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

Lihe Yang[1] Bingyi Kang[2†] Zilong Huang[2] Xiaogang Xu[3,4] Jiashi Feng[2] Hengshuang Zhao[1†]

[1]The University of Hong Kong [2]TikTok [3]Zhejiang Lab [4]Zhejiang University

† corresponding authors

https://depth-anything.github.io

Figure 1. Our model exhibits impressive generalization ability across extensive unseen scenes. **Left two columns:** COCO [36]. **Middle two:** SA-1B [27] (a hold-out unseen set). **Right two:** photos captured by ourselves. Our model works robustly in low-light environments (1st and 3rd column), complex scenes (2nd and 5th column), foggy weather (5th column), and ultra-remote distance (5th and 6th column), *etc*.

Abstract

This work presents Depth Anything[1]*, a highly practical solution for robust monocular depth estimation. Without pursuing novel technical modules, we aim to build a simple yet powerful foundation model dealing with any images under any circumstances. To this end, we scale up the dataset by designing a data engine to collect and automatically annotate large-scale unlabeled data (∼62M), which significantly enlarges the data coverage and thus is able to reduce the generalization error. We investigate two simple yet effective strategies that make data scaling-up promising. First, a more challenging optimization target is created by leveraging data augmentation tools. It compels the model to actively seek extra visual knowledge and acquire robust representations. Second, an auxiliary supervision is developed to enforce the model to inherit rich semantic priors from pre-trained encoders. We evaluate its zero-shot capabilities extensively, including six public datasets and randomly captured photos. It demonstrates impressive generalization ability (Figure 1). Further, through fine-tuning it with metric depth information from NYUv2 and KITTI, new SOTAs are set. Our better depth model also results in a better depth-conditioned ControlNet. Our models are released here.*

The work was done during an internship at TikTok.

[1]While the grammatical soundness of this name may be questionable, we treat it as a whole and pay homage to Segment Anything [27].

1. Introduction

The field of computer vision and natural language processing is currently experiencing a revolution with the emergence of "foundation models" [6] that demonstrate strong zero-/few-shot performance in various downstream scenarios [45, 59]. These successes primarily rely on large-scale training data that can effectively cover the data distribution. Monocular Depth Estimation (MDE), which is a fundamental problem with broad applications in robotics [66], autonomous driving [64, 80], virtual reality [48], *etc*., also requires a foundation model to estimate depth information from a single image. However, this has been underexplored due to the difficulty of building datasets with tens of millions of depth labels. MiDaS [46] made a pioneering study along this direction by training an MDE model on a collection of mixed labeled datasets. Despite demonstrating a certain level of zero-shot ability, MiDaS is limited by its data coverage, thus suffering disastrous performance in some scenarios.

In this work, our goal is to *build a foundation model for MDE* capable of producing high-quality depth information for any images under any circumstances. We approach this target from the perspective of dataset scaling-up. Traditionally, depth datasets are created mainly by acquiring depth data from sensors [18, 55], stereo matching [15], or SfM [33], which is costly, time-consuming, or even intractable in particular situations. We instead, for the first time, pay attention to large-scale unlabeled data. Compared with stereo images or

labeled images from depth sensors, our used monocular unlabeled images exhibit three advantages: (i) (*simple and cheap to acquire*) Monocular images exist almost everywhere, thus they are easy to collect, without requiring specialized devices. (ii) (*diverse*) Monocular images can cover a broader range of scenes, which are critical to the model generalization ability and scalability. (iii) (*easy to annotate*) We can simply use a pre-trained MDE model to assign depth labels for unlabeled images, which only takes a feedforward step. More than efficient, this also produces denser depth maps than LiDAR [18] and omits the computationally intensive stereo matching process.

We design a data engine to automatically generate depth annotations for unlabeled images, enabling data scaling-up to arbitrary scale. It collects 62M diverse and informative images from eight public large-scale datasets, *e.g.*, SA-1B [27], Open Images [30], and BDD100K [82]. We use their raw unlabeled images without any forms of labels. Then, in order to provide a reliable annotation tool for our unlabeled images, we collect 1.5M labeled images from six public datasets to train an initial MDE model. The unlabeled images are then automatically annotated and jointly learned with labeled images in a self-training manner [31].

Despite all the aforementioned advantages of monocular unlabeled images, it is indeed not trivial to make positive use of such large-scale unlabeled images [73, 90], especially in the case of sufficient labeled images and strong pre-training models. In our preliminary attempts, directly combining labeled and pseudo labeled images failed to improve the baseline of solely using labeled images. We conjecture that, the additional knowledge acquired in such a naive self-teaching manner is rather limited. To address the dilemma, we propose to challenge the student model with a more difficult optimization target when learning the pseudo labels. The student model is enforced to seek extra visual knowledge and learn robust representations under various strong perturbations to better handle unseen images.

Furthermore, there have been some works [9, 21] demonstrating the benefit of an auxiliary semantic segmentation task for MDE. We also follow this research line, aiming to equip our model with better high-level scene understanding capability. However, we observed when an MDE model is already powerful enough, it is hard for such an auxiliary task to bring further gains. We speculate that it is due to severe loss in semantic information when decoding an image into a discrete class space. Therefore, considering the excellent performance of DINOv2 in semantic-related tasks, we propose to maintain the rich semantic priors from it with a simple feature alignment loss. This not only enhances the MDE performance, but also yields a multi-task encoder for both middle-level and high-level perception tasks.

Our contributions are summarized as follows:

- We highlight the value of data scaling-up of massive, cheap, and diverse unlabeled images for MDE.
- We point out a key practice in jointly training large-scale labeled and unlabeled images. Instead of learning raw unlabeled images directly, we challenge the model with a harder optimization target for extra knowledge.
- We propose to inherit rich semantic priors from pre-trained encoders for better scene understanding, rather than using an auxiliary semantic segmentation task.
- Our model exhibits stronger zero-shot capability than MiDaS-BEiT$_{\text{L-512}}$ [5]. Further, fine-tuned with metric depth, it outperforms ZoeDepth [4] significantly.

2. Related Work

Monocular depth estimation (MDE). Early works [23, 37, 51] primarily relied on handcrafted features and traditional computer vision techniques. They were limited by their reliance on explicit depth cues and struggled to handle complex scenes with occlusions and textureless regions.

Deep learning-based methods have revolutionized monocular depth estimation by effectively learning depth representations from delicately annotated datasets [18, 55]. Eigen *et al*. [17] first proposed a multi-scale fusion network to regress the depth. Following this, many works consistently improve the depth estimation accuracy by carefully designing the regression task as a classification task [3, 34], introducing more priors [32, 54, 76, 83], and better objective functions [68, 78], *etc*. Despite the promising performance, they are hard to generalize to unseen domains.

Zero-shot depth estimation. Our work belongs to this research line. We aim to train an MDE model with a diverse training set and thus can predict the depth for any given image. Some pioneering works [10, 67] explored this direction by collecting more training images, but their supervision is very sparse and is only enforced on limited pairs of points.

To enable effective multi-dataset joint training, a milestone work MiDaS [46] utilizes an affine-invariant loss to ignore the potentially different depth scales and shifts across varying datasets. Thus, MiDaS provides relative depth information. Recently, some works [4, 22, 79] take a step further to estimate the metric depth. However, in our practice, we observe such methods exhibit poorer generalization ability than MiDaS, especially its latest version [5]. Besides, as demonstrated by ZoeDepth [4], a strong relative depth estimation model can also work well in generalizable metric depth estimation by fine-tuning with metric depth information. Therefore, we still follow MiDaS in relative depth estimation, but further strengthen it by highlighting the value of large-scale monocular unlabeled images.

Leveraging unlabeled data. This belongs to the research area of semi-supervised learning [31, 56, 90], which is popular with various applications [71, 75]. However, existing

works typically assume only limited images are available. They rarely consider the challenging but realistic scenario where there are already sufficient labeled images but also larger-scale unlabeled images. We take this challenging direction for zero-shot MDE. We demonstrate that unlabeled images can significantly enhance the data coverage and thus improve model generalization and robustness.

3. Depth Anything

Our work utilizes both labeled and unlabeled images to facilitate better monocular depth estimation (MDE). Formally, the labeled and unlabeled sets are denoted as $\mathcal{D}^l = \{(x_i, d_i)\}_{i=1}^M$ and $\mathcal{D}^u = \{u_i\}_{i=1}^N$ respectively. We aim to learn a teacher model T from $\mathcal{D}^l$. Then, we utilize T to assign pseudo depth labels for $\mathcal{D}^u$. Finally, we train a student model S on the combination of labeled set and pseudo labeled set. A brief illustration is provided in Figure 2.

