

## A Implementation Details

### A.1 Network architectures

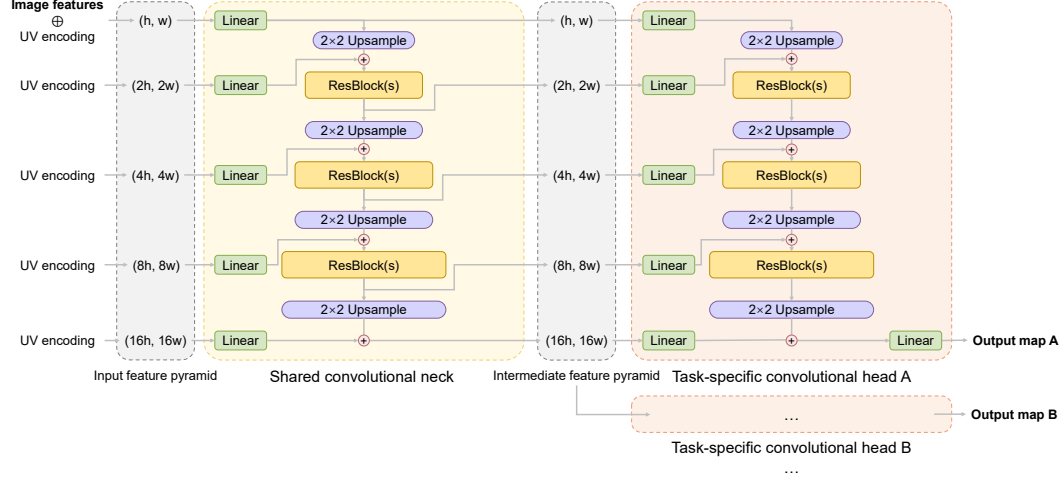


Figure A.1: Illustration of the convolutional neck and head module architectures.

The detailed architectures of our model components are described as follows.

**DINOv2 Image Encoder.** Our model supports variable input resolutions by leveraging the interpolatable positional embeddings of DINOv2 [18]. The native resolution is determined by a user-specified number of image tokens. Given an input image of arbitrary size and a target number of tokens, we compute a patch-level resolution  $h \times w$  that best matches the desired token count. The image is then resized to  $(14h, 14w)$  to match DINOv2’s input requirement, and encoded into  $h \times w$  image tokens along with one classification token. We extract four intermediate feature layers from DINOv2—specifically, the 6th, 12th, 18th, and final transformer layers—project them to a common dimension, reshape their spatial size to  $(h, w)$ , and sum them to form the input for the dense prediction decoder.

**Convolutional Neck and Heads.** Inspired by prior multi-task dense prediction architectures [22, 13, 29], we design a lightweight decoder consisting of a shared convolutional neck and multiple task-specific heads, as illustrated in Fig A.1. Both the neck and the heads are composed of progressive residual convolution blocks (ResBlocks) [10] interleaved with transpose convolution layers (kernel size 2, stride 2) for progressive upsampling from resolution  $(h, w)$  to  $(16h, 16w)$ . Finally, the output map is resized through bilinear interpolation to match the raw image size. To reduce inference latency on modern GPUs, all normalization layers are simply removed from the ResBlocks, without affecting performance or training stability.

At each scale level of the neck, we inject a UV positional encoding, defined as a mapping of the image’s rectangular domain into a unit circle, preserving the raw aspect ratio information. The resulting intermediate feature pyramid is shared across all heads, each of which independently decodes its respective output map. This design enables multi-scale feature sharing while maintaining head-specific decoding tailored to each prediction task.

**CLS-token-conditioned MLP Head.** For scalar prediction, we use a two-layer MLP that takes the CLS token feature from DINOv2 as input and outputs a single scale factor, followed by an exponential mapping to ensure a positive scale output. The hidden layer size is equal to the input feature dimension.

### A.2 Training Data

The datasets used for training our model are listed in Tab. A.1. All datasets are publicly available for academic use, and their sampling weights follow the protocol established in MoGe [29].

Tab. A.2 provides a rough summary of the number of training frames used by several representative monocular geometry estimation methods. As there is no shared or standardized training set in this

field, this table serves to contextualize the scale of training data across methods. Notably, model performance does not necessarily correlate with the amount of training data used.

Table A.1: List of datasets used to train our model.

Name	Domain	#Frames	Type	Weight	Metric Scale
A2D2[8]	Outdoor/Driving	196K	LiDAR	0.8%	✓
Argoverse2[31]	Outdoor/Driving	1.1M	LiDAR	7.1%	✓
ARKitScenes[2]	Indoor	449K	SfM	8.3%	✓
BlendedMVS[35]	In-the-wild	115K	SfM	11.5%	
MegaDepth[16]	Outdoor/In-the-wild	92K	SfM	5.4%	
ScanNet++[5]	Indoor	176K	SfM	4.6%	✓
Taskonomy[37]	Indoor	3.6M	SfM	14.1%	✓
Waymo[25]	Outdoor/Driving	788K	LiDAR	6.2%	✓
ApolloSynthetic[1]	Outdoor/Driving	194K	Synthetic	3.8%	✓
EDEN[36]	Outdoor/Garden	369K	Synthetic	1.2%	
GTA-SfM[27]	Outdoor/In-the-wild	19K	Synthetic	2.7%	✓
Hypersim[23]	Indoor	75K	Synthetic	4.8%	✓
IRS[28]	Indoor	101K	Synthetic	5.4%	✓
KenBurns[17]	In-the-wild	76K	Synthetic	1.5%	
MatrixCity[15]	Outdoor/Driving	390K	Synthetic	1.3%	✓
MidAir[7]	Outdoor/In-the-wild	423K	Synthetic	3.8%	✓
MVS-Synth[12]	Outdoor/Driving	12K	Synthetic	1.2%	✓
Structured3D[38]	Indoor	77K	Synthetic	4.6%	✓
Synthia[24]	Outdoor/Driving	96K	Synthetic	1.1%	✓
Synscapes[32]	Outdoor/Driving	25K	Synthetic	1.9%	✓
UnrealStereo4K [26]	In-the-wild	8K	Synthetic	1.6%	✓
TartanAir[30]	In-the-wild	306K	Synthetic	4.8%	✓
UrbanSyn[9]	Outdoor/Driving	7K	Synthetic	2.0%	✓
ObjaverseV1[6]	Object	167K	Synthetic	4.6%	

Table A.2: Summary of labeled training frame counts and pretrained backbones for the models compared in this paper.

