Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data
Supplementary Material

Lihe Yang[1] Bingyi Kang[2†] Zilong Huang[2] Xiaogang Xu[3,4] Jiashi Feng[2] Hengshuang Zhao[1‡]
[1]HKU [2]TikTok [3]CUHK [4]ZJU
† project lead ‡ corresponding author
https://depth-anything.github.io

1. More Implementation Details

We resize the shorter side of all images to 518 and keep the original aspect ratio. All images are cropped to 518×518 during training. During inference, we do not crop images and only ensure both sides are multipliers of 14, since the pre-defined patch size of DINOv2 encoders [12] is 14. Evaluation is performed at the original resolution by interpolating the prediction. Following MiDaS [3, 13], in zero-shot evaluation, the scale and shift of our prediction are manually aligned with the ground truth.

When fine-tuning our pre-trained encoder to metric depth estimation, we adopt the ZoeDepth codebase [2]. We merely replace the original MiDaS-based encoder with our stronger Depth Anything encoder, with a few hyper-parameters modified. Concretely, the training resolution is 392×518 on NYUv2 [15] and 384×768 on KITTI [9] to match the patch size of our encoder. The encoder learning rate is set as 1/50 of the learning rate of the randomly initialized decoder, which is much smaller than the 1/10 adopted for MiDaS encoder, due to our strong initialization. The batch size is 16 and the model is trained for 5 epochs.

When fine-tuning our pre-trained encoder to semantic segmentation, we use the MMSegmentation codebase [6]. The training resolution is set as 896×896 on both ADE20K [17] and Cityscapes [7]. The encoder learning rate is set as 3e-6 and the decoder learning rate is 10× larger. We use Mask2Former [5] as our semantic segmentation model. The model is trained for 160K iterations on ADE20K and 80K iterations on Cityscapes both with batch size 16, without any COCO [11] or Mapillary [1] pre-training. Other training configurations are the same as the original codebase.

2. More Ablation Studies

All ablation studies here are conducted on the ViT-S model.

The necessity of tolerance margin for feature alignment. As shown in Table 1, the gap between the tolerance margin of 1.00 and 0.85 or 0.70 clearly demonstrates the necessity of this design (mean AbsRel: 0.188 *vs.* 0.175).

α	KITTI	NYU	Sintel	DDAD	ETH3D	DIODE	**Mean**
1.00	0.085	0.055	0.523	0.250	0.134	0.079	0.188
0.85	0.080	**0.053**	**0.464**	**0.247**	**0.127**	**0.076**	**0.175**
0.70	**0.079**	0.054	0.482	0.248	0.127	0.077	0.178

Table 1. Ablation studies on different values of the tolerance margin α for the feature alignment loss $\mathcal{L}_{feat}$. Limited by space, we only report the AbsRel (↓) metric here.

$\mathcal{L}_{feat}$		Unseen datasets (AbsRel ↓)						**Mean**
U	L	KITTI	NYU	Sintel	DDAD	ETH3D	DIODE	
		0.083	0.055	0.478	0.249	0.133	0.080	0.180
✓		**0.080**	**0.053**	**0.464**	**0.247**	**0.127**	**0.076**	**0.175**
	✓	0.084	0.054	0.472	0.252	0.133	0.081	0.179

Table 2. Ablation studies of applying our feature alignment loss $\mathcal{L}_{feat}$ to unlabeled data (**U**) or labeled data (**L**).

Applying feature alignment to labeled data. Previously, we enforce the feature alignment loss $\mathcal{L}_{feat}$ on unlabeled data. Indeed, it is technically feasible to also apply this constraint to labeled data. In Table 2, apart from applying $\mathcal{L}_{feat}$ on unlabeled data, we explore to apply it to labeled data. We find that adding this auxiliary optimization target to labeled data is not beneficial to our baseline that does not involve any feature alignment (their mean AbsRel values are almost the same: 0.180 *vs.* 0.179). We conjecture that this is because the labeled data has relatively higher-quality depth annotations. The involvement of semantic loss may interfere with the learning of these informative manual labels. In comparison, our pseudo labels are noisier and less informative. Therefore, introducing the auxiliary constraint to unlabeled data can combat the noise in pseudo depth labels, as well as arm our model with semantic capability.