3.1. Learning Labeled Images

This process is similar to the training of MiDaS [5, 46]. However, since MiDaS did not release its code, we first reproduced it. Concretely, the depth value is first transformed into the disparity space by $d = 1/t$ and then normalized to 0∼1 on each depth map. To enable multi-dataset joint training, we adopt the affine-invariant loss to ignore the unknown scale and shift of each sample:

$$\mathcal{L}_l = \frac{1}{HW} \sum_{i=1}^{HW} \rho(d_i^*, d_i), \tag{1}$$

where d_i^* and d_i are the prediction and ground truth, respectively. And ρ is the affine-invariant mean absolute error loss: $\rho(d_i^*, d_i) = |\hat{d}_i^* - \hat{d}_i|$, where $\hat{d}_i^*$ and $\hat{d}_i$ are the scaled and shifted versions of the prediction d_i^* and ground truth d_i:

$$\hat{d}_i = \frac{d_i - t(d)}{s(d)}, \tag{2}$$

where $t(d)$ and $s(d)$ are used to align the prediction and ground truth to have zero translation and unit scale:

$$t(d) = \text{median}(d), \quad s(d) = \frac{1}{HW} \sum_{i=1}^{HW} |d_i - t(d)|. \tag{3}$$

To obtain a robust monocular depth estimation model, we collect 1.5M labeled images from 6 public datasets. Details of these datasets are listed in Table 1. We use fewer labeled datasets than MiDaS v3.1 [5] (12 training datasets), because 1) we do not use NYUv2 [55] and KITTI [18] datasets to ensure zero-shot evaluation on them, 2) some datasets are not available (anymore), *e.g.*, Movies [46] and WSVD [61], and 3) some datasets exhibit poor quality, *e.g.*, RedWeb (also low resolution) [67]. Despite using fewer labeled images, our easy-to-acquire and diverse unlabeled images will comprehend the data coverage and greatly enhance the model generalization ability and robustness.

Dataset	Indoor	Outdoor	Label	# Images
Labeled Datasets				
BlendedMVS [77]	✓	✓	Stereo	115K
DIML [13]	✓	✓	Stereo	927K
HRWSI [68]	✓	✓	Stereo	20K
IRS [62]	✓		Stereo	103K
MegaDepth [33]		✓	SfM	128K
TartanAir [63]	✓	✓	Stereo	306K
Unlabeled Datasets				
BDD100K [82]		✓	None	8.2M
Google Landmarks [65]		✓	None	4.1M
ImageNet-21K [50]	✓	✓	None	13.1M
LSUN [81]	✓		None	9.8M
Objects365 [53]	✓	✓	None	1.7M
Open Images V7 [30]	✓	✓	None	7.8M
Places365 [88]	✓	✓	None	6.5M
SA-1B [27]	✓	✓	None	11.1M

Table 1. In total, our Depth Anything is trained on 1.5M labeled images and **62M unlabeled images** jointly.

Furthermore, to strengthen the teacher model T learned from these labeled images, we adopt the DINOv2 [43] pre-trained weights to initialize our encoder. In practice, we apply a pre-trained semantic segmentation model [70] to detect the `sky` region, and set its disparity value as 0 (farthest).

3.2. Unleashing the Power of Unlabeled Images

This is the main point of our work. Distinguished from prior works that laboriously construct diverse labeled datasets, we highlight the value of unlabeled images in enhancing the data coverage. Nowadays, we can practically build a diverse and large-scale unlabeled set from the Internet or public datasets of various tasks. Also, we can effortlessly obtain the dense depth map of monocular unlabeled images simply by forwarding them to a pre-trained well-performed MDE model. This is much more convenient and efficient than performing stereo matching or SfM reconstruction for stereo images or videos. We select eight large-scale public datasets as our unlabeled sources for their diverse scenes. They contain more than 62M images in total. The details are provided in the bottom half of Table 1.

Technically, given the previously obtained MDE teacher model T, we make predictions on the unlabeled set $\mathcal{D}^u$ to obtain a pseudo labeled set $\hat{\mathcal{D}}^u$:

$$\hat{\mathcal{D}}^u = \{(u_i, T(u_i)) | u_i \in \mathcal{D}^u\}_{i=1}^N. \tag{4}$$

With the combination set $\mathcal{D}^l \cup \hat{\mathcal{D}}^u$ of labeled images and pseudo labeled images, we train a student model S on it.

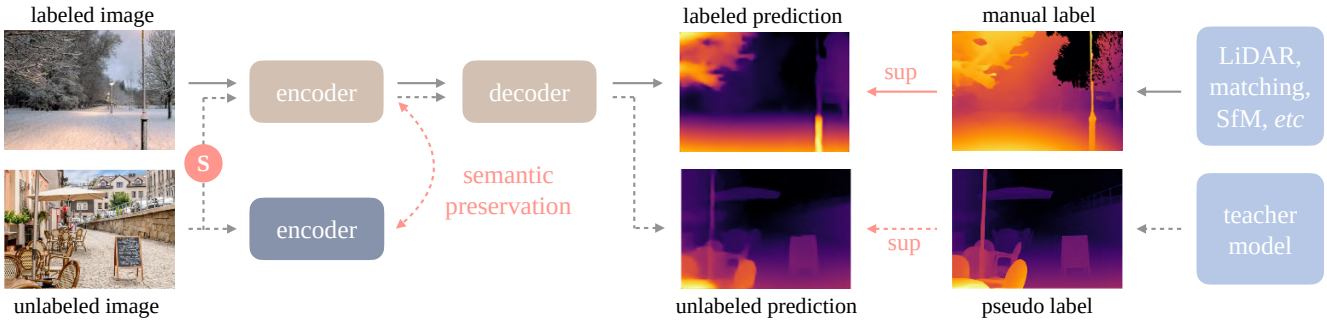

Figure 2. Our pipeline. Solid line: flow of labeled images, dotted line: unlabeled images. We especially highlight the value of large-scale unlabeled images. The S denotes adding strong perturbations (Section 3.2). To equip our depth estimation model with rich semantic priors, we enforce an auxiliary constraint between the online student model and a frozen encoder to preserve the semantic capability (Section 3.3).

Following prior works [74], instead of fine-tuning S from T, we re-initialize S for better performance.

Unfortunately, in our pilot studies, we failed to gain improvements with such a self-training pipeline, which indeed contradicts the observations when there are only a few labeled images [56]. We conjecture that, with already sufficient labeled images in our case, the extra knowledge acquired from additional unlabeled images is rather limited. Especially considering the teacher and student share the same pre-training and architecture, they tend to make similar correct or false predictions on the unlabeled set $\mathcal{D}^u$, even without the explicit self-training procedure.

To address the dilemma, we propose to challenge the student with a more difficult optimization target for additional visual knowledge on unlabeled images. We inject strong perturbations to unlabeled images during training. It compels our student model to actively seek extra visual knowledge and acquire invariant representations from these unlabeled images. These advantages help our model deal with the open world more robustly. We introduce two forms of perturbations: one is strong color distortions, including color jittering and Gaussian blurring, and the other is strong spatial distortion, which is CutMix [84]. Despite the simplicity, the two modifications make our large-scale unlabeled images significantly improve the baseline of labeled images.

We provide more details about CutMix. It was originally proposed for image classification, and is rarely explored in monocular depth estimation. We first interpolate a random pair of unlabeled images u_a and u_b spatially:

$$u_{ab} = u_a \odot M + u_b \odot (1 - M), \tag{5}$$

where M is a binary mask with a rectangle region set as 1.

The unlabeled loss $\mathcal{L}_u$ is obtained by first computing affine-invariant losses in valid regions defined by M and $1 - M$, respectively:

$$\mathcal{L}_u^M = \rho\big(S(u_{ab}) \odot M,\, T(u_a) \odot M\big), \tag{6}$$

$$\mathcal{L}_u^{1-M} = \rho\big(S(u_{ab}) \odot (1 - M), T(u_b) \odot (1 - M)\big), \tag{7}$$

where we omit the $\sum$ and pixel subscript i for simplicity. Then we aggregate the two losses via weighted averaging:

$$\mathcal{L}_u = \frac{\sum M}{HW}\mathcal{L}_u^M + \frac{\sum(1 - M)}{HW}\mathcal{L}_u^{1-M}. \tag{8}$$

We use CutMix with 50% probability. The unlabeled images for CutMix are already strongly distorted in color, but the unlabeled images fed into the teacher model T for pseudo labeling are clean, without any distortions.

3.3. Semantic-Assisted Perception

There exist some works [9, 21, 28, 72] improving depth estimation with an auxiliary semantic segmentation task. We believe that arming our depth estimation model with such high-level semantic-related information is beneficial. Besides, in our specific context of leveraging unlabeled images, these auxiliary supervision signals from other tasks can also combat the potential noise in our pseudo depth label.