Method	#Total Training Frames	Pretrained Backbone
ZoeDepth [3]	~ 2M	MiDaS BEiT384-L [21]
DA V1 [33]	1.5M (+ 62M pseudo-labeled)	DINOv2 ViT-Large
DA V2 [34]	595K (+ 62M pseudo-labeled)	DINOv2 ViT-Large
Metric3D V2 [11]	16M	DINOv2 ViT-Large
UniDepth V1 [19]	3.7M	DINOv2 ViT-Large
UniDepth V2 [20]	16M	DINOv2 ViT-Large
Depth Pro [4]	~ 6M	DINOv2 ViT-Large
MoGe [29]	9M	DINOv2 ViT-Large
<i>Ours</i>	8.9M	DINOv2 ViT-Large

### A.3 Evaluation Protocol

**Relative Geometry** We follow the evaluation protocol of alignment in MoGe [29]. Predictions and ground truth are aligned in scale (and shift, if applicable) for each image before measuring errors as specified below

- **Scale-invariant point map.** The scale  $a^*$  to align prediction with ground truth is computed as:

$$a^* = \operatorname{argmin}_a \sum_{i \in \mathcal{M}} \frac{1}{z_i} \|a\hat{\mathbf{p}}_i - \mathbf{p}_i\|_1, \quad (1)$$

- **Affine-invariant point map.** The scale  $a^*$  and shift  $\mathbf{b}^*$  are computed as:

$$(a^*, \mathbf{b}^*) = \operatorname{argmin}_{a, \mathbf{b}} \sum_{i \in \mathcal{M}} \frac{1}{z_i} \|a\hat{\mathbf{p}}_i + \mathbf{b} - \mathbf{p}_i\|_1. \quad (2)$$

- **Scale-invariant depth map,** the scale  $a^*$  is computed as

$$a^* = \operatorname{argmin}_s \sum_{i \in \mathcal{M}} \frac{1}{z_i} |a\hat{z}_i - z_i|. \quad (3)$$

45 • **Affine-invariant depth map.** The scale  $a^*$  and shift  $b^*$  are computed as

$$(a^*, b^*) = \operatorname{argmin}_s \sum_{i \in \mathcal{M}} \frac{1}{z_i} |a\hat{z}_i + b - z_i|. \quad (4)$$

46 • **Affine-invariant disparity map.** We follow the established protocol for affine disparity  
47 alignment [21], using least-squares to align predictions in disparity space:

$$(a^*, b^*) = \operatorname{argmin}_s \sum_{i \in \mathcal{M}} (a\hat{d}_i + b - d_i)^2, \quad (5)$$

48 where  $\hat{d}_i$  is the predicted disparity and  $d_i$  is the ground truth, defined as  $d_i = 1/z_i$ . To  
49 prevent aligned disparities from taking excessively small or negative values, the aligned  
50 disparity is truncated by the inverted maximum depth  $1/z_{\max}$  before inversion. The final  
51 aligned depth  $\hat{z}_i^*$  is computed as:

$$\hat{z}_i^* := \frac{1}{\max(a^*\hat{d}_i + b^*, 1/z_{\max})}. \quad (6)$$

## 52 Metric Geometry

- 53 • **Metric depth.** The output is evaluated without alignment and clamping range of values for  
54 all methods, unless specific post-processing is hard-coded in its model inference pipeline.
- 55 • **Metric point map.** The point map prediction is aligned with the ground truth by the optimal  
56 translation:

$$\mathbf{b}^* = \operatorname{argmin}_{\mathbf{b}} \sum_{i \in \mathcal{M}} \frac{1}{z_i} \|\hat{\mathbf{p}}_i + \mathbf{b} - \mathbf{p}_i\|_1. \quad (7)$$

## 57 B Additional Experiments and Results

### 58 B.1 Test-time Resolution Scaling

59 In ViT-based models, the native input resolution is determined by the number of image tokens derived  
60 from fixed-size patches, specifically,  $14^2$  for DINOv2 models. As such, resolution scaling can be  
61 effectively studied through varying token counts. Our model is trained across a wide range of token  
62 counts from 1200 to 3600, corresponding to native input resolutions ranging approximately from  
63  $484^2$  to  $1188^2$ . This training setup enables robust generalization to a broad range of resolutions and  
64 flexible usage with details as follows.

65 **Geometry Accuracy** MoGe [29] and UniDepth V2 [19] are both trained on diverse input resolutions  
66 and aspect ratios, which helps them maintain accuracy under resolution shifts within a moderate  
67 range (1200 - 3000). In contrast, models such as Depth Anything [33, 34] and Metric3D V2 [11] are  
68 trained with fixed input resolution and exhibit substantial performance degradation when evaluated at  
69 resolutions that diverge from their training setting. Our method, trained over a broader resolution  
70 spectrum, remains robust under test-time scaling. As shown in Fig. B.3a, it maintains the top accuracy  
71 when scaled up for improved detail or down for faster inference—even beyond the training range.

72 **Boundary Sharpness** Higher input resolutions and more image tokens generally lead to sharper  
73 boundaries in dense prediction tasks, as observed in prior works [34, 22, 14] and also shown in  
74 Fig. B.2. In Fig. B.3b, we evaluate several DINOv2-based methods for boundary sharpness at  
75 different test-time resolutions. Note that Depth Pro operates at a fixed high resolution due to its  
76 specialized multi-scale, patch-based architecture. Our approach consistently delivers the sharpest  
77 predictions at each resolution and outperforms Depth Pro using significantly fewer tokens to reach  
78 similar levels of detail.

79 **Latency Trade-off** As shown in Fig. B.3c, inference latency scales roughly linearly with the  
80 number of tokens. Although all compared methods share the same ViT backbone, overall runtime  
81 can vary due to differences in decoder complexity and architectural choices. Our model adopts  
82 a lightweight design that enables fast inference while maintaining strong accuracy, achieving a  
83 favorable trade-off between latency and performance across a wide range of resolutions—within a  
84 single unified framework.

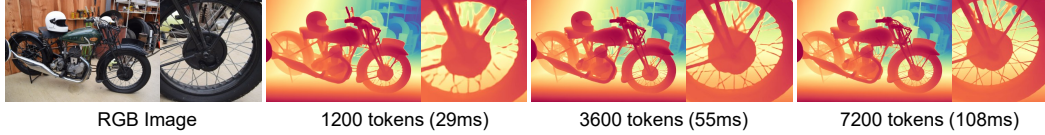


Figure B.2: Increasing the number of image tokens to trade-off latency for visual sharpness.