3. Limitations and Future Works

Currently, the largest model size is only constrained to ViT-Large [8]. Therefore, in the future, we plan to further scale up the model size from ViT-Large to ViT-Giant, which is also well pre-trained by DINOv2 [12]. We can train a more powerful teacher model with the larger model, producing more accurate pseudo labels for smaller models to learn, *e.g.*, ViT-L and ViT-B. Furthermore, to facilitate real-world applications, we believe the widely adopted 512$\times$512 training resolution is not enough. We plan to re-train our model on a larger resolution of 700+ or even 1000+.

4. More Qualitative Results

Please refer to the following pages for comprehensive qualitative results on six unseen test sets (Figure 1 for KITTI [9], Figure 2 for NYUv2 [15], Figure 3 for Sintel [4], Figure 4 for DDAD [10], Figure 5 for ETH3D [14], and Figure 6 for DIODE [16]). We compare our model with the strongest MiDaS model [3], *i.e.*, DPT-BEiT$_{\text{L-512}}$. Our model exhibits higher depth estimation accuracy and stronger robustness. Please refer to our project page for more visualizations.

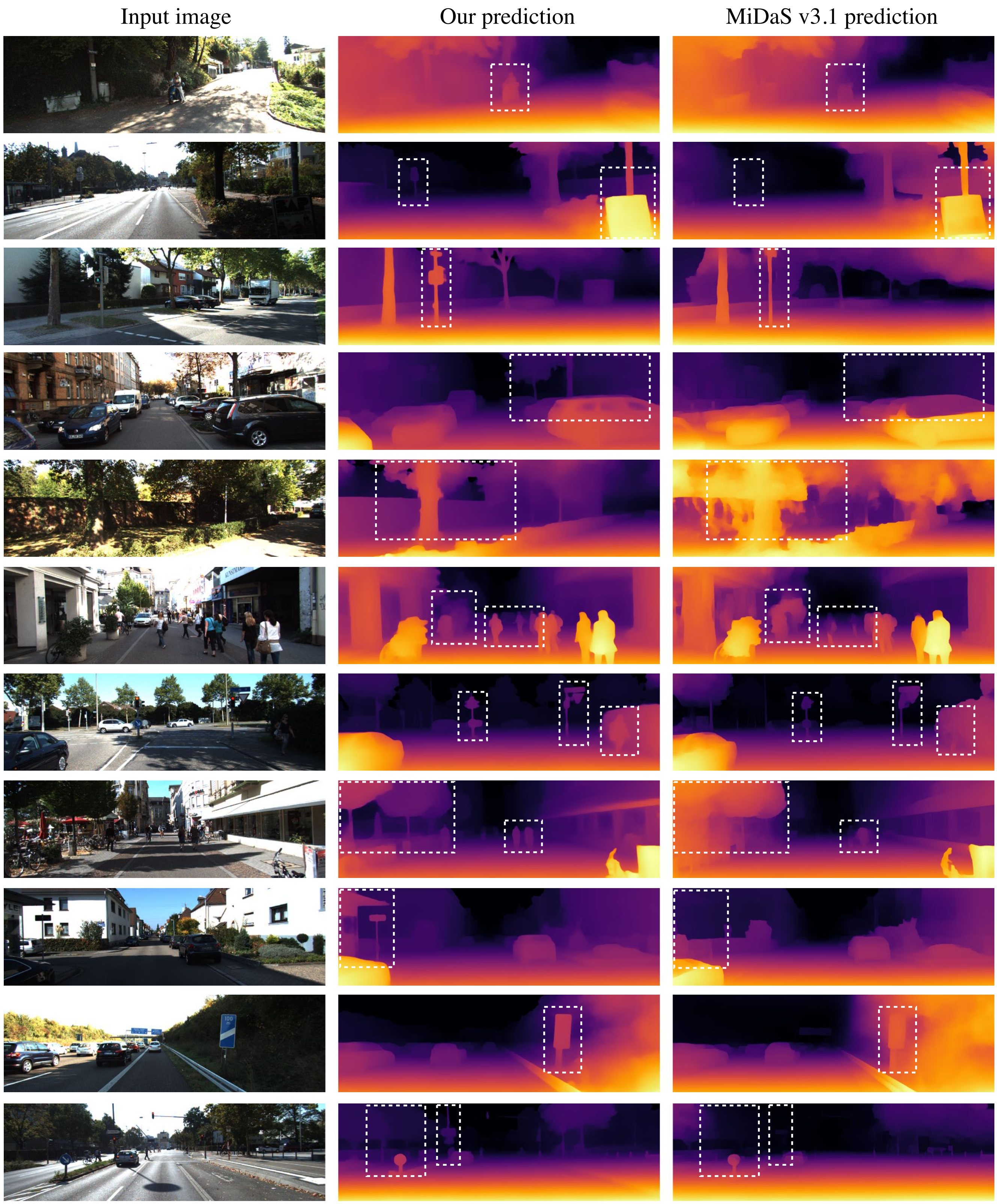

Figure 1. Qualitative results on KITTI. Due to the extremely sparse ground truth which is hard to visualize, we here compare our prediction with the most advanced MiDaS v3.1 [3] prediction. The brighter color denotes the closer distance.

