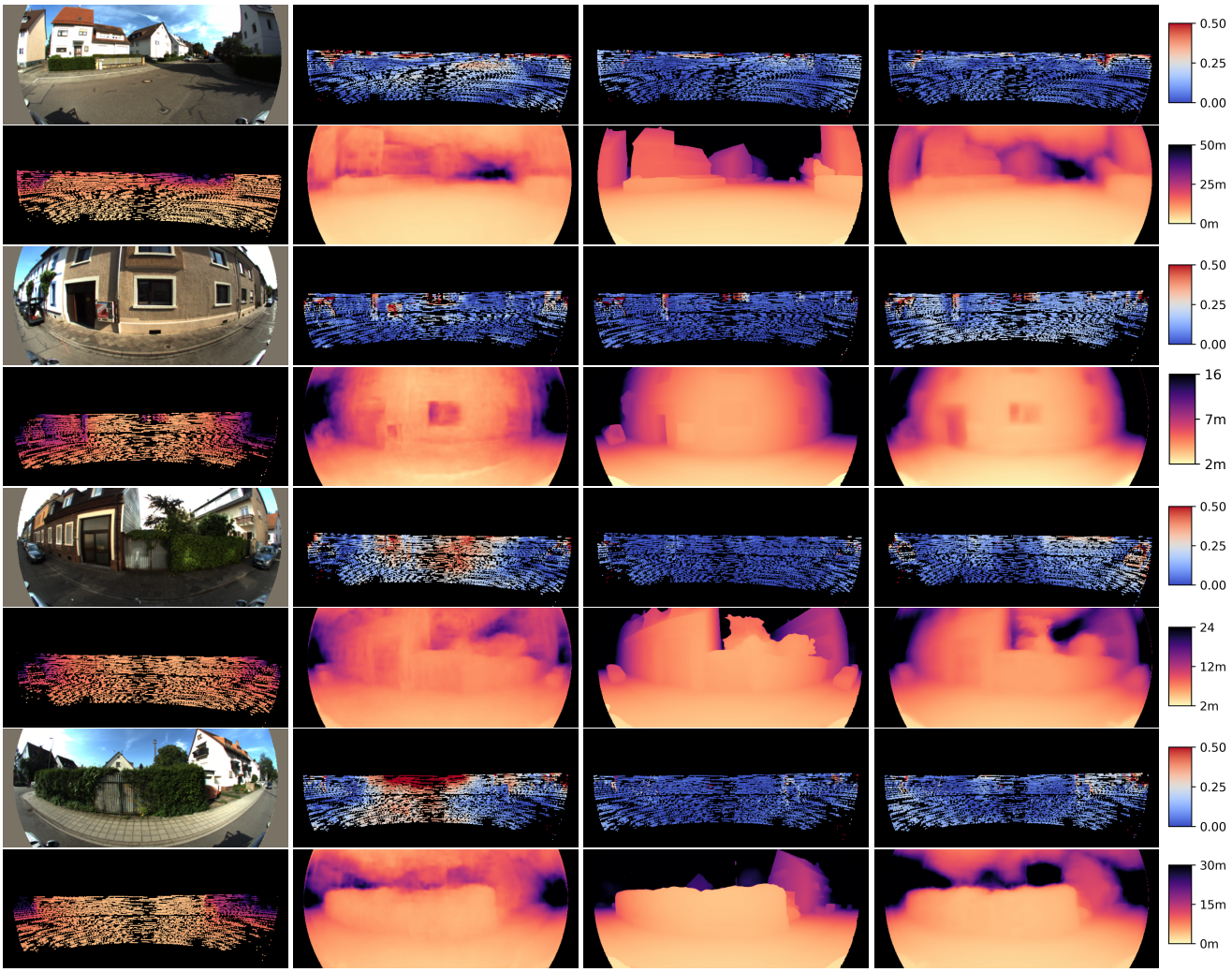


Figure 7. **Qualitative Results on ScanNet++ [63].** Every pair of consecutive rows corresponds to a single sample. Odd rows display the input RGB image, and A.Rel error between predicted and GT depth maps. Even rows display the GT depth map and predicted depth maps.



RGB & GT | DAC_U [19] | UniK3D [41] | UniDAC

Figure 8. **Qualitative Results on Pano3D-GV2 [2].** Every pair of consecutive rows corresponds to a single sample. Odd rows display the input RGB image, and A.Rel error between predicted and GT depth maps. Even rows display the GT depth map and predicted depth maps.



RGB & GT | DAC_U [19] | UniK3D [41] | UniDAC

Figure 9. **Qualitative Results on KITTI-360 [31].** Every pair of consecutive rows corresponds to a single sample. Odd rows display the input RGB image, and A.Rel error between predicted and GT depth maps. Even rows display the GT depth map and predicted depth maps.