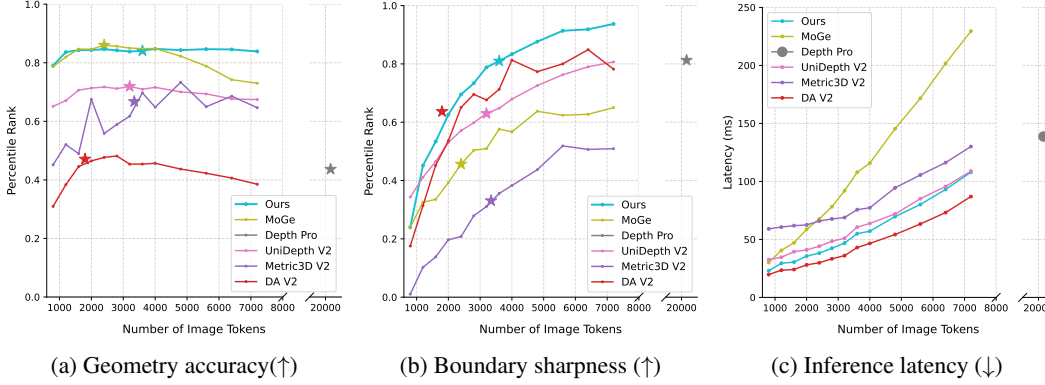


Figure B.3: Performance comparison under test-time resolution scaling. ★ denotes the default configuration for each method. (a) Percentile rank ( $\frac{x - \text{worst}}{\text{best} - \text{worst}}$ ) averaged across all evaluated datasets and two geometry metrics (metric and relative geometry accuracy). (b) Average percentile rank for boundary sharpness. Both are evaluated on a 1/10 subset uniformly sampled from the evaluation benchmarks. (c) Inference latency measured on an NVIDIA A100 GPU with FP16 precision. Our method demonstrates the most favorable balance between latency and performance across different resolutions.

## B.2 Runtime Analysis

As shown in Table B.3, we evaluate the runtime performance of each method under their representative test-time configurations. Specifically, we measure single-frame inference latency and peak GPU memory usage on an NVIDIA A100 GPU. These metrics provide a practical comparison of computational efficiency and resource requirements across different architectures.

Table B.3: Runtime statistics measured on a single NVIDIA A100 GPU for single-frame inference.

Method	#Parameters	#Tokens	Native Resolution	Latency (ms)		Memory (GB)	
				FP16	FP32	FP16	FP32
DA V2	335M	1369	518 <sup>2</sup>	24	86	0.91	1.8
Metric3D V2	412M	3344	1064×616	87	255	1.4	2.3
UniDepth	354M	1020	448 <sup>2</sup>	33	84	1.1	1.8
		3061	774 <sup>2</sup>	50	206	1.8	2.5
Depth Pro	504M	20160	1536 <sup>2</sup>	139	906	8.0	3.7
MoGe	314M	1200	484 <sup>2</sup>	40	93	0.74	1.4
		2500	700 <sup>2</sup>	70	192	0.88	1.6
<i>Ours</i>	326M	1200	484 <sup>2</sup>	29	82	0.96	1.7
		2500	700 <sup>2</sup>	39	157	1.1	2.1
		3600	840 <sup>2</sup>	55	238	1.3	2.5
		7200	1188 <sup>2</sup>	108	565	1.9	3.8

## B.3 More Visual Results

More visual results for qualitative comparison are included in Fig. B.4 and Fig. B.5.



## 92 B.4 Complete Evaluation on Individual Datasets

93 In the paper, we only listed the average performance across multiple datasets for qualitative compari-  
 94 son and ablation study. Table B.4 and Table B.5 list all the results for each individual datasets.

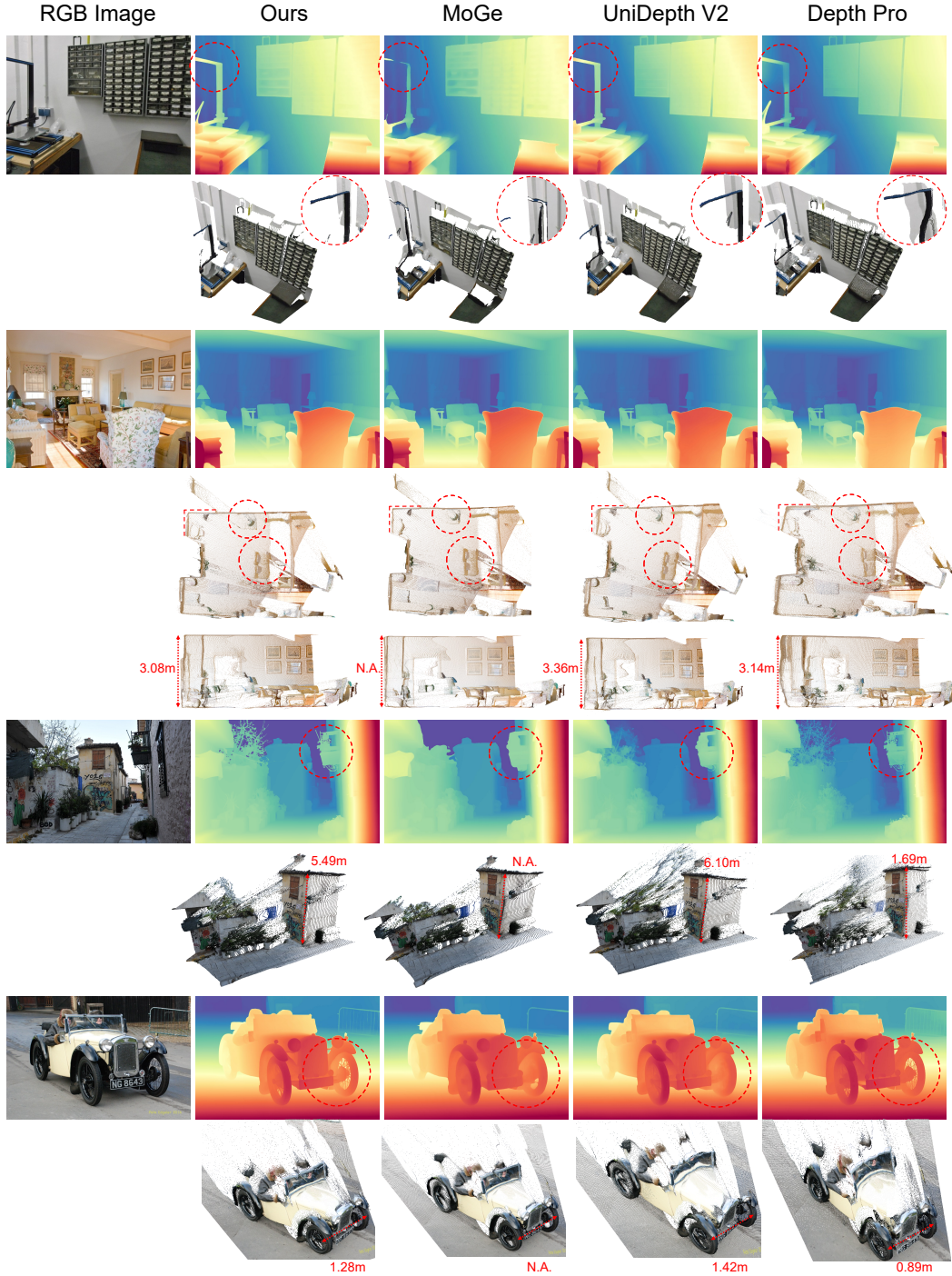


Figure B.4: More visual results on open-domain images (1/2). *Best viewed zoomed in.*

