Supplemental Material

Reproducibility Statement

All results presented in this work are reproducible using the code associated with our research, available at https://jugit.fz-juelich.de/ias-8/mdeaux.

Ethics Statement

Our research relies on publicly available datasets, ensuring transparency and reproducibility in our experiments. Additionally, for datasets obtained through agreements, such as Matterport, we adhere to the respective terms and conditions outlined in the agreements [44].

Environmental Impact

Recognizing the environmental impact of computational resources, we are mindful of the compute resources used in our experiments. The experiments, conducted on the aforementioned hardware setup, resulted in an estimated environmental impact of approximately 29.25 million mWh, equivalent to 0.0117 metric tons of carbon dioxide. This is comparable to the emissions from driving 5.1 miles in the average gasoline-powered passenger vehicle in the US [2]. For a comprehensive overview of the environmental impact of compute [49], we refer to metrics provided by platforms like the ML CO2 Impact calculator [28].

A. Training Details

As mentioned in Section 4 of the main paper, we adopt the DINOv2 training procedure [46]. In particular, we use AdamW with initial learning rate of 0.0001 and weight decay of 0.01, and a cosine scheduler with linear warmup for $1/3$ of the iterations. In total we train for 38400 steps with a batch size of 4 (2 images per 2 GPUs). For Taskonomy we use a batch size of 16 (across 8 GPUs), whereas for Matterport we train twice longer. When using our method, we duplicate the optimizer and learning rate schedulers, and scale the DPT decoder learning rates by our parameter α.

For the experiment using the Depth Anything backbone, we adapt the Depth Anything training procedure to our scheme. In particular we use an initial learning rate of 0.000161 and a cosine scheduler without linear warmup. We train for 38400 steps with a batch size of 16 (2 images per 8 GPUs).

B. Hardware Details

In our experiments, we leveraged high-performance computing (HPC) nodes equipped with 4 NVIDIA A100 GPUs with 40GB VRAM and 48 CPU cores. For most of the experiments the training process used only 2 of the GPUs, while for some we used 8 GPUs.

C. Additional ablations

C.1. Single-Label Dense Classification

Motivated by the promising outcomes observed with our MLDC task as an auxiliary task for Monocular Depth Estimation, we simplify the problem to Single-Label Dense Classification to investigate the viability of using straightforward classification datasets as auxiliaries for MDE. This involves extracting the dominant class for each image segmentation mask and output during training and computing the CrossEntropyLoss based on the dominant classes. As depicted in Table 5, the results for single-label and MLDC are comparable. This suggests that classifying the dominant class could potentially suffice as an auxiliary task for MDE. We encourage further exploration with additional single-label classification datasets, such as ImageNet [16], to validate whether they can contribute further improvements in MDE quality.

C.2. Comparison of Single and Multi Source Auxiliary Tasks

We extend our investigation to ascertain if our approach can also be applied within a single dataset, even though this is not the primary focus of our research. For SUN RGBD we use the original provided semantic segmentation labels, whereas for the other datasets we use pseudo labels generated using the approach used in PolyMax [70] for Taskonomy. Note that the quality of the pseudo-labels is not guaranteed to be high, especially when images contain objects unknown to the chosen pseudo-labeler model. This can lead to noisy labels and potentially degrade the performance. The results reported in Table 6 reveal that our method can also improve the MDE quality when only using a single dataset and pre-processing for both tasks, while the quality gains for each MDE dataset are worse compared to using multiple auxiliary data sources. On one hand, this demonstrates the versatility of our method, showing that it can provide improvements in both single and multiple sources scenarios. On the other hand, these results suggest that using multiple and diverse auxiliary sources should be preferable to ensure higher quality gains.

D. Additional qualitative results

We present additional qualitative results for each dataset to demonstrate the effectiveness of our approach. Each row should be considered individually.