Therefore, we made an initial attempt by carefully assigning semantic segmentation labels to our unlabeled images with a combination of RAM [86] + GroundingDINO [38] + HQ-SAM [26] models. After post-processing, this yields a class space containing 4K classes. In the joint-training stage, the model is enforced to produce both depth and segmentation predictions with a shared encoder and two individual decoders. Unfortunately, after trial and error, we still could not boost the performance of the original MDE model. We speculated that, decoding an image into a discrete class space indeed loses too much semantic information. The limited information in these semantic masks is hard to further boost our depth model, especially when our depth model has established very competitive results.

Therefore, we aim to seek more informative semantic signals to serve as auxiliary supervision for our depth estimation task. We are greatly astonished by the strong performance of DINOv2 models [43] in semantic-related tasks, *e.g.*, image retrieval and semantic segmentation, even with frozen weights without any fine-tuning. Motivated by these clues, we propose to transfer its strong semantic capability to our

Method	Encoder	KITTI [18]		NYUv2 [55]		Sintel [7]		DDAD [20]		ETH3D [52]		DIODE [60]	
		AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1
MiDaS v3.1 [5]	ViT-L	0.127	0.850	0.048	0.980	0.587	0.699	0.251	0.766	0.139	0.867	0.075	0.942
Depth Anything	ViT-S	0.080	0.936	0.053	0.972	0.464	0.739	0.247	0.768	0.127	**0.885**	0.076	0.939
	ViT-B	0.080	0.939	0.046	0.979	**0.432**	0.756	0.232	0.786	**0.126**	0.884	0.069	0.946
	ViT-L	**0.076**	**0.947**	**0.043**	**0.981**	0.458	**0.760**	**0.230**	**0.789**	0.127	0.882	**0.066**	**0.952**

Table 2. **Zero-shot relative** depth estimation. **Better:** AbsRel ↓ , δ_1 ↑. We compare with the best model from MiDaS v3.1. Note that MiDaS ***does not*** strictly follow the zero-shot evaluation on KITTI and NYUv2, because it uses their training images. We provide three model scales for different purposes, based on ViT-S (24.8M), ViT-B (97.5M), and ViT-L (335.3M), respectively. **Best**, second best results.

depth model with an auxiliary feature alignment loss. The feature space is high-dimensional and continuous, thus containing richer semantic information than discrete masks. The feature alignment loss is formulated as:

$$\mathcal{L}_{feat} = 1 - \frac{1}{HW} \sum_{i=1}^{HW} \cos(f_i, f_i'), \tag{9}$$

where $\cos(\cdot, \cdot)$ measures the cosine similarity between two feature vectors. f is the feature extracted by the depth model S, while f' is the feature from a frozen DINOv2 encoder. We do not follow some works [19] to project the online feature f into a new space for alignment, because a randomly initialized projector makes the large alignment loss dominate the overall loss in the early stage.

Another key point in feature alignment is that, semantic encoders like DINOv2 tend to produce similar features for different parts of an object, *e.g.*, car front and rear. In depth estimation, however, different parts or even pixels within the same part, can be of varying depth. Thus, it is not beneficial to *exhaustively* enforce our depth model to produce exactly the same features as the frozen encoder.

To solve this issue, we set a tolerance margin α for the feature alignment. If the cosine similarity of f_i and f_i' has surpassed α, this pixel will not be considered in our $\mathcal{L}_{feat}$. This allows our method to enjoy both the semantic-aware representation from DINOv2 and the part-level discriminative representation from depth supervision. As a side effect, our produced encoder not only performs well in downstream MDE datasets, but also achieves strong results in the semantic segmentation task. It also indicates the potential of our encoder to serve as a universal multi-task encoder for both middle-level and high-level perception tasks.

Finally, our overall loss is an average combination of the three losses $\mathcal{L}_l$, $\mathcal{L}_u$, and $\mathcal{L}_{feat}$.

4. Experiment

4.1. Implementation Details

We adopt the DINOv2 encoder [43] for feature extraction. Following MiDaS [5, 46], we use the DPT [47] decoder for depth regression. All labeled datasets are simply combined together without re-sampling. In the first stage, we train a teacher model on labeled images for 20 epochs. In the second stage of joint training, we train a student model to sweep across all unlabeled images for one time. The unlabeled images are annotated by a best-performed teacher model with a ViT-L encoder. The ratio of labeled and unlabeled images is set as 1:2 in each batch. In both stages, the base learning rate of the pre-trained encoder is set as 5e-6, while the randomly initialized decoder uses a 10× larger learning rate. We use the AdamW optimizer and decay the learning rate with a linear schedule. We only apply horizontal flipping as our data augmentation for labeled images. The tolerance margin α for feature alignment loss is set as 0.15. For more details, please refer to our appendix.

4.2. Zero-Shot Relative Depth Estimation

As aforementioned, this work aims to provide accurate depth estimation for any image. Therefore, we comprehensively validate the zero-shot depth estimation capability of our Depth Anything model on six representative unseen datasets: KITTI [18], NYUv2 [55], Sintel [7], DDAD [20], ETH3D [52], and DIODE [60]. We compare with the best DPT-BEiT$_{\text{L-512}}$ model from the latest MiDaS v3.1 [5], which uses more labeled images than us. As shown in Table 2, both with a ViT-L encoder, our Depth Anything surpasses the strongest MiDaS model tremendously across extensive scenes in terms of both the AbsRel (absolute relative error: $|d^* - d|/d$) and δ_1 (percentage of $\max(d^*/d, d/d^*) < 1.25$) metrics. For example, when tested on the well-known autonomous driving dataset DDAD [20], we improve the AbsRel (↓) from 0.251 → 0.230 and improve the δ_1 (↑) from 0.766 → 0.789.

Besides, our ViT-B model is already clearly superior to the MiDaS based on a much larger ViT-L. Moreover, our ViT-S model, whose scale is less than 1/10 of the MiDaS model, even outperforms MiDaS on several unseen datasets, including Sintel, DDAD, and ETH3D. The performance advantage of these small-scale models demonstrates their great potential in computationally-constrained scenarios.

It is also worth noting that, on the most widely used MDE

Method	*Higher is better* ↑			*Lower is better* ↓		
	δ_1	δ_2	δ_3	AbsRel	RMSE	log10
AdaBins [3]	0.903	0.984	0.997	0.103	0.364	0.044
DPT [47]	0.904	0.988	0.998	0.110	0.357	0.045
P3Depth [44]	0.898	0.981	0.996	0.104	0.356	0.043
SwinV2-L [40]	0.949	0.994	0.999	0.083	0.287	0.035
AiT [42]	0.954	0.994	0.999	0.076	0.275	0.033
VPD [87]	0.964	0.995	0.999	0.069	0.254	0.030
ZoeDepth* [4]	0.951	0.994	0.999	0.077	0.282	0.033
Ours	**0.984**	**0.998**	**1.000**	**0.056**	**0.206**	**0.024**

Table 3. **Fine-tuning and evaluating on NYUv2** [55] with our pre-trained MDE encoder. We highlight **best**, second best results, as well as **most discriminative metrics**. ∗: Reproduced by us.

Method	*Higher is better* ↑			*Lower is better* ↓		
	δ_1	δ_2	δ_3	AbsRel	RMSE	RMSE log
AdaBins [3]	0.964	0.995	0.999	0.058	2.360	0.088
DPT [47]	0.959	0.995	0.999	0.062	2.573	0.092
P3Depth [44]	0.953	0.993	0.998	0.071	2.842	0.103
NeWCRFs [83]	0.974	0.997	0.999	0.052	2.129	0.079
SwinV2-L [40]	0.977	0.998	1.000	0.050	1.966	0.075
NDDepth [54]	0.978	0.998	0.999	0.050	2.025	0.075
GEDepth [76]	0.976	0.997	0.999	0.048	2.044	0.076
ZoeDepth* [4]	0.971	0.996	0.999	0.054	2.281	0.082
Ours	**0.982**	**0.998**	**1.000**	**0.046**	**1.896**	**0.069**

Table 4. **Fine-tuning and evaluating on KITTI** [18] with our pre-trained MDE encoder. ∗: Reproduced by us.

benchmarks KITTI and NYUv2, although MiDaS v3.1 uses the corresponding training images (*not zero-shot anymore*), our Depth Anything is still evidently superior to it *without training with any KITTI or NYUv2 images*, *e.g.*, 0.127 *vs.* 0.076 in AbsRel and 0.850 *vs.* 0.947 in δ_1 on KITTI.