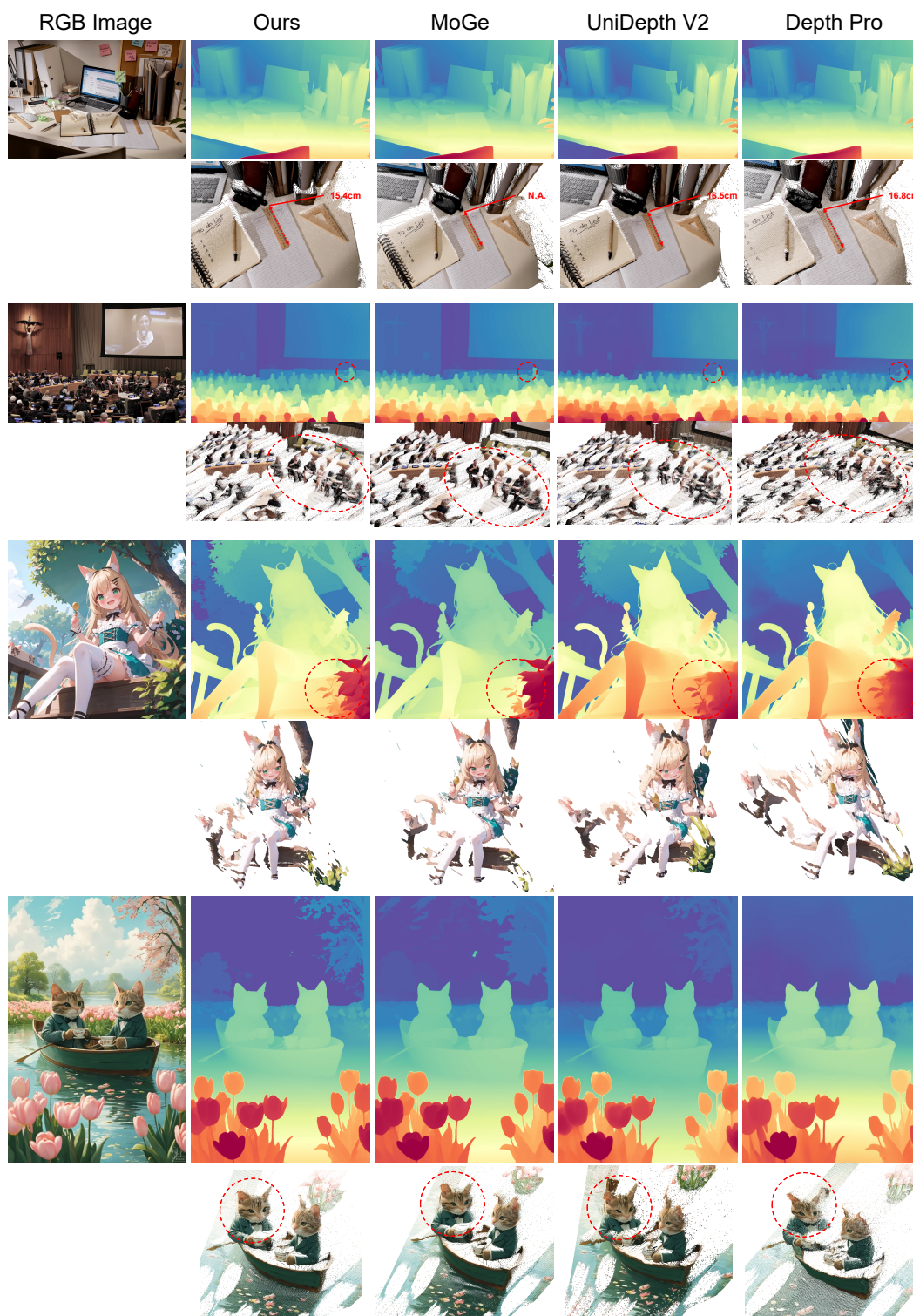


Figure B.5: More visual results on open-domain images (2/2). *Best viewed zoomed in.*