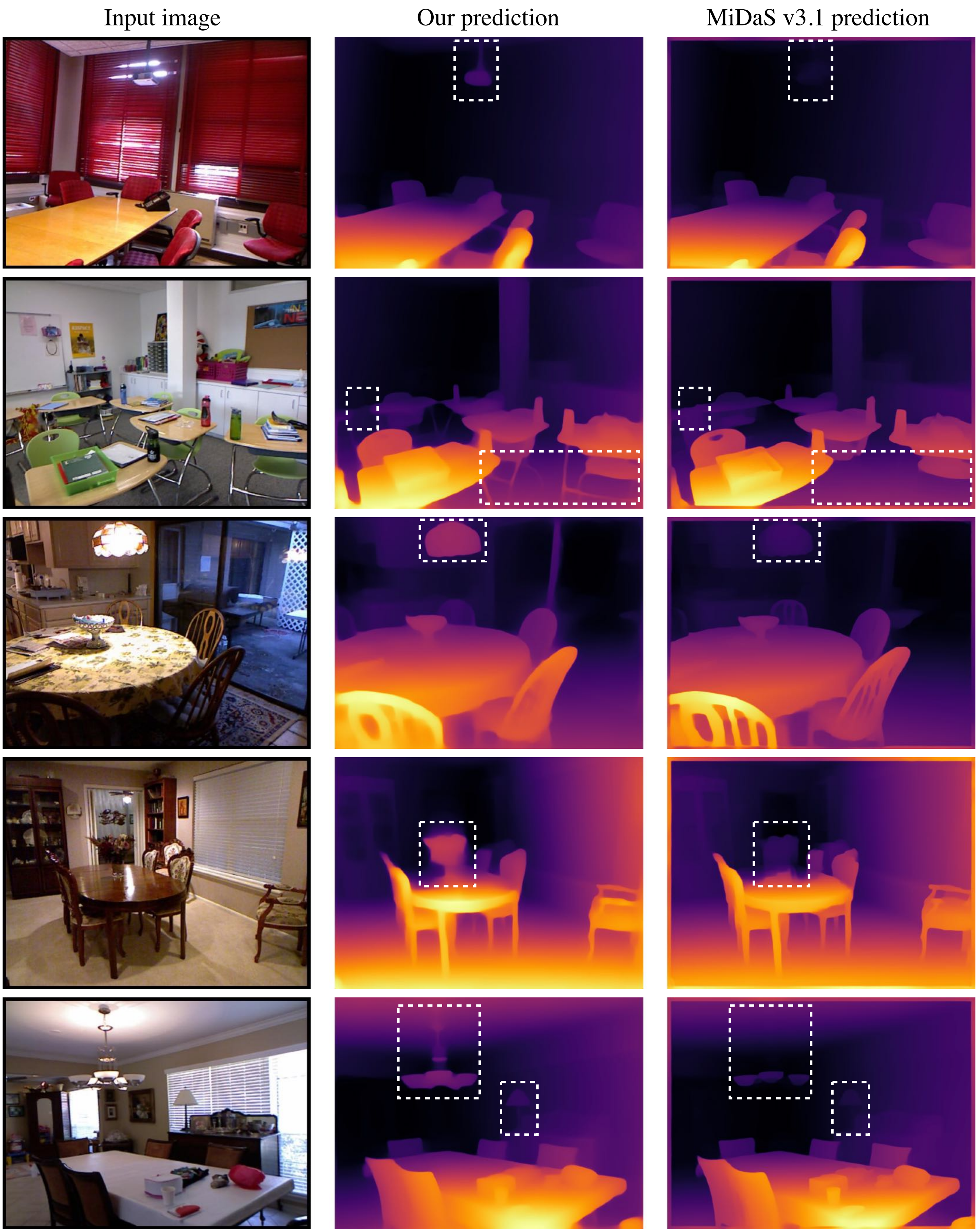

Figure 2. Qualitative results on NYUv2. It is worth noting that MiDaS [3] uses NYUv2 