4.3. Fine-tuned to *Metric* Depth Estimation

Apart from the impressive performance in zero-shot relative depth estimation, we further examine our Depth Anything model as a promising weight initialization for downstream *metric* depth estimation. We initialize the encoder of downstream MDE models with our pre-trained encoder parameters and leave the decoder randomly initialized. The model is fine-tuned with correponding metric depth information. In this part, we use our ViT-L encoder for fine-tuning.

We examine two representative scenarios: 1) *in-domain* metric depth estimation, where the model is trained and evaluated on the same domain (Section 4.3.1), and 2) *zero-shot* metric depth estimation, where the model is trained on one domain, *e.g.*, NYUv2 [55], but evaluated in different domains, *e.g.*, SUN RGB-D [57] (Section 4.3.2).

4.3.1 In-Domain Metric Depth Estimation

As shown in Table 3 of NYUv2 [55], our model outperforms the previous best method VPD [87] remarkably, improving the δ_1 (↑) from 0.964 → 0.984 and AbsRel (↓) from 0.069 to 0.056. Similar improvements can be observed in Table 4 of the KITTI dataset [18]. We improve the δ_1 (↑) on KITTI from 0.978 → 0.982. It is worth noting that we adopt the ZoeDepth framework for this scenario with a relatively basic depth model, and we believe our results can be further enhanced if equipped with more advanced architectures.

4.3.2 Zero-Shot Metric Depth Estimation

We follow ZoeDepth [4] to conduct zero-shot metric depth estimation. ZoeDepth fine-tunes the MiDaS pre-trained encoder with metric depth information from NYUv2 [55] (for indoor scenes) or KITTI [18] (for outdoor scenes). Therefore, we simply replace the MiDaS encoder with our better Depth Anything encoder, leaving other components unchanged. As shown in Table 5, across a wide range of unseen datasets of indoor and outdoor scenes, our Depth Anything results in a better metric depth estimation model than the original ZoeDepth based on MiDaS.

4.4. Fine-tuned to Semantic Segmentation

In our method, we design our MDE model to inherit the rich semantic priors from a pre-trained encoder via a simple feature alignment constraint. Here, we examine the semantic capability of our MDE encoder. Specifically, we fine-tune our MDE encoder to downstream semantic segmentation datasets. As exhibited in Table 7 of the Cityscapes dataset [15], our encoder from large-scale MDE training (86.2 mIoU) is superior to existing encoders from large-scale ImageNet-21K pre-training, *e.g.*, Swin-L [39] (84.3) and ConvNeXt-XL [41] (84.6). Similar observations hold on the ADE20K dataset [89] in Table 8. We improve the previous best result from 58.3 → 59.4.

We hope to highlight that, witnessing the superiority of our pre-trained encoder on both monocular depth estimation and semantic segmentation tasks, we believe it has great potential to serve as a generic multi-task encoder for both middle-level and high-level visual perception systems.

4.5. Ablation Studies

Unless otherwise specified, we use the ViT-L encoder for our ablation studies here.

Zero-shot transferring of *each* training dataset. In Table 6, we provide the zero-shot transferring performance of *each* training dataset, which means that we train a relative MDE model on *one* training set and evaluate it on the six unseen datasets. With these results, we hope to offer more insights for future works that similarly aim to build a general

Method	SUN RGB-D [57]		iBims-1 [29]		HyperSim [49]		Virtual KITTI 2 [8]		DIODE Outdoor [60]	
	AbsRel (↓)	δ_1 (↑)	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1
ZoeDepth [4]	0.520	0.545	0.169	0.656	0.407	0.302	0.106	0.844	0.814	0.237
Depth Anything	**0.500**	**0.660**	**0.150**	**0.714**	**0.363**	**0.361**	**0.085**	**0.913**	**0.794**	**0.288**

Table 5. **Zero-shot metric** depth estimation. The first three test sets in the header are indoor scenes, while the last two are outdoor scenes. Following ZoeDepth, we use the model trained on NYUv2 for indoor generalization, while use the model trained on KITTI for outdoor evaluation. For fair comparisons, we report the ZoeDepth results reproduced in our environment.

Training set	KITTI [18]		NYUv2 [55]		Sintel [7]		DDAD [20]		ETH3D [52]		DIODE [60]		**Mean**	
	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1
BlendedMVS [77]	*0.089*	0.918	0.068	0.958	*0.556*	0.689	*0.305*	*0.731*	0.148	0.845	0.092	0.921	*0.210*	*0.844*
DIML [13]	0.099	0.907	0.055	*0.969*	0.573	0.722	0.381	0.657	0.142	0.859	0.107	0.908	0.226	0.837
HRWSI [68]	0.095	0.917	0.062	0.966	0.502	0.731	0.270	0.750	0.186	0.775	0.087	0.935	0.200	0.846
IRS [62]	0.105	0.892	*0.057*	0.970	0.568	0.714	0.328	0.691	0.143	0.845	0.088	0.926	0.215	0.840
MegaDepth [33]	0.217	0.741	0.071	0.953	0.632	0.660	0.479	0.566	*0.142*	*0.852*	0.104	0.910	0.274	0.780
TartanAir [63]	0.088	0.920	0.061	0.964	0.602	*0.723*	0.332	0.690	0.160	0.818	*0.088*	*0.928*	0.222	0.841
All labeled data	**0.085**	**0.934**	**0.053**	**0.971**	**0.492**	**0.748**	**0.245**	**0.771**	**0.134**	**0.874**	**0.070**	**0.945**	**0.180**	**0.874**

Table 6. Examine the zero-shot transferring performance of *each labeled training set* (left) to six unseen datasets (top). **Better performance:** AbsRel ↓ , δ_1 ↑. We highlight the **best**, second, and *third best* results for each test dataset in **bold**, underline, and *italic*, respectively.

Method	Encoder	mIoU (s.s.)	m.s.
Segmenter [58]	ViT-L [16]	-	82.2
SegFormer [70]	MiT-B5 [70]	82.4	84.0
Mask2Former [12]	Swin-L [39]	83.3	84.3
OneFormer [24]	Swin-L [39]	83.0	84.4
OneFormer [24]	ConvNeXt-XL [41]	83.6	84.6
DDP [25]	ConvNeXt-L [41]	83.2	83.9
Ours	ViT-L [16]	**84.8**	**86.2**

Table 7. Transferring our MDE pre-trained encoder to **Cityscapes** for semantic segmentation. We ***do not*** use Mapillary [1] for pre-training. s.s./m.s.: single-/multi-scale evaluation.

Method	Encoder	mIoU
Segmenter [58]	ViT-L [16]	51.8
SegFormer [70]	MiT-B5 [70]	51.0
Mask2Former [12]	Swin-L [39]	56.4
UperNet [69]	BEiT-L [2]	56.3
ViT-Adapter [11]	BEiT-L [2]	58.3
OneFormer [24]	Swin-L [39]	57.4
OneFormer [24]	ConNeXt-XL [41]	57.4
Ours	ViT-L [16]	**59.4**

Table 8. Transferring our MDE encoder to **ADE20K** for semantic segmentation. We use Mask2Former as our segmentation model.

monocular depth estimation system. Among the six training datasets, HRWSI [68] fuels our model with the strongest generalization ability, even though it only contains 20K images. This indicates the data diversity counts a lot, which is well aligned with our motivation to utilize unlabeled images. Some labeled datasets may not perform very well, *e.g.*, MegaDepth [33], however, it has its own preferences that are not reflected in these six test datasets. For example, we find models trained with MegaDepth data are specialized at estimating the distance of ultra-remote buildings (Figure 1), which will be very beneficial for aerial vehicles.

Effectiveness of 1) challenging the student model when learning unlabeled images, and 2) semantic constraint. As shown in Table 9, simply adding unlabeled images with pseudo labels does not necessarily bring gains to our model, since the labeled images are already sufficient. However, with strong perturbations ($\mathcal{S}$) applied to unlabeled images during re-training, the student model is challenged to seek additional visual knowledge and learn more robust representations. Consequently, the large-scale unlabeled images enhance the model generalization ability significantly.

Moreover, with our used semantic constraint $\mathcal{L}_{feat}$, the power of unlabeled images can be further amplified for the depth estimation task. More importantly, as emphasized in Section 4.4, this auxiliary constraint also enables our trained encoder to serve as a key component in a multi-task visual system for both middle-level and high-level perception.

Comparison with MiDaS trained encoder in downstream tasks. Our Depth Anything model has exhibited stronger zero-shot capability than MiDaS [5, 46]. Here, we further

Figure 3. Qualitative results on six unseen datasets.