Method	NYUv2 Rel↓ $\delta_1$ ↑	KITTI Rel↓ $\delta_1$ ↑	ETH3D Rel↓ $\delta_1$ ↑	iBims-1 Rel↓ $\delta_1$ ↑	GSO Rel↓ $\delta_1$ ↑	Sintel Rel↓ $\delta_1$ ↑	DDAD Rel↓ $\delta_1$ ↑	DIODE Rel↓ $\delta_1$ ↑	Spring Rel↓ $\delta_1$ ↑	HAMMER Rel↓ $\delta_1$ ↑	Avg. Rel↓ $\delta_1$ ↑ Rank↓
Metric point map											
MASt3R	7.11 95.6	26.0 45.8	27.4 43.1	10.1 89.3	- -	- -	35.4 28.7	21.7 66.3	- -	56.0 18.3	26.2 55.3 4.93
UniDepth V1	4.80 98.3	4.52 98.5	22.4 63.1	10.8 92.8	- -	- -	11.4 89.5	12.8 88.9	- -	18.0 79.5	12.1 87.2 2.71
UniDepth V2	4.83 98.0	5.88 97.5	9.46 95.0	5.23 97.9	- -	- -	13.3 90.3	17.0 80.8	- -	15.0 83.9	10.1 91.9 2.43
Depth Pro	6.13 97.3	11.1 85.3	21.2 64.9	6.89 96.9	- -	- -	22.6 61.3	13.5 81.8	- -	14.5 86.0	13.7 81.9 3.29
Ours	4.44 98.3	7.44 94.4	7.19 97.7	5.63 97.4	- -	- -	11.4 87.9	7.85 92.3	- -	13.4 87.0	8.19 93.6 1.64
Metric depth map (wo/ GT intrinsics)											
ZoeDepth	11.0 91.9	17.0 85.4	57.1 33.7	17.4 67.2	- -	- -	38.9 38.6	39.3 29.3	- -	94.3 3.23	39.3 49.9 5.90
MASt3R	10.8 99.7	56.7 98.4	47.2 20.1	18.7 61.5	- -	- -	62.4 5.51	54.9 19.0	- -	97.2 6.74	49.7 30.3 6.71
DA V1	10.5 94.9	11.6 94.5	40.2 24.0	12.9 81.8	- -	- -	34.5 44.7	58.0 16.2	- -	54.8 27.3	31.8 54.8 5.50
DA V2	16.4 80.9	10.6 88.6	36.1 36.3	11.1 91.7	- -	- -	41.7 37.5	41.2 22.1	- -	52.1 38.9	29.9 56.6 4.43
UniDepth V1	7.59 97.6	4.69 98.4	56.9 14.9	23.8 57.6	- -	- -	13.8 85.1	17.1 71.9	- -	38.2 46.7	23.2 67.5 3.32
UniDepth V2	10.6 92.8	8.58 95.4	20.7 69.5	9.52 93.2	- -	- -	18.4 77.6	43.0 51.8	- -	38.2 46.8	21.3 75.3 2.54
Depth Pro	10.7 91.9	23.5 38.3	38.5 32.8	15.9 81.5	- -	- -	33.4 35.3	31.9 37.7	- -	39.1 63.0	27.6 54.4 4.36
Ours	7.33 96.1	18.1 62.9	10.4 90.8	13.6 83.0	- -	- -	15.8 73.0	17.5 66.4	- -	26.9 65.6	15.7 76.8 2.21
Metric depth map (w/ GT intrinsics)											
Metric3D V2	7.16 96.5	5.25 98.0	11.8 88.8	9.96 94.1	- -	- -	9.21 93.7	49.1 1.98	- -	35.7 44.3	18.3 73.9 2.75
UniDepth V1	5.98 97.9	4.43 98.5	44.5 26.7	22.6 60.5	- -	- -	13.0 87.2	21.0 63.5	- -	38.6 45.9	21.4 68.6 2.50
UniDepth V2	7.81 96.0	5.98 97.7	15.0 85.2	7.71 95.5	- -	- -	14.1 89.3	41.0 67.1	- -	37.7 47.1	18.5 82.6 2.57
Ours	6.46 96.9	8.64 93.7	10.5 92.2	9.92 92.4	- -	- -	13.1 85.6	16.2 77.1	- -	30.4 74.2	13.6 87.4 2.00
Scale-invariant point map											
MASt3R	6.26 96.0	10.0 93.8	6.28 95.5	7.55 95.1	5.03 99.0	31.5 50.2	15.9 77.6	12.8 85.0	39.3 33.7	10.7 95.0	14.5 82.1 5.45
UniDepth V1	5.33 98.4	5.96 98.5	18.5 77.6	5.29 97.4	6.58 99.6	33.0 48.9	11.4 90.2	12.3 91.0	33.1 49.8	4.83 98.5	13.6 85.0 3.83
UniDepth V2	5.59 98.3	5.41 98.0	6.58 97.2	5.56 98.1	4.53 99.7	27.2 56.3	13.4 91.2	12.0 93.4	31.9 46.0	4.20 99.2	11.6 87.7 2.98
Depth Pro	5.04 97.7	10.6 95.1	11.2 92.0	5.84 97.1	4.94 99.8	26.9 63.9	15.8 81.0	8.52 91.6	28.1 60.5	6.82 98.7	12.4 87.7 3.83
MoGe	4.86 98.4	5.47 97.4	4.58 98.9	4.63 97.1	2.58 100	22.3 69.5	12.3 90.3	6.58 94.5	4.84 96.4	6.45 98.1	7.46 94.1 2.14
Ours	3.94 98.3	8.27 97.5	5.45 98.6	5.34 98.3	2.55 100	23.1 66.8	11.0 90.7	8.42 93.7	31.1 42.4	8.77 98.4	10.8 88.5 2.40
Affine-invariant point map											
MASt3R	5.30 96.3	8.32 92.3	5.48 96.6	5.72 95.0	3.50 99.2	26.3 62.8	14.7 79.6	8.10 90.1	33.3 51.1	5.34 96.6	11.6 86.0 5.45
UniDepth V1	3.93 98.4	4.29 98.6	12.2 89.6	4.65 98.0	2.99 99.8	28.5 58.4	10.3 90.5	8.56 90.9	29.6 58.5	4.15 98.7	10.9 88.1 3.95
UniDepth V2	3.66 98.4	4.75 98.0	4.35 98.4	4.05 98.1	2.91 99.9	17.9 76.5	12.0 90.8	7.45 92.4	25.1 66.9	3.45 99.4	8.56 91.9 2.55
Depth Pro	4.36 97.9	9.15 90.7	7.73 94.0	4.34 97.4	3.16 99.7	19.6 74.5	14.4 81.2	6.28 93.7	25.0 66.0	5.31 98.8	9.93 89.4 4.30
MoGe	3.68 98.3	4.86 97.2	3.57 99.0	3.61 97.3	1.14 100	16.8 77.8	10.5 91.4	4.37 96.4	4.45 96.4	3.88 98.1	5.69 95.2 2.14
Ours	3.33 98.4	6.47 96.4	3.89 98.7	3.65 98.5	1.16 100	17.4 77.0	10.1 90.3	5.13 94.9	24.5 63.7	4.19 99.1	7.98 91.7 2.23
Local point map											
MASt3R	- -	- -	5.54 95.3	6.19 95.0	- -	- -	11.4 87.9	8.58 91.8	8.75 90.9	- -	8.09 92.2 5.40
UniDepth V1	- -	- -	8.61 92.6	5.92 96.0	- -	- -	13.4 84.3	8.18 92.0	9.95 90.0	- -	9.21 91.0 5.55
UniDepth V2	- -	- -	3.99 97.4	4.02 97.3	- -	- -	9.35 92.2	8.18 92.4	6.15 95.3	- -	6.34 94.9 3.10