training data (*not zero-shot*), while we do not.

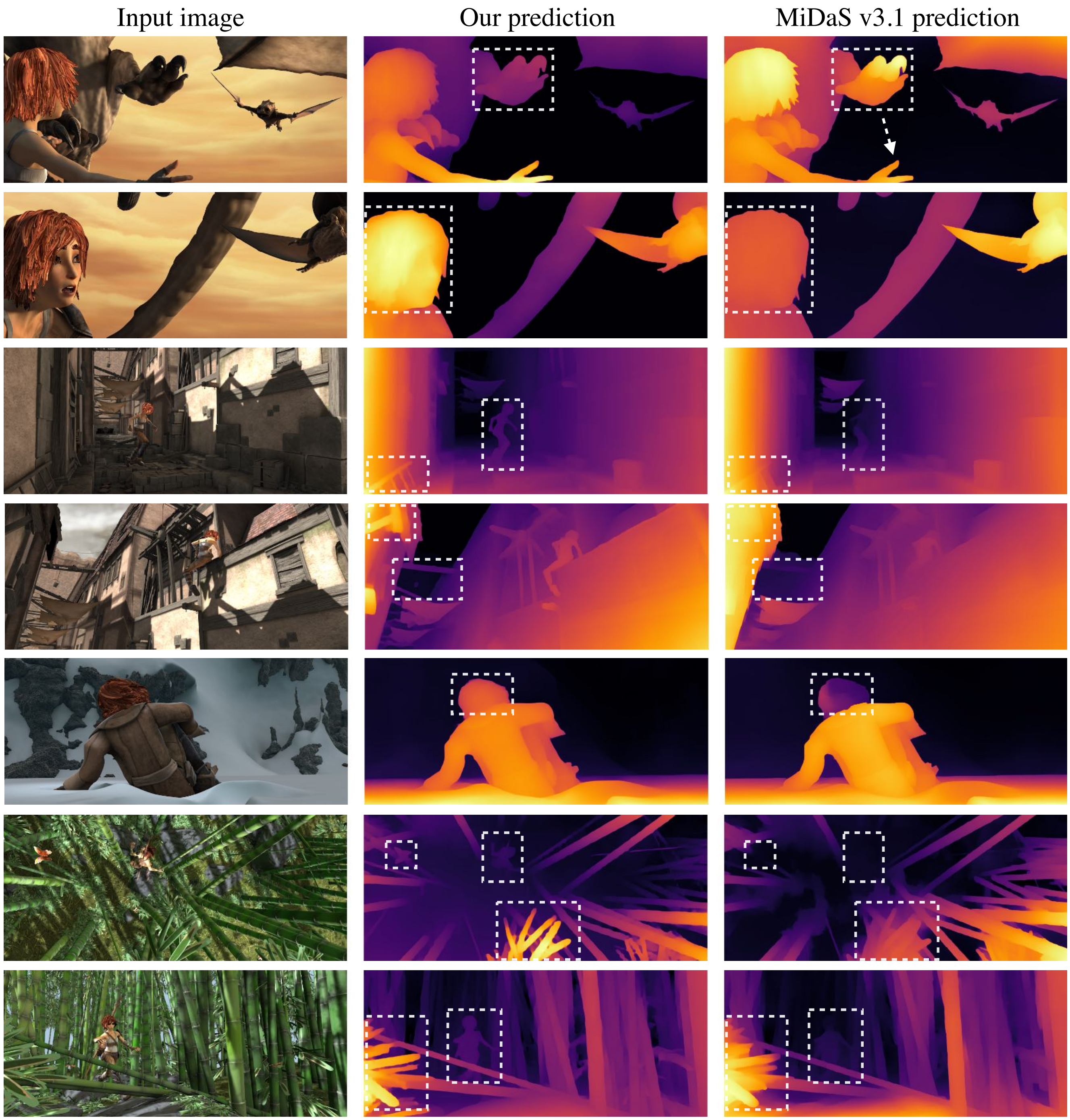

Figure 3. Qualitative results on Sintel.