$\mathcal{L}_l$	$\mathcal{L}_u$	$\mathcal{S}$	$\mathcal{L}_{feat}$	KI	NY	SI	DD	ET	DI
✓				0.085	0.053	0.492	0.245	0.134	0.070
✓	✓			0.085	0.054	0.481	0.242	0.138	0.073
✓	✓	✓		0.081	0.048	0.469	0.235	0.134	0.068
✓	✓	✓	✓	**0.076**	**0.043**	**0.458**	**0.230**	**0.127**	**0.066**

Table 9. Ablation studies of: 1) challenging the student with strong perturbations ($\mathcal{S}$) when learning unlabeled images, and 2) semantic constraint ($\mathcal{L}_{feat}$). Limited by space, we only report the AbsRel (↓) metric, and shorten the dataset name with its first two letters.

Method	NYUv2		KITTI		Cityscapes	ADE20K
	AbsRel	δ_1	AbsRel	δ_1	mIoU	mIoU
MiDaS	0.077	0.951	0.054	0.971	82.1	52.4
Ours	**0.056**	**0.984**	**0.046**	**0.982**	**84.8**	**59.4**

Table 10. Comparison between our trained encoder and MiDaS [5] trained encoder in terms of downstream fine-tuning performance. **Better performance:** AbsRel ↓ , δ_1 ↑ , mIoU ↑ .

compare our trained encoder with MiDaS v3.1 [5] trained encoder in terms of the downstream fine-tuning performance. As demonstrated in Table 10, on both the downstream depth estimation task and semantic segmentation task, our produced encoder outperforms the MiDaS encoder remarkably, *e.g.*, 0.951 *vs.* 0.984 in the δ_1 metric on NYUv2, and 52.4 *vs.* 59.4 in the mIoU metric on ADE20K.

Comparison with DINOv2 in downstream tasks. We have demonstrated the superiority of our trained encoder when fine-tuned to downstream tasks. Since our finally produced encoder (from large-scale MDE training) is fine-tuned from DINOv2 [43], we compare our encoder with the original DINOv2 encoder in Table 11. It can be observed that our encoder performs better than the original DINOv2 encoder in both the downstream metric depth estimation task and semantic segmentation task. Although the DINOv2 weight has provided a very strong initialization (also much better than the MiDaS encoder as reported in Table 10), our large-scale and high-quality MDE training can further enhance it impressively in downstream transferring performance.

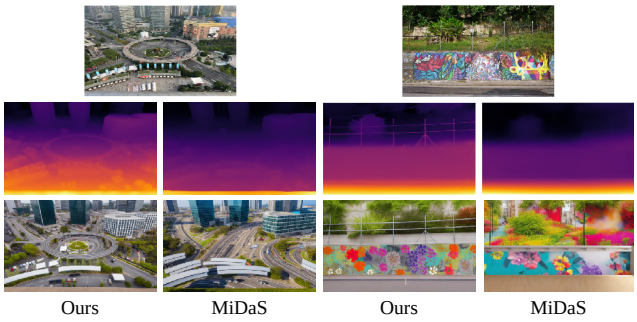

Figure 4. We compare our depth prediction with MiDaS. Meantime, we use ControlNet to synthesize new images from the depth map (the last row). First row: input image, second row: depth prediction.

Encoder	NYUv2		KITTI		ADE20K
	AbsRel (↓)	δ_1 (↑)	AbsRel	δ_1	mIoU (↑)
DINOv2	0.066	0.973	0.058	0.971	58.8
Ours	**0.056**	**0.984**	**0.046**	**0.982**	**59.4**

Table 11. Comparison between the original DINOv2 and our produced encoder in terms of downstream fine-tuning performance.

4.6. Qualitative Results

We visualize our model predictions on the six unseen datasets in Figure 3. Our model is robust to test images from various domains. In addition, we compare our model with MiDaS in Figure 4. We also attempt to synthesis new images conditioned on the predicted depth maps with ControlNet [85]. Our model produces more accurate depth estimation than MiDaS, as well as better synthesis results, although the ControlNet is trained with MiDaS depth. For more accurate synthesis, we have also re-trained a better depth-conditioned ControlNet based on our Depth Anything, aiming to provide better control signals for image synthesis and video editing. Please refer to our project page for more qualitative results on video editing [35] with our Depth Anything.

5. Conclusion

In this work, we present Depth Anything, a highly practical solution to robust monocular depth estimation. Different from prior arts, we especially highlight the value of cheap and diverse unlabeled images. We design two simple yet highly effective strategies to fully exploit their value: 1) posing a more challenging optimization target when learning unlabeled images, and 2) preserving rich semantic priors from pre-trained models. As a result, our Depth Anything model exhibits excellent zero-shot depth estimation ability, and also serves as a promising initialization for downstream metric depth estimation and semantic segmentation tasks.

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

Supplementary Material

6. 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 [43] is 14. Evaluation is performed at the original resolution by interpolating the prediction. Following MiDaS [5, 46], 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 [4]. 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 [55] and 384×768 on KITTI [18] 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 [14]. The training resolution is set as 896×896 on both ADE20K [89] and Cityscapes [15]. The encoder learning rate is set as 3e-6 and the decoder learning rate is 10× larger. We use Mask2Former [12] 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 [36] or Mapillary [1] pre-training. Other training configurations are the same as the original codebase.

7. 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 12, the gap between the tolerance margin of 0 and 0.15 or 0.30 clearly demonstrates the necessity of this design (mean AbsRel: 0.188 *vs.* 0.175).

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 13, 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.

α	KITTI	NYU	Sintel	DDAD	ETH3D	DIODE	**Mean**
0.00	0.085	0.055	0.523	0.250	0.134	0.079	0.188
0.15	0.080	**0.053**	**0.464**	**0.247**	**0.127**	**0.076**	**0.175**
0.30	**0.079**	0.054	0.482	0.248	0.127	0.077	0.178

Table 12. 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 13. Ablation studies of applying our feature alignment loss $\mathcal{L}_{feat}$ to unlabeled data (**U**) or labeled data (**L**).

8. Limitations and Future Works

Currently, the largest model size is only constrained to ViT-Large [16]. 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 [43]. 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 512×512 training resolution is not enough. We plan to re-train our model on a larger resolution of 700+ or even 1000+.

9. More Qualitative Results

Please refer to the following pages for comprehensive qualitative results on six unseen test sets (Figure 5 for KITTI [18], Figure 6 for NYUv2 [55], Figure 7 for Sintel [7], Figure 8 for DDAD [20], Figure 9 for ETH3D [52], and Figure 10 for DIODE [60]). We compare our model with the strongest MiDaS model [5], *i.e.*, DPT-BEiT$_{\text{L-512}}$. Our model exhibits higher depth estimation accuracy and stronger robustness.