Table 5. AbsRel $\times 10^4$ (↓) scores of the best task w.r.t. the DINOv2 baseline. The last column shows the result of Single-Label Dense Classification, while others are taken from Table 3. The indoor-outdoor dominated MIX6 dataset with $\alpha = 0.9$ is used for all the experiments. The best and second-best results are highlighted in **bold** and *italic*, respectively.

MDE Datasets	In MIX6	DINOv2	Aux Tasks				Gain %
			Classification	Segmentation	Reconstruction	S-Classification	
NYUv2	✗	809 ± 10	**696 ± 3**	755 ± 6	838 ± 7	*727 ± 8*	13.9
SUN RGBD	✓	1128 ± 11	*1024 ± 10*	1069 ± 9	1111 ± 5	**1007 ± 11**	10.3
Matterport3D	✗	1874 ± 19	*1728 ± 9*	1793 ± 13	1805 ± 6	**1720 ± 5**	8.2
Taskonomy	✗	1506 ± 13	**1481 ± 4**	1567 ± 12	1543 ± 6	*1491 ± 20*	1.7
DIODE In	✗	3588 ± 19	**3239 ± 26**	3451 ± 47	3368 ± 31	*3289 ± 59*	9.7
DIODE Out	✗	5820 ± 223	4965 ± 162	5085 ± 172	**4530 ± 91**	*4926 ± 121*	22.2

Table 6. AbsRel $\times 10^4$ (↓) scores on various depth datasets using different auxiliary tasks and percentage gain of the best task w.r.t. the DINOv2 baseline. For every dataset, the same dataset is used as auxiliary with $\alpha = 0.9$, either with original or pseudo labels. The best and second-best results are highlighted in **bold** and *italic*, respectively.

MDE Datasets	Pseudo Labels	DINOv2	Aux Tasks – Same Dataset Only			Gain %
			Classification	Segmentation	Reconstruction	
NYUv2	✓	*809 ± 10*	**762 ± 25**	838 ± 32	846 ± 29	5.8
SUN RGBD	✗	*1128 ± 11*	**1105 ± 32**	1161 ± 7	1167 ± 12	2
Matterport3D	✓	1874 ± 19	**1841 ± 21**	*1856 ± 17*	1864 ± 15	1.8
Taskonomy	✓	**1506 ± 13**	*1576 ± 16*	1606 ± 5	1607 ± 8	-4.6
DIODE In	✓	3588 ± 19	**3447 ± 67**	*3554 ± 90*	3620 ± 26	3.1
DIODE Out	✓	*5820 ± 223*	5918 ± 88	6040 ± 60	**5682 ± 96**	2.4