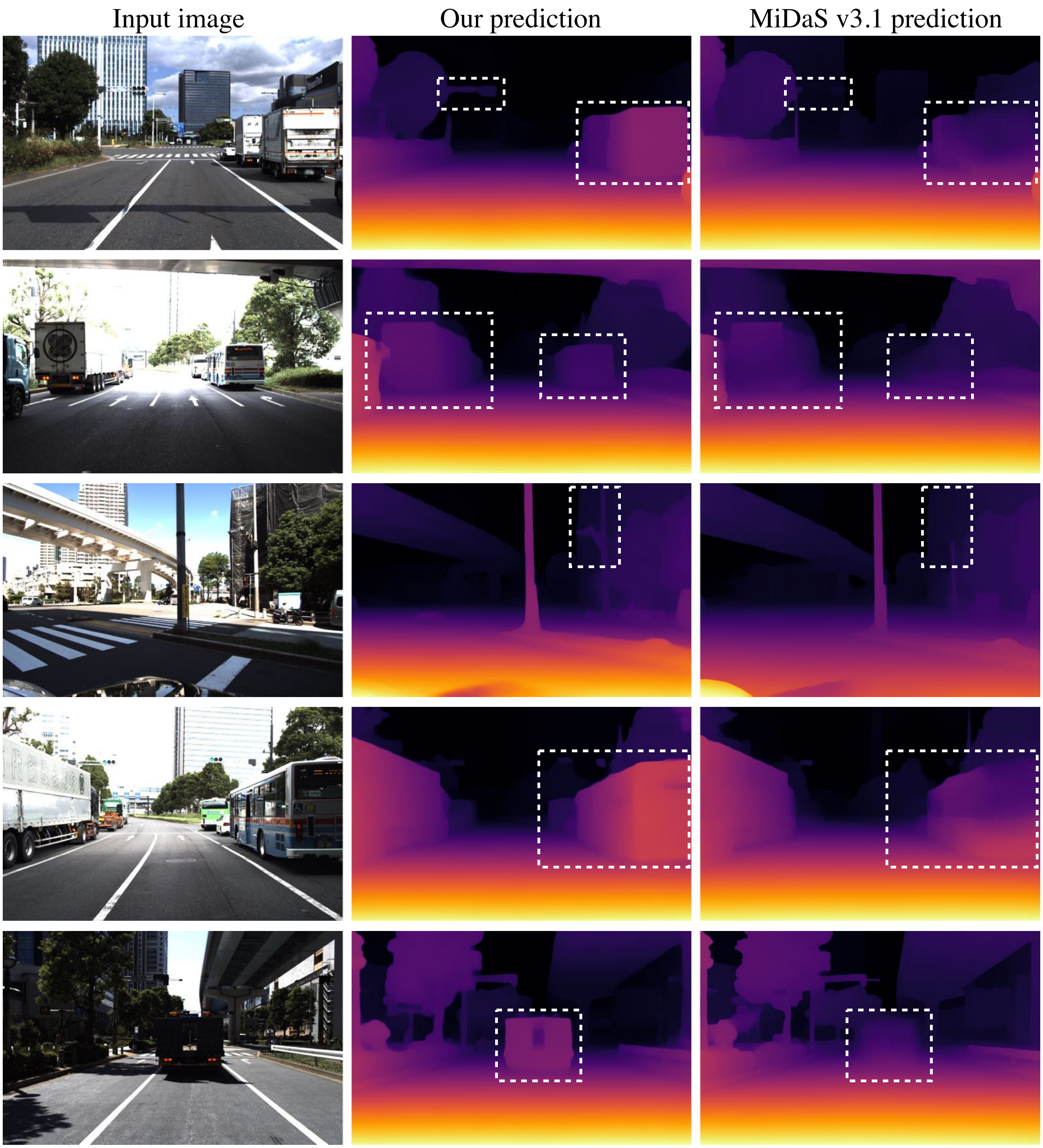

Figure 4. Qualitative results on DDAD.