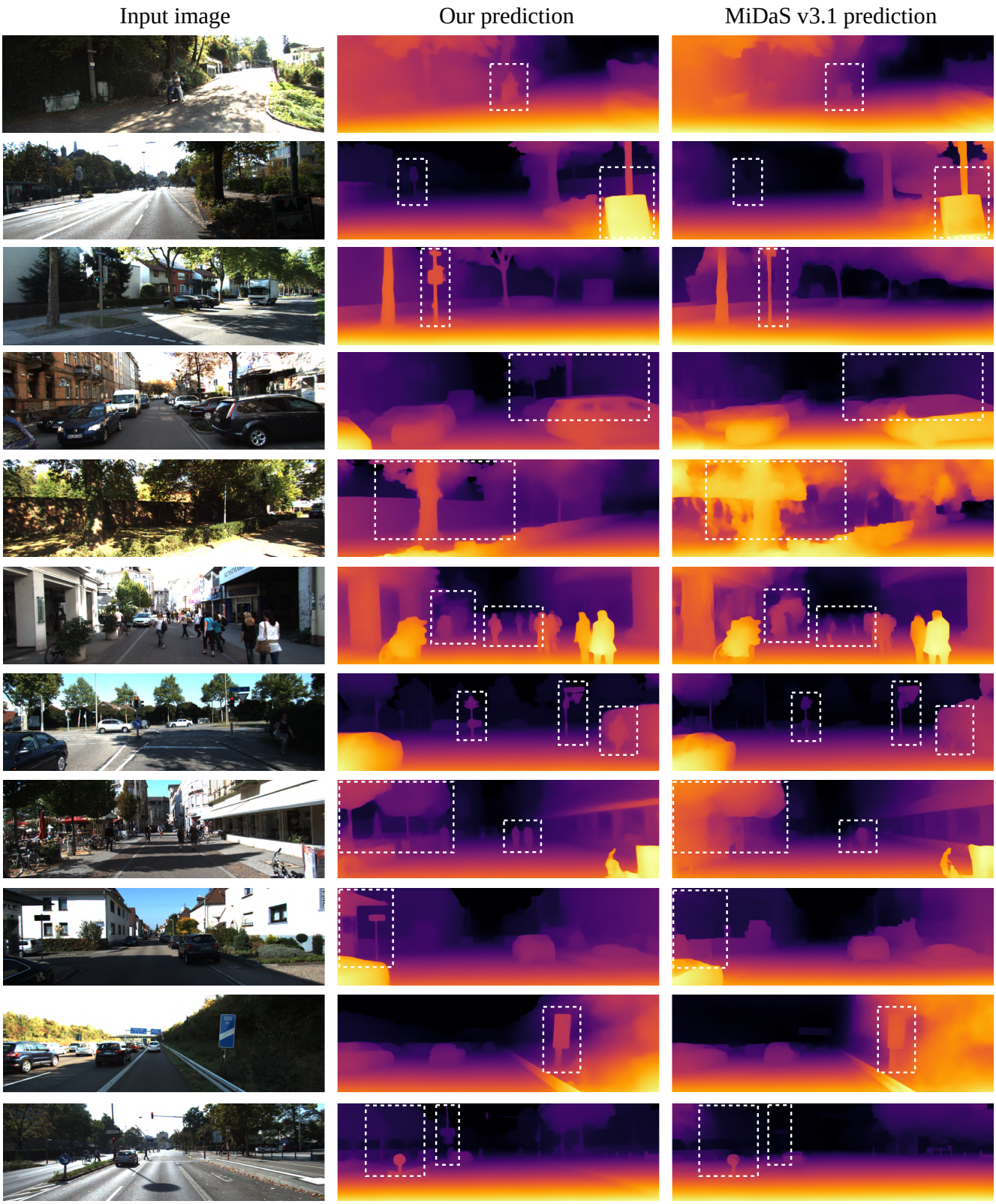

Figure 5. Qualitative results on KITTI. Due to the extremely sparse ground truth which is hard to visualize, we here compare our prediction with the most advanced MiDaS v3.1 [5] prediction. The brighter color denotes the closer distance.

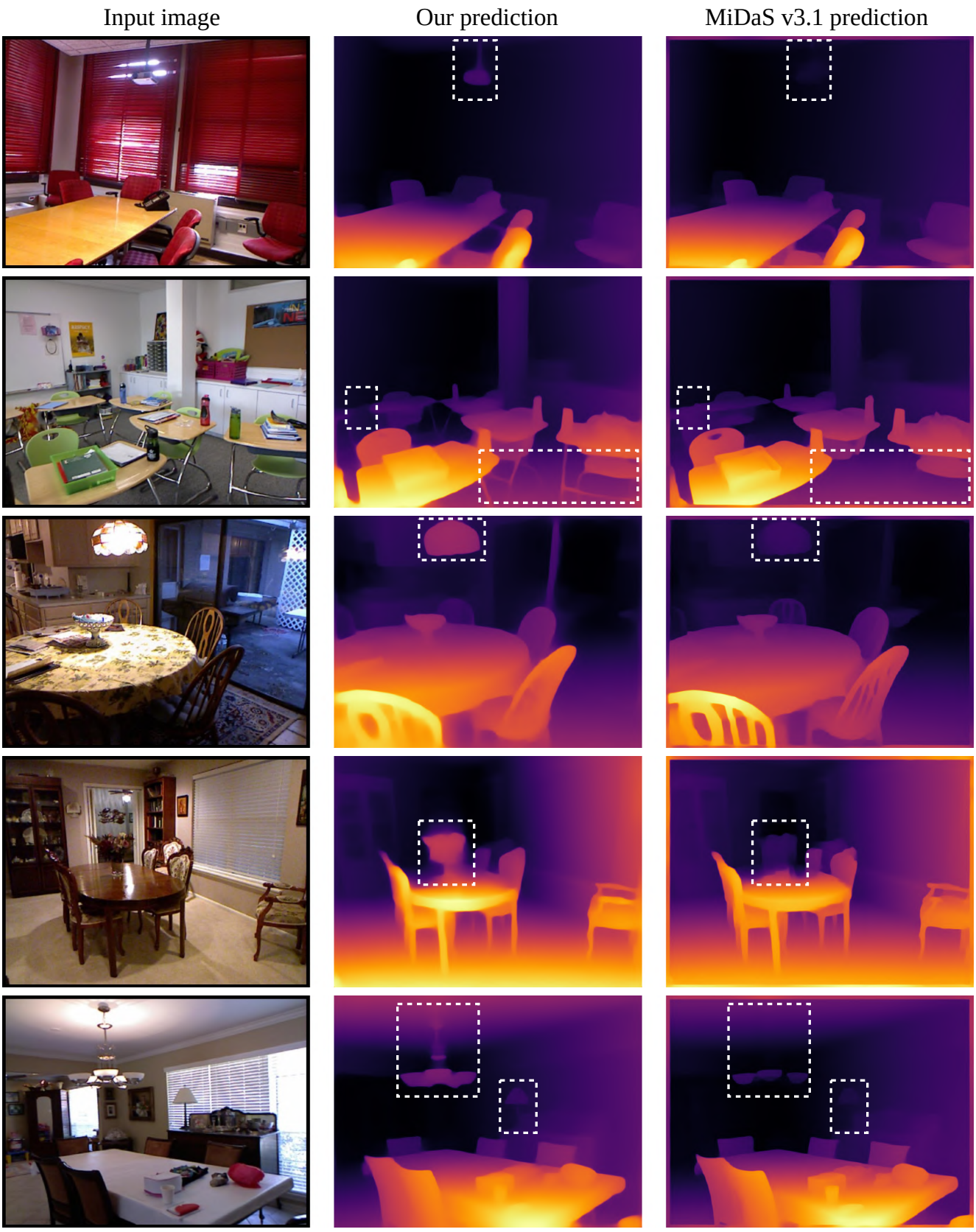

Figure 6. Qualitative results on NYUv2. It is worth noting that MiDaS [5] uses NYUv2 training data (*not zero-shot*), while we do not.

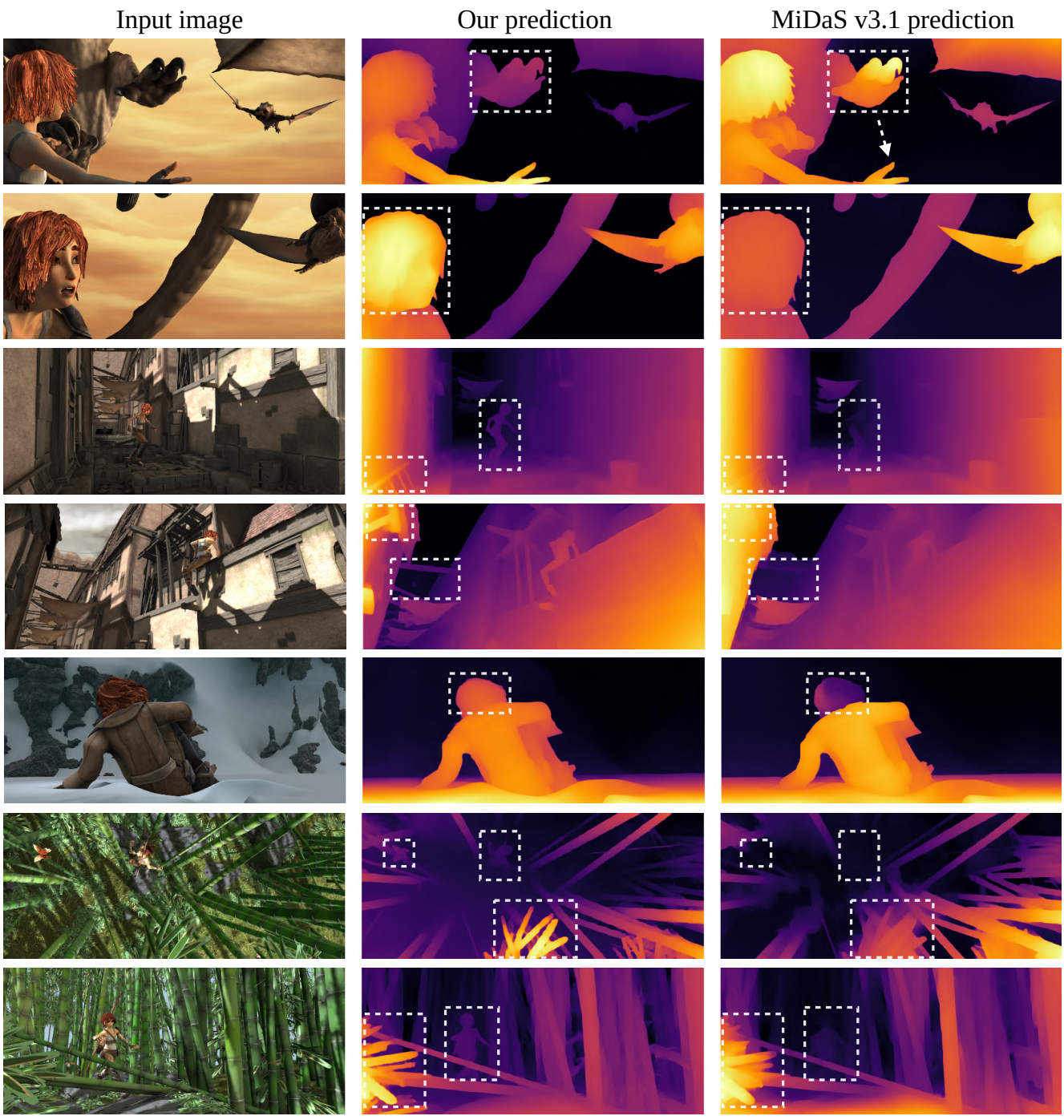

Figure 7. Qualitative results on Sintel.