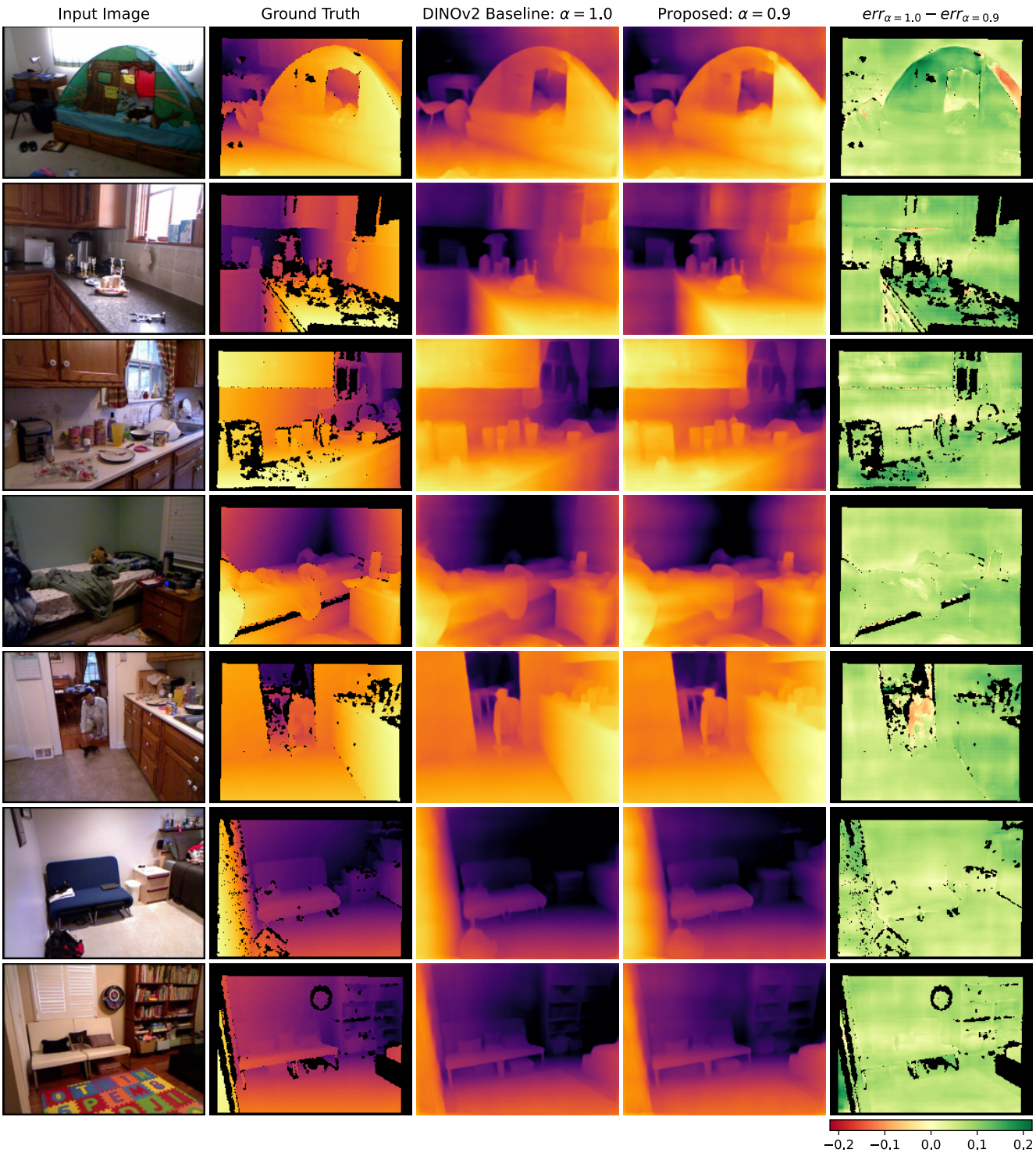

Figure 6. Results on NYU with MIX6 auxiliary MLDC task. From left to right: image and respective ground truth, baseline and our method predictions, and error difference between the last two w.r.t. to the ground truth.