Depth Pro	- -	- -	4.76 96.9	4.11 97.5	- -	- -	10.8 89.5	8.08 92.4	6.80 94.1	- -	6.91 94.1 3.55
MoGe	- -	- -	3.21 98.1	4.16 96.8	- -	- -	8.63 92.7	6.74 94.3	4.78 96.3	- -	5.50 95.6 2.05
Ours	- -	- -	3.27 98.2	3.61 97.7	- -	- -	8.13 93.2	6.57 94.3	5.09 96.1	- -	5.33 95.9 1.35
Scale-invariant depth map											
ZoeDepth	5.62 96.3	7.27 91.9	10.4 87.3	7.45 93.2	3.23 99.9	27.4 61.8	17.0 72.8	11.3 85.2	30.3 55.9	7.42 94.7	12.7 83.9 8.75
MASt3R	5.37 96.0	6.24 94.5	5.68 95.5	5.58 95.2	3.72 99.1	26.3 63.7	13.5 81.5	8.37 89.4	32.2 53.5	5.50 96.5	11.2 86.5 7.65
DA V1	4.77 97.5	5.61 95.6	9.41 88.9	5.53 95.8	5.49 99.3	28.3 56.7	13.2 81.5	10.3 87.5	27.3 59.1	6.88 96.4	11.7 85.8 8.22
DA V2	5.03 97.3	7.23 93.7	6.12 95.5	4.32 97.9	4.38 99.3	23.0 65.2	14.7 78.0	7.95 90.0	28.0 61.1	5.92 97.7	10.7 87.6 6.80
Metric3D V2	4.69 97.4	4.00 98.5	3.84 98.5	4.23 97.7	2.46 99.9	20.7 69.8	7.41 94.6	3.29 98.4	24.4 64.4	4.19 99.1	7.92 91.8 3.39
UniDepth V1	3.86 98.4	3.73 98.6	5.67 97.0	4.79 97.4	4.18 99.7	28.3 58.8	10.1 90.5	6.83 92.8	29.2 59.3	4.19 98.4	10.1 89.1 5.12
UniDepth V2	3.65 98.4	4.24 98.0	3.23 98.9	3.45 98.1	3.16 99.7	23.1 65.3	11.0 91.5	5.92 94.1	24.9 65.1	3.48 99.1	8.61 90.8 3.10
Depth Pro	4.42 97.6	5.47 96.2	7.54 94.1	4.13 97.4	2.18 99.9	23.9 68.7	14.0 82.0	7.05 92.0	25.1 63.8	4.36 98.9	9.81 89.1 5.33
MoGe	3.44 98.4	4.25 97.8	3.36 98.9	3.46 97.0	1.47 100	19.3 73.4	9.17 90.5	4.89 94.7	4.63 96.4	3.77 98.1	5.77 94.5 2.72
Ours	3.44 98.2	4.11 98.0	3.55 98.7	3.16 98.2	1.49 100	19.6 71.6	8.91 91.2	5.30 94.6	20.0 72.4	3.96 99.2	7.35 92.2 2.12
Affine-invariant depth											
ZoeDepth	4.76 97.3	5.59 95.1	7.27 94.2	5.85 95.7	2.54 99.9	21.8 69.2	14.2 80.1	7.80 90.9	24.3 66.6	6.65 95.7	10.1 88.5 9.09
MASt3R	4.67 96.7	5.79 95.1	4.64 97.0	4.62 95.6	2.85 99.4	21.3 70.3	12.5 83.4	5.79 94.1	27.4 62.8	4.21 96.8	9.38 89.1 7.97
DA V1	3.82 98.3	5.04 96.4	6.23 95.2	4.23 97.3	1.98 100	20.1 71.8	11.3 86.1	6.75 92.6	22.4 68.9	5.77 97.3	8.76 90.4 6.91
DA V2	4.16 97.9	6.77 94.3	4.63 97.2	3.44 98.3	1.44 100	17.1 76.6	13.4 81.8	5.41 94.6	23.7 68.7	4.73 98.9	8.48 90.8 6.15
Metric3D V2	3.94 97.6	3.50 98.4	3.24 99.0	3.28 98.3	2.10 99.4	26.6 71.7	7.15 94.8	2.75 98.7	21.0 72.5	3.02 99.0	7.66 92.9 4.53
UniDepth V1	3.40 98.6	3.55 98.7	4.92 97.5	3.76 98.2	2.48 99.9	24.9 64.1	9.46 90.8	4.90 96.2	25.2 67.3	3.55 98.9	8.61 91.0 5.67
UniDepth V2	2.96 98.6	3.85 98.1	2.95 98.5	2.64 98.4	1.37 100	13.3 83.2	10.5 90.9	4.05 96.5	20.1 75.4	2.48 99.6	6.42 93.9 2.80
Depth Pro	3.67 98.2	5.12 96.8	4.97 96.4	3.23 98.3	1.46 100	15.8 80.1	12.6 84.1	4.66 95.6	21.7 70.5	3.30 99.6	7.65 92.0 5.05
MoGe	2.92 98.6	3.94 98.0	2.69 99.2	2.74 97.9	0.94 100	13.0 83.2	8.40 92.1	3.16 97.5	4.34 96.4	3.00 98.3	4.51 96.1 2.94
Ours	2.89 98.6	3.75 98.1	2.80 99.1	2.36 98.8	0.94 100	13.3 82.5	8.26 92.5	3.14 97.4	15.9 81.2	2.85 99.3	5.62 94.8 2.02
Affine-invariant disparity											
ZoeDepth	5.21 97.7	5.84 95.6	8.07 94.0	6.19 96.1	2.60 99.9	26.9 66.3	14.1 81.7	8.17 92.0	27.2 63.0	6.84 96.4	11.1 88.3 8.78
DA V1	4.20 98.4	5.40 97.0	4.68 98.2	4.18 97.6	1.54 100	20.2 77.6	12.7 86.9	5.69 95.7	22.2 72.5	5.56 98.0	8.63 92.2 5.62
DA V2	4.14 98.3	5.61 96.7	4.71 97.9	3.47 98.5	1.24 100	21.4 72.8	13.1 86.4	5.29 96.1	24.3 70.6	4.97 99.1	8.82 91.6 5.42
Metric3D V2	13.4 81.5	3.76 98.2	4.30 97.7	8.55 92.3	1.80 100	21.8 72.4	7.35 94.1	7.70 90.2	23.3 68.1	3.17 99.2	9.51 89.4 6.17
MASt3R	5.07 96.8	5.93 95.5	5.25 96.4	5.39 95.7	2.98 99.7	30.2 65.1	13.0 83.6	6.41 94.3	37.3 53.2	4.41 97.2	11.6 87.8 8.60
UniDepth V1	3.78 98.7	3.64 98.7	5.34 97.2	4.06 98.1	2.56 99.9	28.6 60.7	9.94 89.1	5.95 95.5	30.0 61.6	3.64 99.1	9.75 89.9 5.92
UniDepth V2	3.38 98.7	3.99 98.0	2.97 99.0	3.15 98.3	1.30 100	17.2 79.9	10.2 90.2	4.43 96.4	24.4 69.6	2.51 99.6	7.35 93.0 2.75
Depth Pro	4.21 98.1	5.10 97.0	4.94 96.7	3.74 98.2	1.49 100	17.4 79.1	11.7 87.1	4.84 96.4	27.5 64.5	3.31 99.6	8.42 91.7 5.08
MoGe	3.38 98.6	4.05 98.1	3.11 98.9	3.23 98.0	0.96 100	18.4 79.5	8.99 91.5	3.98 97.2	6.43 93.7	3.30 98.8	5.58 95.4 3.17
Ours	3.35 98.6	3.92 98.1	3.21 98.9	2.85 98.7	0.96 100	18.0 78.7	8.69 92.1	4.03 97.2	18.7 76.6	2.90 99.5	6.66 93.8 2.17