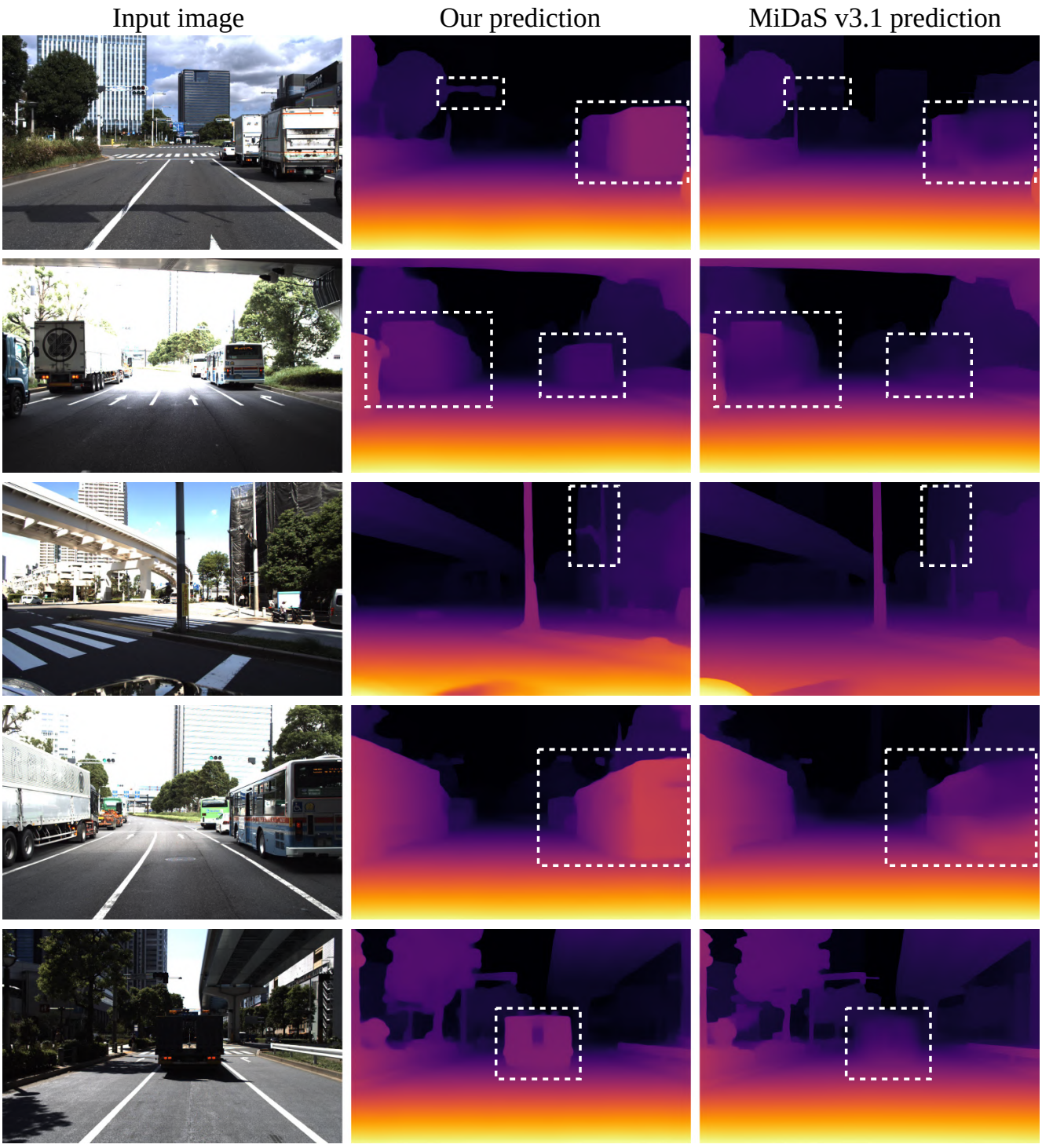

Figure 8. Qualitative results on DDAD.

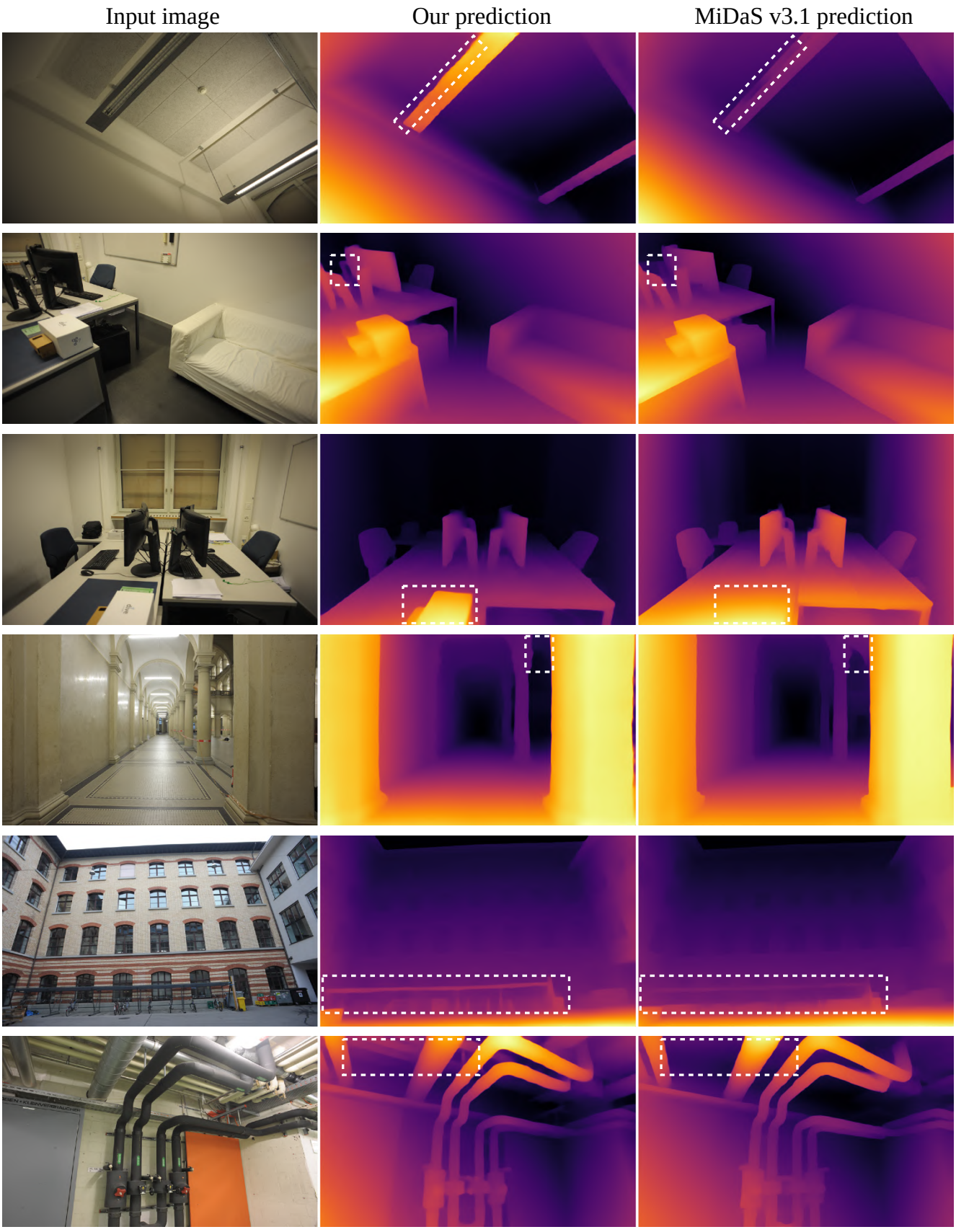

Figure 9. Qualitative results on ETH3D.