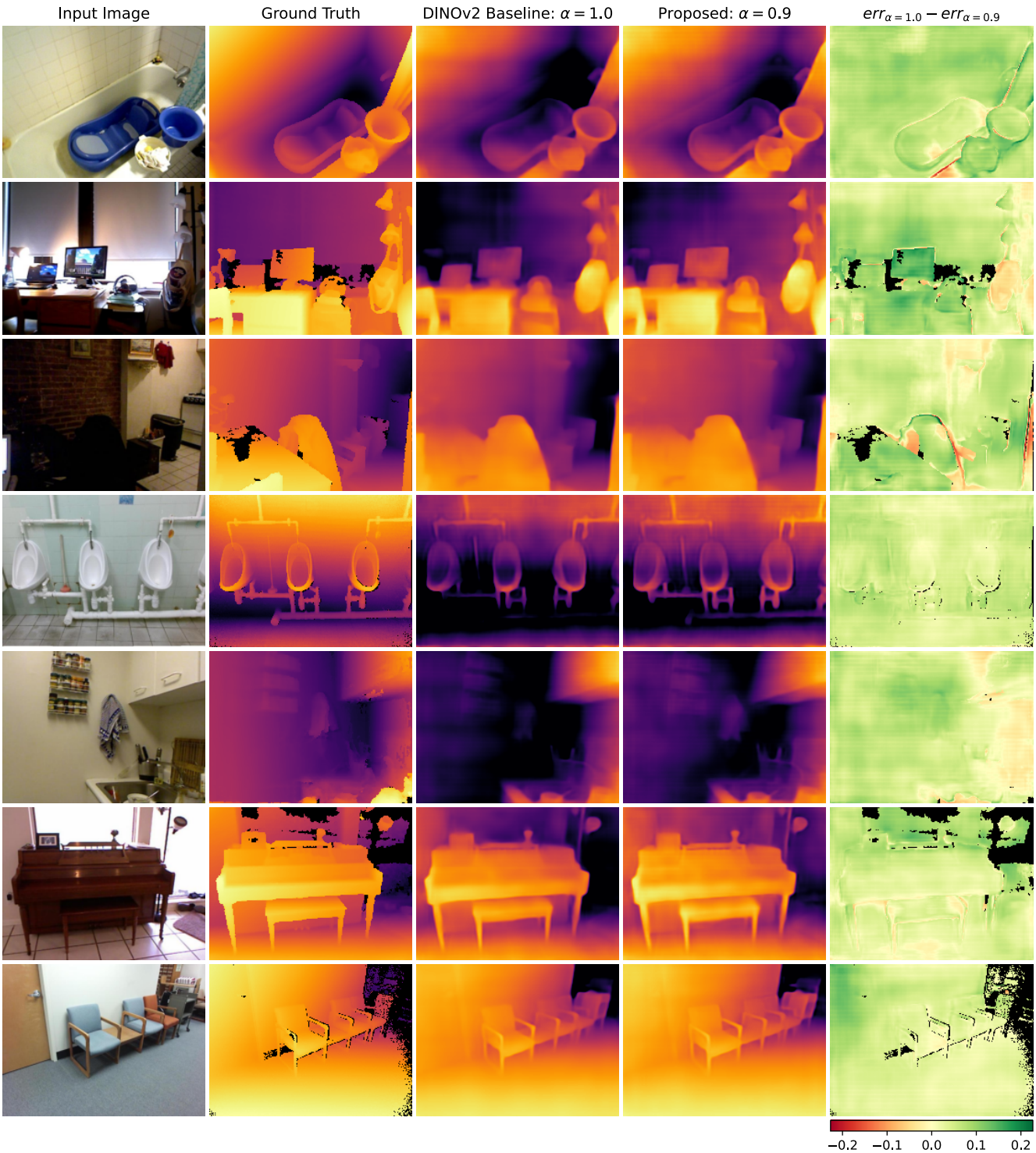

Figure 7. Results on SUNRGBD with MIX6 auxiliary MLDC task. From left to right: image and respective ground truth, baseline and our method predictions, and error difference between the last two w.r.t. to the ground truth.