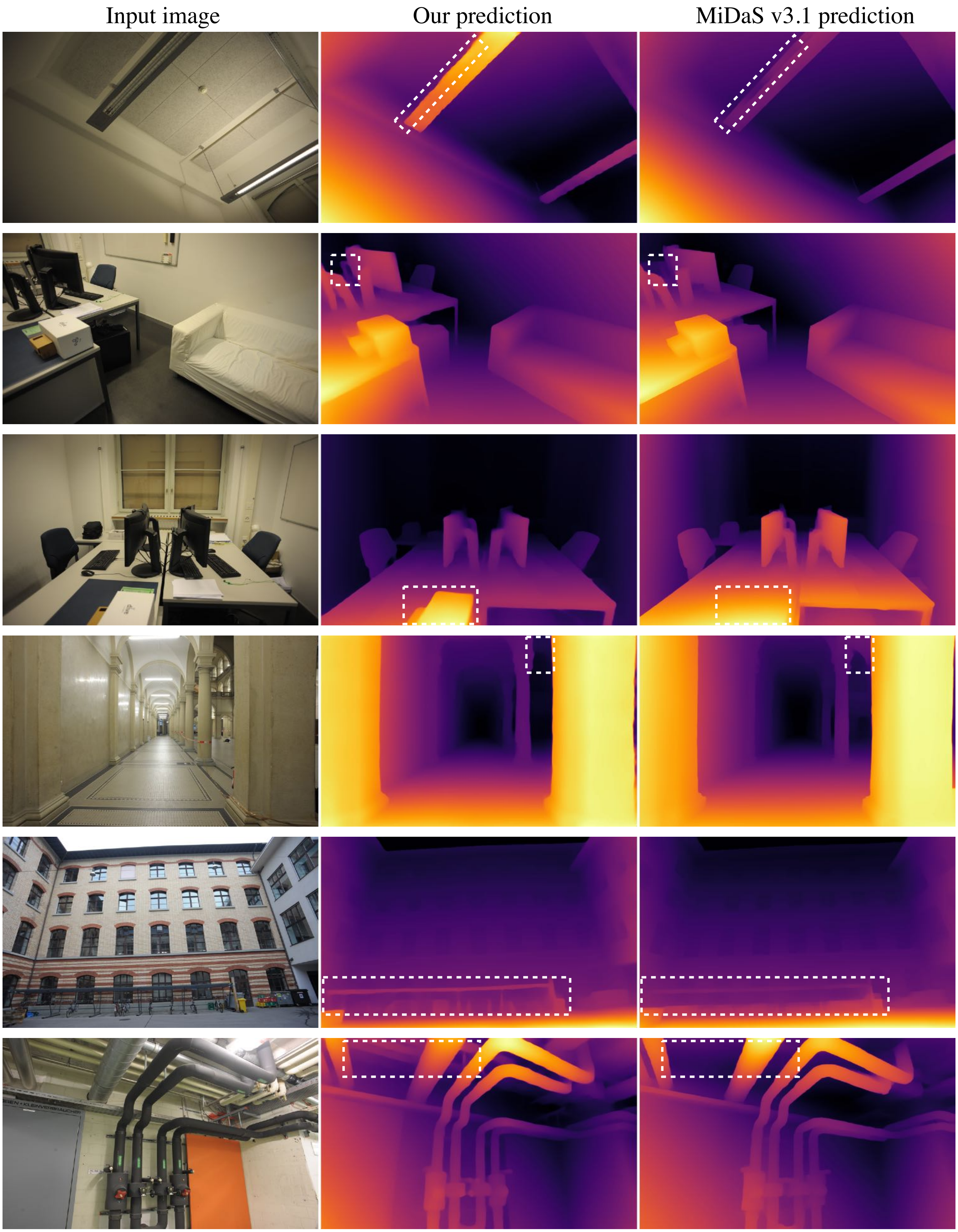

Figure 5. Qualitative results on ETH3D.