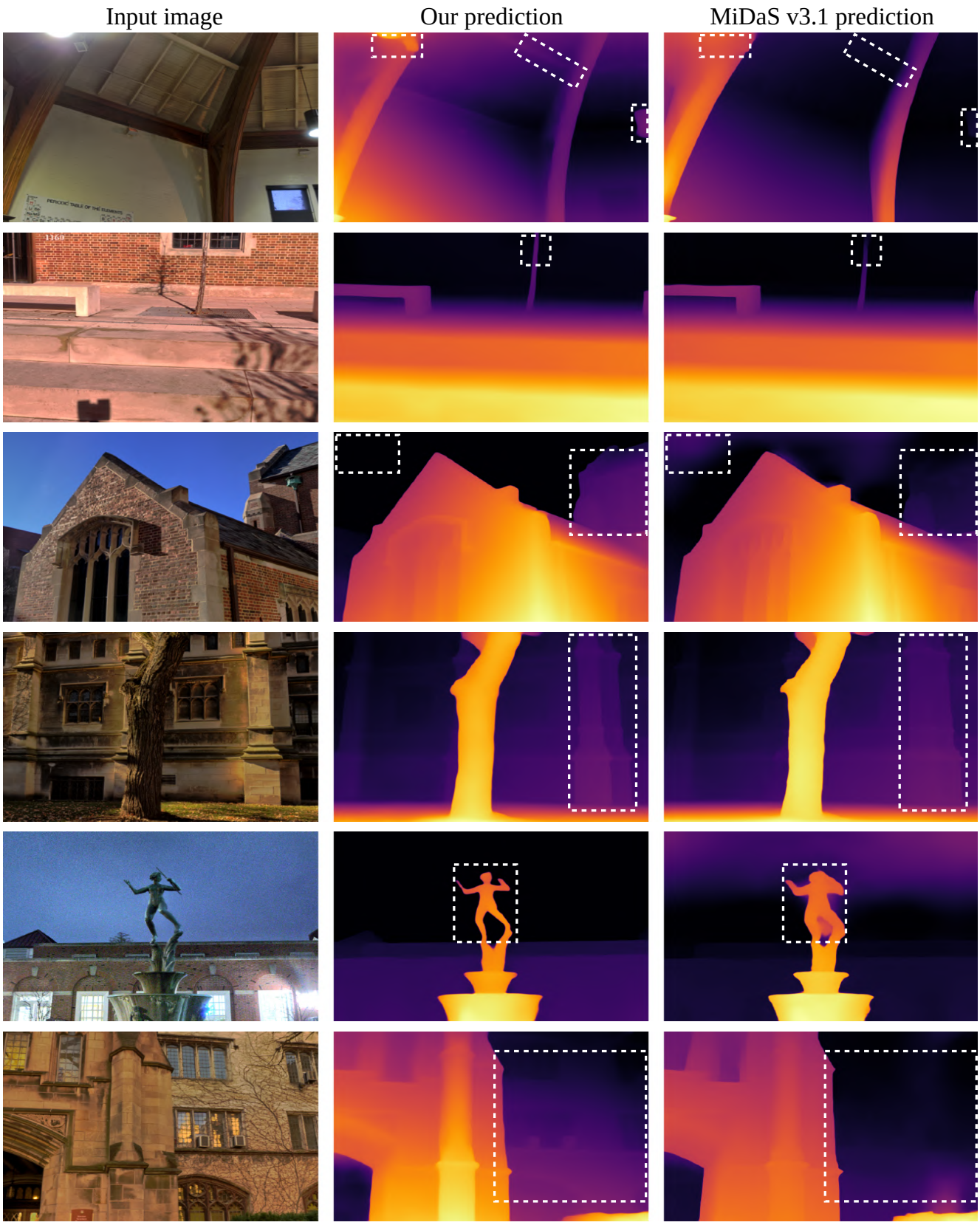

Figure 10. Qualitative results on DIODE.

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. 7, 9

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2022. 7

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, 2021. 2, 6

[4] 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. 2, 6, 7, 9

[5] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv:2307.14460*, 2023. 2, 3, 5, 7, 8, 9, 10, 11

[6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv:2108.07258*, 2021. 1

[7] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 5, 7, 9

[8] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv:2001.10773*, 2020. 7

[9] Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *CVPR*, 2019. 2, 4

[10] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *NeurIPS*, 2016. 2

[11] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *ICLR*, 2023. 7

[12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 7, 9

[13] Jaehoon Cho, Dongbo Min, Youngjung Kim, and Kwanghoon Sohn. Diml/cvl rgb-d dataset: 2m rgb-d images of natural indoor and outdoor scenes. *arXiv:2110.11590*, 2021. 3, 7

[14] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 9

[15] 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, 6, 9

[16] 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. 7, 9

[17] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014. 2

[18] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013. 1, 2, 3, 5, 6, 7, 9

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, 2020. 5

[20] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, 2020. 5, 7, 9

[21] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *ICLR*, 2020. 2, 4

[22] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *ICCV*, 2023. 2

[23] Derek Hoiem, Alexei A Efros, and Martial Hebert. Recovering surface layout from an image. *IJCV*, 2007. 2

[24] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *CVPR*, 2023. 7

[25] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. Ddp: Diffusion model for dense visual prediction. In *ICCV*, 2023. 7

[26] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023. 4

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 1, 2, 3

[28] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *ECCV*, 2020. 4

[29] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *ECCVW*, 2018. 7

[30] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 2, 3

[31] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICMLW*, 2013. 2

[32] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *CVPR*, 2015. 2

[33] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 1, 3, 7

[34] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv:2204.00987*, 2022. 2

[35] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv:2308.14749*, 2023. 8

[36] 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, 9

[37] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T Freeman. Sift flow: Dense correspondence across different scenes. In *ECCV*, 2008. 2

[38] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv:2303.05499*, 2023. 4

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 6, 7

[40] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022. 6

[41] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 6, 7

[42] Jia Ning, Chen Li, Zheng Zhang, Chunyu Wang, Zigang Geng, Qi Dai, Kun He, and Han Hu. All in tokens: Unifying output space of visual tasks via soft token. In *ICCV*, 2023. 6

[43] 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. 3, 4, 5, 8, 9

[44] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3depth: Monocular depth estimation with a piecewise planarity prior. In *CVPR*, 2022. 6

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[46] 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, 2, 3, 5, 7, 9

[47] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 5, 6

[48] Alex Rasla and Michael Beyeler. The relative importance of depth cues and semantic edges for indoor mobility using simulated prosthetic vision in immersive virtual reality. In *VRST*, 2022. 1

[49] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021. 7

[50] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 3

[51] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *TPAMI*, 2008. 2

[52] 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. 5, 7, 9

[53] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 3

[54] Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, and Zhengguo Li. Nddepth: Normal-distance assisted monocular depth estimation. In *ICCV*, 2023. 2, 6

[55] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1, 2, 3, 5, 6, 7, 9

[56] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 2, 4

[57] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 6, 7

[58] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 7

[59] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023. 1

[60] 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. 5, 7, 9

[61] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In *3DV*, 2019. 3

[62] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. In *ICME*, 2021. 3, 7

[63] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and

Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *IROS*, 2020. 3, 7

[64] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, 2019. 1

[65] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *CVPR*, 2020. 3

[66] Diana Wofk, Fangchang Ma, Tien-Ju Yang, Sertac Karaman, and Vivienne Sze. Fastdepth: Fast monocular depth estimation on embedded systems. In *ICRA*, 2019. 1

[67] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *CVPR*, 2018. 2, 3

[68] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *CVPR*, 2020. 2, 3, 7

[69] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 7

[70] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 3, 7

[71] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *ICCV*, 2021. 2

[72] Xiaogang Xu, Hengshuang Zhao, Vibhav Vineet, Ser-Nam Lim, and Antonio Torralba. Mtformer: Multi-task learning via transformer and cross-task reasoning. In *ECCV*, 2022. 4

[73] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv:1905.00546*, 2019. 2

[74] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *CVPR*, 2022. 4

[75] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *CVPR*, 2023. 2

[76] Xiaodong Yang, Zhuang Ma, Zhiyu Ji, and Zhe Ren. Gedepth: Ground embedding for monocular depth estimation. In *ICCV*, 2023. 2, 6

[77] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *CVPR*, 2020. 3, 7

[78] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *ICCV*, 2019. 2

[79] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *ICCV*, 2023. 2

[80] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. In *ICLR*, 2020. 1

[81] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv:1506.03365*, 2015. 3

[82] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 2, 3

[83] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation. *arXiv:2203.01502*, 2022. 2, 6

[84] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 4

[85] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 8

[86] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv:2306.03514*, 2023. 4

[87] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *ICCV*, 2023. 6

[88] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 2017. 3

[89] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 6, 9

[90] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. In *NeurIPS*, 2020. 2