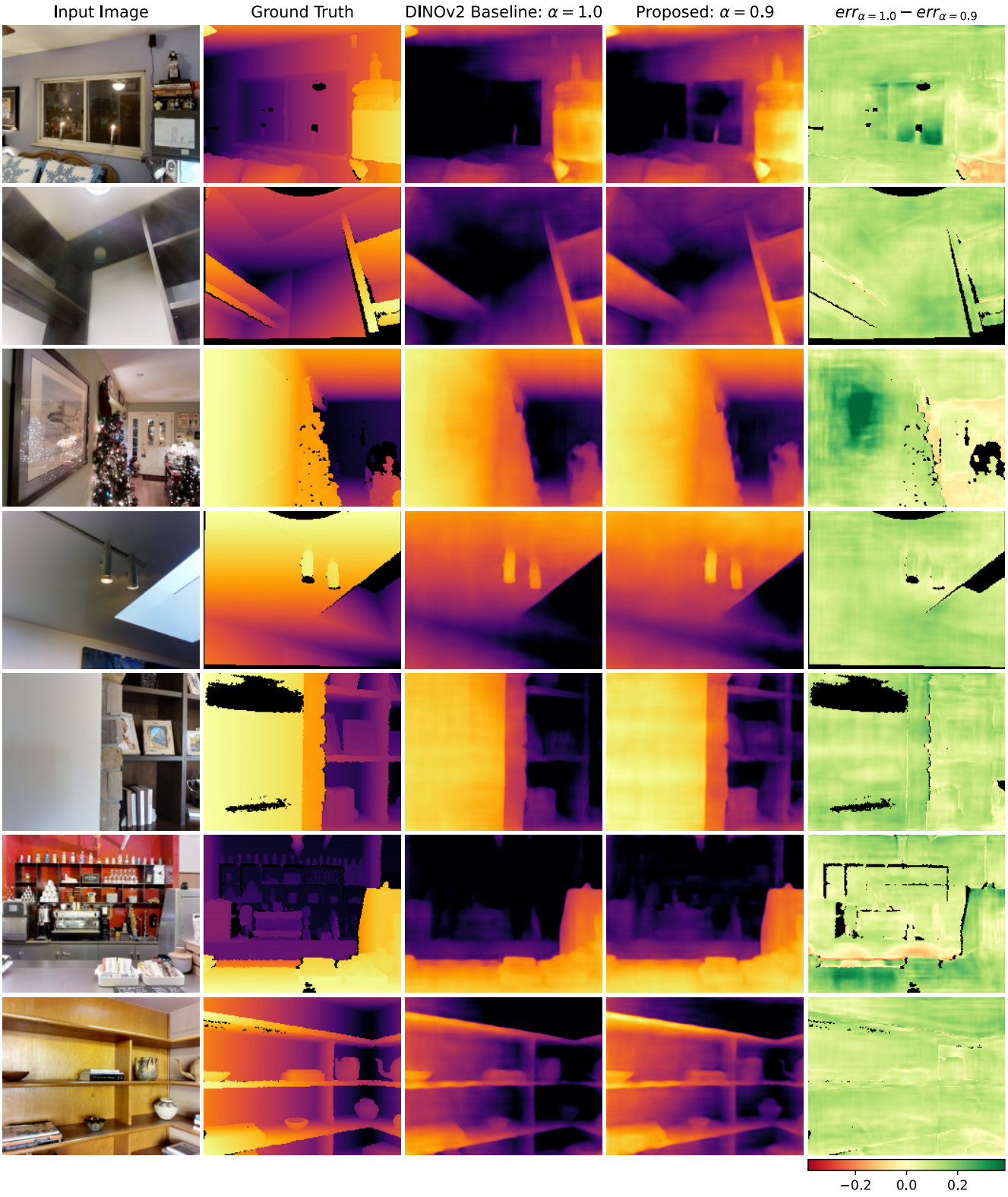

Figure 8. Results on Matterport with MIX6 auxiliary MLDC task. From left to right: input image and respective ground truth, baseline and our method predictions, and error difference between the last two w.r.t. to the ground truth.