Table B.4: Evaluation results of baselines and our method on each dataset.

Ablation		NYUv2	KITTI	ETH3D	iBims-1	GSO	Sintel	DDAD	DIODE	Spring	HAMMER	Avg.
Data	Scale Prediction	Rel↓ $\delta_1$ ↑	Rel↓ $\delta_1$ ↑	Rel↓ $\delta_1$ ↑	Rel↓ $\delta_1$ ↑	Rel↓ $\delta_1$ ↑	Rel↓ $\delta_1$ ↑	Rel↓ $\delta_1$ ↑	Rel↓ $\delta_1$ ↑	Rel↓ $\delta_1$ ↑	Rel↓ $\delta_1$ ↑	Rel↓ $\delta_1$ ↑
Metric point map												
Improved real	Entangled (SI-Log)	6.00 97.3	8.33 93.4	11.6 89.4	7.78 94.6	- -	- -	14.4 83.1	10.6 88.4	- -	11.4 88.7	10.0 90.7
Improved real	Entangled (Shift inv.)	5.26 97.6	8.81 92.6	10.4 91.7	6.14 96.6	- -	- -	13.0 84.8	8.97 91.0	- -	10.4 90.3	9.00 92.1
Improved real	Decoupled (Conv)	5.37 97.8	9.56 91.8	9.46 94.1	6.49 95.6	- -	- -	13.3 83.5	8.93 91.9	- -	14.2 85.1	9.62 91.4
Synthetic only	Decoupled (MLP)	8.58 94.7	9.48 91.9	14.9 83.4	8.20 94.2	- -	- -	16.5 80.4	11.0 88.8	- -	18.4 78.2	12.4 87.4
Raw real	Decoupled (MLP)	5.36 97.8	7.70 94.5	8.58 94.6	6.60 95.7	- -	- -	12.2 85.5	9.01 91.5	- -	13.7 85.4	9.02 92.1
Improved real	Decoupled (MLP)	5.47 97.6	8.98 92.6	8.75 94.3	6.24 96.1	- -	- -	12.8 84.6	9.26 90.9	- -	12.9 87.4	9.20 91.9
Metric depth map (w/ GT intrinsics)												
Improved real	Entangled (SI-Log)	9.65 91.4	14.5 77.3	16.4 73.7	20.1 56.2	- -	- -	19.1 67.7	22.3 54.3	- -	23.0 59.8	17.9 68.6
Improved real	Entangled (Shift inv.)	9.04 93.1	19.1 56.8	15.5 76.8	15.1 72.1	- -	- -	18.0 67.9	19.9 59.8	- -	22.0 55.3	16.9 68.8
Improved real	Decoupled (Conv)	9.22 92.7	20.3 51.8	13.8 79.8	15.8 71.0	- -	- -	18.1 66.6	19.0 61.8	- -	27.5 54.9	17.7 68.4
Synthetic only	Decoupled (MLP)	18.1 73.7	15.8 71.0	24.7 53.2	15.8 76.6	- -	- -	21.9 62.4	22.7 56.5	- -	32.8 62.0	21.7 65.1
Raw real	Decoupled (MLP)	9.22 92.9	13.8 80.5	13.8 82.1	16.7 72.4	- -	- -	16.5 72.3	19.7 61.5	- -	20.8 67.9	15.8 75.7
Improved real	Decoupled (MLP)	9.48 92.2	18.7 59.2	13.5 82.6	13.6 79.3	- -	- -	17.0 69.2	20.0 59.7	- -	23.4 67.4	16.5 72.8
Scale-invariant point map												
Improved real	Entangled (SI-Log)	6.03 97.4	9.68 95.3	8.13 95.3	8.63 96.8	4.01 100	26.6 59.4	13.8 84.8	10.3 89.8	31.2 48.0	10.4 95.6	12.9 86.2
Improved real	Entangled (Shift inv.)	5.00 97.8	10.7 95.8	7.02 96.7	7.42 97.4	3.42 100	26.0 58.2	12.7 86.6	8.97 92.2	28.9 49.8	10.5 97.5	12.1 87.2
Improved real	Decoupled (Conv)	4.84 97.8	12.1 94.8	6.55 96.9	7.15 96.9	3.19 100	26.3 55.8	12.9 85.5	9.11 91.8	29.9 46.9	10.4 97.0	12.2 86.3
Synthetic only	Decoupled (MLP)	6.66 96.9	11.3 93.0	6.85 95.8	5.99 96.7	3.14 100	25.4 61.3	15.0 81.1	10.5 89.4	30.9 46.8	7.39 97.5	12.3 85.8
Raw real	Decoupled (MLP)	4.88 98.0	9.15 96.0	6.08 97.1	7.31 96.8	3.06 100	24.8 60.8	11.8 87.7	8.34 92.3	28.1 53.2	10.9 96.2	11.4 87.8
Improved real	Decoupled (MLP)	5.00 97.8	11.2 95.0	6.21 97.4	6.52 97.3	2.97 100	25.6 60.3	12.6 87.0	8.76 92.3	28.3 51.0	9.10 98.3	11.6 87.6
Affine-invariant point map												
Improved real	Entangled (SI-Log)	5.00 97.7	8.22 93.0	6.72 96.1	5.71 96.8	2.57 100	21.1 71.0	12.9 84.5	7.36 91.9	26.7 59.7	6.37 97.1	10.3 88.8
Improved real	Entangled (Shift inv.)	4.17 97.9	8.58 93.0	5.42 97.0	4.74 96.8	1.78 100	19.7 72.7	11.7 86.6	6.07 93.5	23.3 66.3	5.02 98.6	9.05 90.2
Improved real	Decoupled (Conv)	4.08 98.0	9.65 90.8	5.12 97.0	4.67 97.0	1.66 100	19.6 72.2	12.0 85.4	6.11 93.4	23.4 67.9	5.24 98.0	9.15 90.0
Synthetic only	Decoupled (MLP)	5.48 97.0	9.11 90.2	5.93 96.2	4.94 96.6	1.56 100	20.0 73.1	13.7 81.9	7.01 91.6	25.5 63.7	4.44 98.7	9.77 88.9
Raw real	Decoupled (MLP)	4.06 98.2	7.37 94.5	4.89 97.5	4.74 96.8	1.61 100	19.0 73.6	10.9 87.9	5.89 93.7	23.4 67.0	5.09 98.2	8.70 90.7
Improved real	Decoupled (MLP)	4.14 98.0	8.95 92.0	4.94 97.5	4.50 97.2	1.62 100	19.6 73.6	11.7 86.6	6.06 93.3	22.8 68.7	4.40 98.9	8.87 90.6
Local point map												
Improved real	Entangled (SI-Log)	- -	- -	6.30 95.6	5.96 96.6	- -	12.0 87.5	8.14 92.5	8.63 92.5	- -	- -	8.21 92.9
Improved real	Entangled (Shift inv.)	- -	- -	4.61 97.2	4.56 97.2	- -	10.3 90.2	7.38 93.5	6.61 94.7	- -	- -	6.69 94.6
Improved real	Decoupled (Conv)	- -	- -	4.25 97.5	4.34 97.3	- -	9.72 91.0	7.24 93.6	6.17 95.1	- -	- -	6.34 94.9
Synthetic only	Decoupled (MLP)	- -	- -	4.37 97.4	4.45 97.2	- -	9.33 91.7	7.51 93.3	6.44 94.9	- -	- -	6.42 94.9
Raw real	Decoupled (MLP)	- -	- -	4.28 97.4	4.55 97.1	- -	9.64 91.2	7.11 93.7	6.28 94.9	- -	- -	6.37 94.9
Improved real	Decoupled (MLP)	- -	- -	4.20 97.5	4.31 97.3	- -	9.34 91.9	7.21 93.7	6.21 95.0	- -	- -	6.25 95.1
Scale-invariant depth map												
Improved real	Entangled (SI-Log)	4.99 97.4	5.18 96.7	6.48 95.2	5.26 97.3	2.64 100	23.2 65.7	11.6 86.2	7.76 91.3	24.8 63.4	6.40 96.7	9.83 89.0
Improved real	Entangled (Shift inv.)	4.17 97.8	4.57 97.5	5.03 96.7	4.42 97.6	2.09 100	22.4 66.7	10.3 88.0	6.16 93.5	20.8 69.2	4.69 98.6	8.46 90.6
Improved real	Decoupled (Conv)	4.08 97.9	4.61 97.3	4.80 96.9	4.32 97.1	1.92 100	22.5 64.9	10.3 87.7	6.26 93.2	21.0 68.8	4.81 98.2	8.46 90.2
Synthetic only	Decoupled (MLP)	5.05 96.9	5.47 96.3	5.64 95.9	4.76 96.9	1.90 100	21.7 68.0	12.0 85.1	7.16 91.2	22.2 67.4	4.66 97.9	9.05 89.6
Raw real	Decoupled (MLP)	4.09 98.0	4.60 97.2	4.82 97.1	4.49 97.0	1.92 100	21.8 66.5	9.79 88.4	6.09 93.2	21.8 68.6	4.67 98.1	8.41 90.4
Improved real	Decoupled (MLP)	4.16 97.8	4.59 97.2	4.62 97.4	4.20 97.4	1.89 100	21.9 67.8	10.1 88.4	6.08 93.3	20.4 71.6	4.34 98.9	8.23 91.0
Affine-invariant depth												
Improved real	Entangled (SI-Log)	4.32 98.0	4.91 97.0	5.21 96.9	4.15 97.8	2.11 100	17.4 76.3	10.8 87.9	5.17 95.7	20.3 72.6	5.26 97.6	7.96 92.0
Improved real	Entangled (Shift inv.)	3.57 98.2	4.26 97.6	3.98 97.9	3.35 98.2	1.44 100	16.0 78.4	9.58 89.7	4.01 96.8	17.7 76.7	3.58 98.7	6.75 93.2
Improved real	Decoupled (Conv)	3.48 98.3	4.31 97.4	3.77 98.1	3.23 98.0	1.35 100	15.2 79.2	9.48 89.5	3.99 96.4	17.7 76.4	3.74 98.5	6.62 93.2
Synthetic only	Decoupled (MLP)	4.19 97.8	5.04 96.6	4.40 97.6	3.73 97.7	1.27 100	15.6 78.9	11.1 86.8	4.50 95.7	19.1 75.0	3.58 98.8	7.25 92.5
Raw real	Decoupled (MLP)	3.48 98.4	4.29 97.3	3.75 98.3	3.37 98.0	1.32 100	15.2 80.1	8.95 90.2	3.91 96.7	18.6 75.7	3.48 98.4	6.63 93.3
Improved real	Decoupled (MLP)	3.53 98.3	4.30 97.3	3.69 98.3	3.16 98.3	1.33 100	15.3 79.3	9.28 90.1	3.95 96.6	17.6 76.8	3.19 99.0	6.53 93.4
Affine-invariant disparity												
Improved real	Entangled (SI-Log)	4.69 98.1	4.99 97.2	5.73 96.7	4.63 97.8	2.14 100	21.9 71.1	11.3 88.0	5.69 95.8	23.8 68.4	5.46 98.3	9.03 91.1
Improved real	Entangled (Shift inv.)	4.03 98.3	4.39 97.7	4.48 97.8	3.85 98.2	1.47 100	20.2 73.8	9.84 89.8	4.75 96.5	21.3 69.5	3.74 99.0	7.80 92.1
Improved real	Decoupled (Conv)	3.96 98.4	4.49 97.5	4.29 97.9	3.75 98.4	1.39 100	19.9 74.4	9.89 89.7	4.75 96.3	21.1 70.0	3.81 98.8	7.73 92.1
Synthetic only	Decoupled (MLP)	4.75 97.8	5.17 96.7	4.86 97.6	4.29 98.0	1.31 100	20.8 74.0	11.5 87.3	5.21 95.9	22.2 69.7	3.66 99.0	8.38 91.6
Raw real	Decoupled (MLP)	3.92 98.4	4.50 97.4	4.23 98.0	3.95 98.2	1.35 100	19.9 74.8	9.22 90.6	4.78 96.3	21.4 69.6	3.61 98.7	7.69 92.2
Improved real	Decoupled (MLP)	4.03 98.3	4.45 97.5	4.11 98.1	3.74 98.2	1.37 100	19.8 75.4	9.71 90.1	4.79 96.2	20.0 72.5	3.30 99.3	7.53 92.6