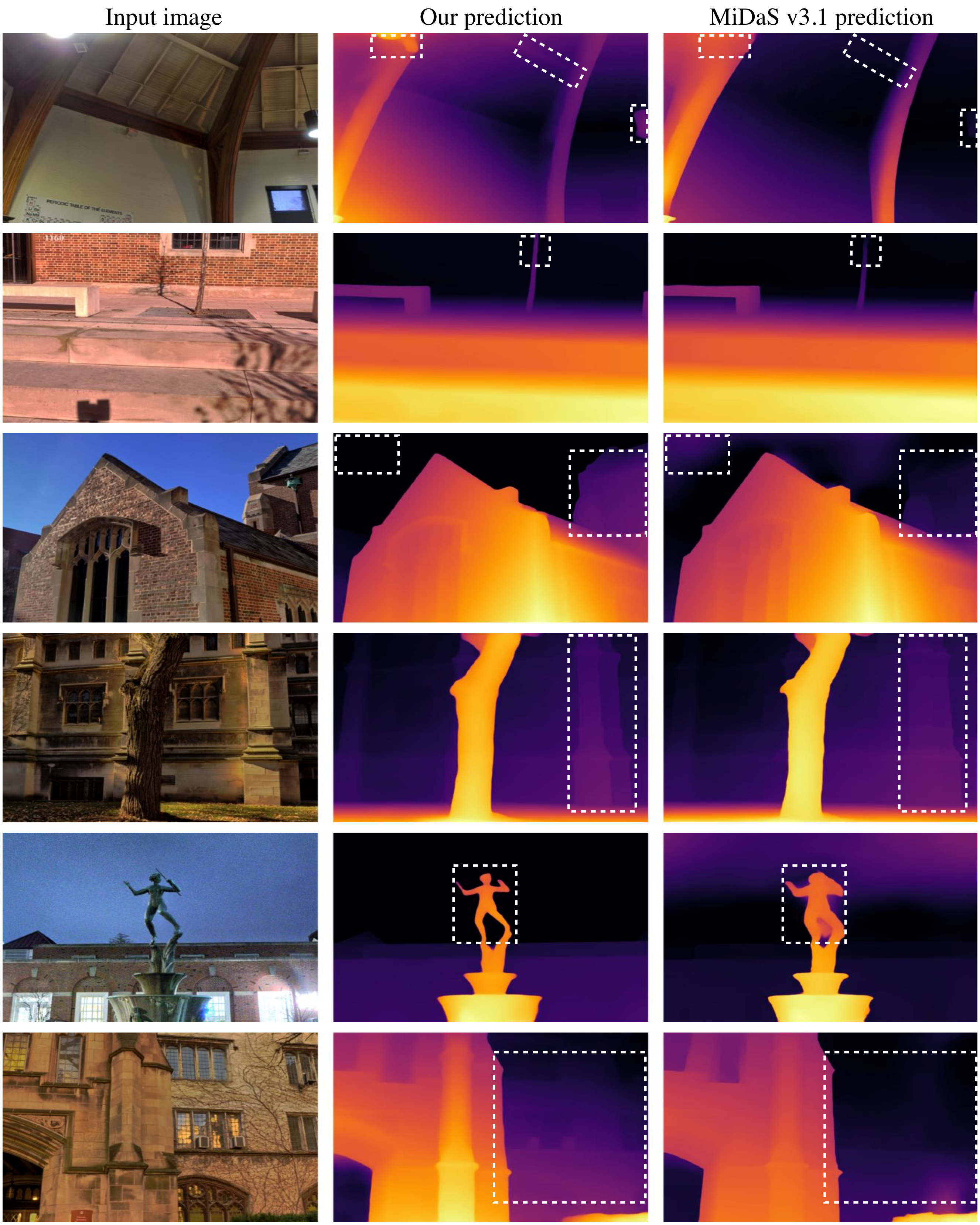

Figure 6. Qualitative results on DIODE.

References

[1] Manuel López Antequera, Pau Gargallo, Markus Hofinger, Samuel Rota Bulò, Yubin Kuang, and Peter Kontschieder. Mapillary planet-scale depth dataset. In *ECCV*, 2020. 1

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv:2302.12288*, 2023. 1

[3] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv:2307.14460*, 2023. 1, 2, 3, 4

[4] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 2

[5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 1

[6] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 1

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013. 1, 2

[10] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, 2020. 2

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[12] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2023. 1, 2

[13] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. 1

[14] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 2017. 2

[15] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1, 2

[16] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv:1908.00463*, 2019. 2

[17] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 1