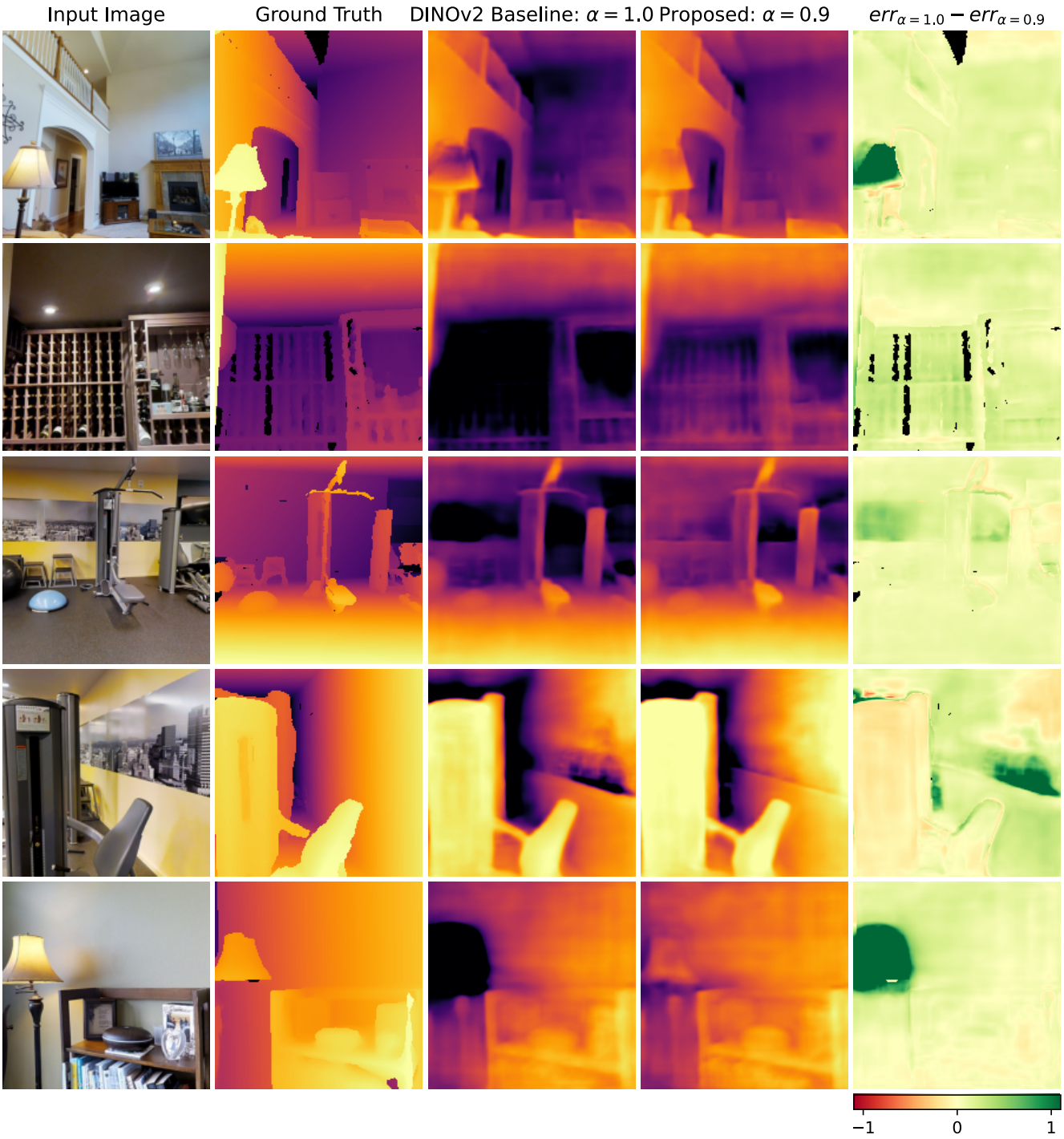

Figure 9. Results on Taskonomy with MIX6 auxiliary MLDC task. From left to right: input image and respective ground truth, baseline and our method predictions, and error difference between the last two w.r.t. to the ground truth.