Table B.5: Evaluation results of ablation study on each sets

## References

- [1] Baidu Apollo. Apollo synthetic dataset, 2019. Accessed: 2025-03-06.
- [2] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.
- [3] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.
- [4] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv*, 2024.
- [5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.
- [6] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023.
- [7] Michael Fonder and Marc Van Droogenbroeck. Mid-air: A multi-modal dataset for extremely low altitude drone flights. In *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2019.
- [8] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. A2D2: Audi Autonomous Driving Dataset. 2020.
- [9] Jose L. Gómez, Manuel Silva, Antonio Seoane, Agnès Borrás, Mario Noriega, Germán Ros, Jose A. Iglesias-Guitián, and Antonio M. López. All for one, and one for all: Urbansyn dataset, the third musketeer of synthetic driving scenes, 2023.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [11] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024.
- [12] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [13] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.
- [14] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.
- [15] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3205–3215, 2023.
- [16] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [17] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Transactions on Graphics*, 38(6):184:1–184:15, 2019.
- [18] 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. *arXiv preprint arXiv:2304.07193*, 2023.

- [19] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [20] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025.
- [21] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.
- [22] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.
- [23] 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 *International Conference on Computer Vision (ICCV) 2021*, 2021.
- [24] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [25] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [26] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [27] Kaixuan Wang and Shaojie Shen. Flow-motion and depth network for monocular stereo and beyond. *CoRR*, abs/1909.05452, 2019.
- [28] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. IRS: A large synthetic indoor robotics stereo dataset for disparity and surface normal estimation. *CoRR*, abs/1912.09678, 2019.
- [29] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. 2024.
- [30] 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. 2020.
- [31] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021.
- [32] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *CoRR*, abs/1810.08705, 2018.
- [33] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.
- [34] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024.
- [35] 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. *Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [36] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Yifan Liu, and Chunhua Shen. Towards accurate reconstruction of 3d scene shape from a single monocular image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6480–6494, 2022.

- 196 [37] Amir R Zamir, Alexander Sax, , William B Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese.  
197 Taskonomy: Disentangling task transfer learning. In *2018 IEEE Conference on Computer Vision and*  
198 *Pattern Recognition (CVPR)*. IEEE, 2018.
- 199 [38] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-  
200 realistic dataset for structured 3d modeling. In *Proceedings of The European Conference on Computer*  
201 *Vision (ECCV)*, 2020.