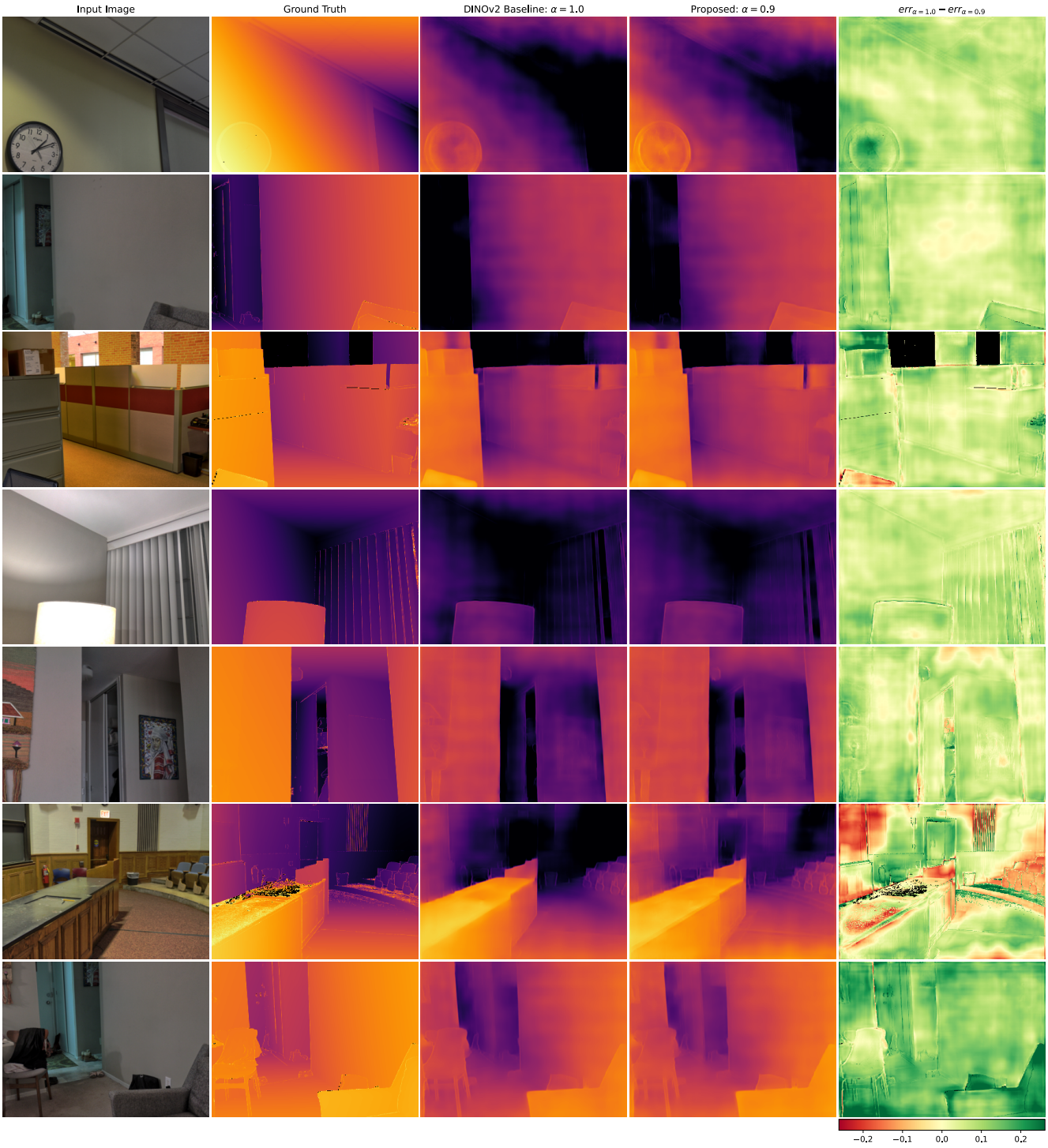

Figure 10. Results on DIODE Indoor with MIX6 auxiliary MLDC task. From left to right: input image and respective ground truth, baseline and our method predictions, and error difference between the last two w.r.t. to the ground truth.

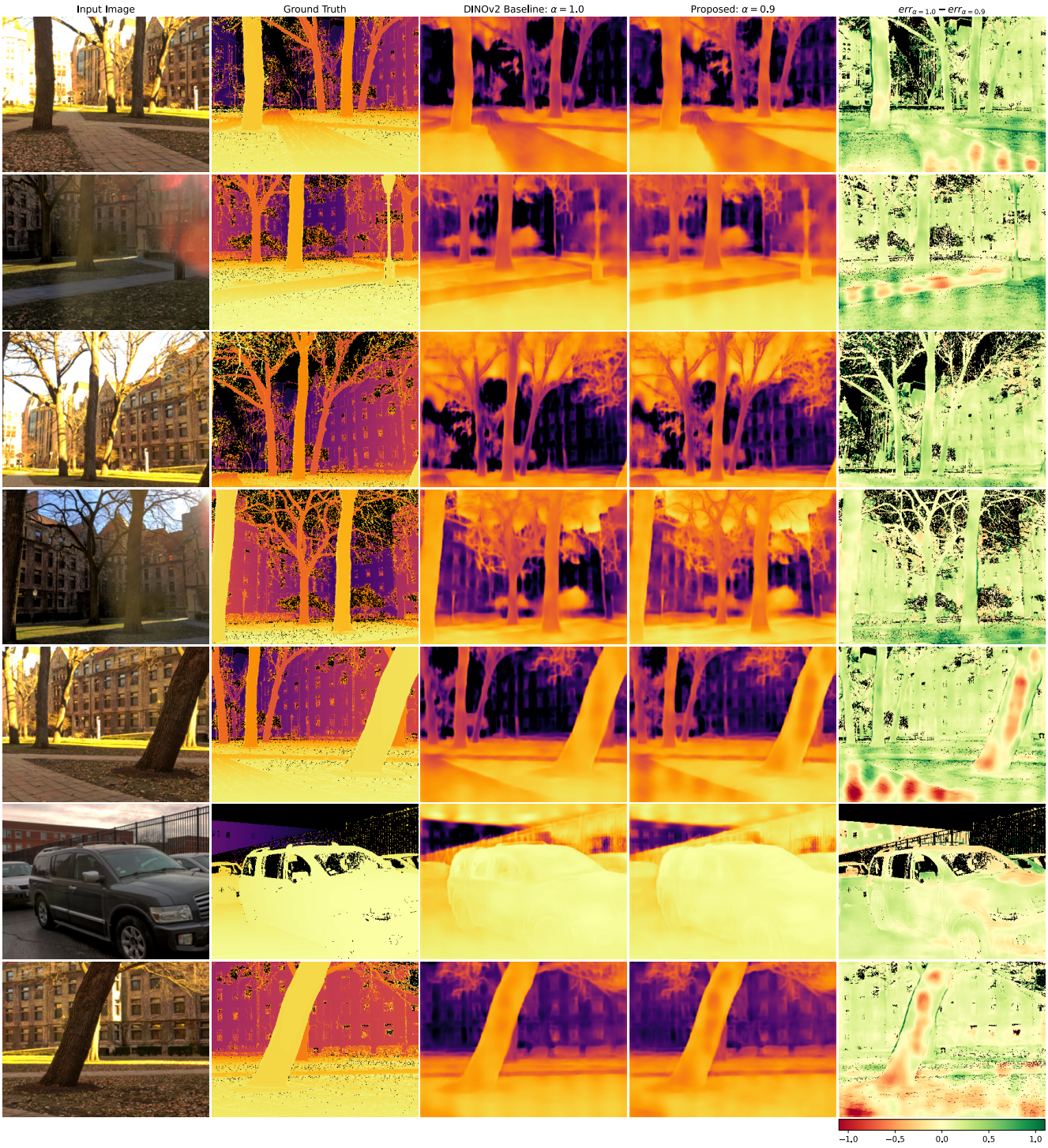

Figure 11. Results on DIODE Outdoor with MIX6 auxiliary MLDC task. Left to right: input image and ground truth; baseline and our method predictions; error difference relative to ground truth highlighting visible improvements in the facade behind the trees.