



MoGe: Unlocking Accurate Monocular Geometry Estimation for Open-Domain Images with Optimal Training Supervision

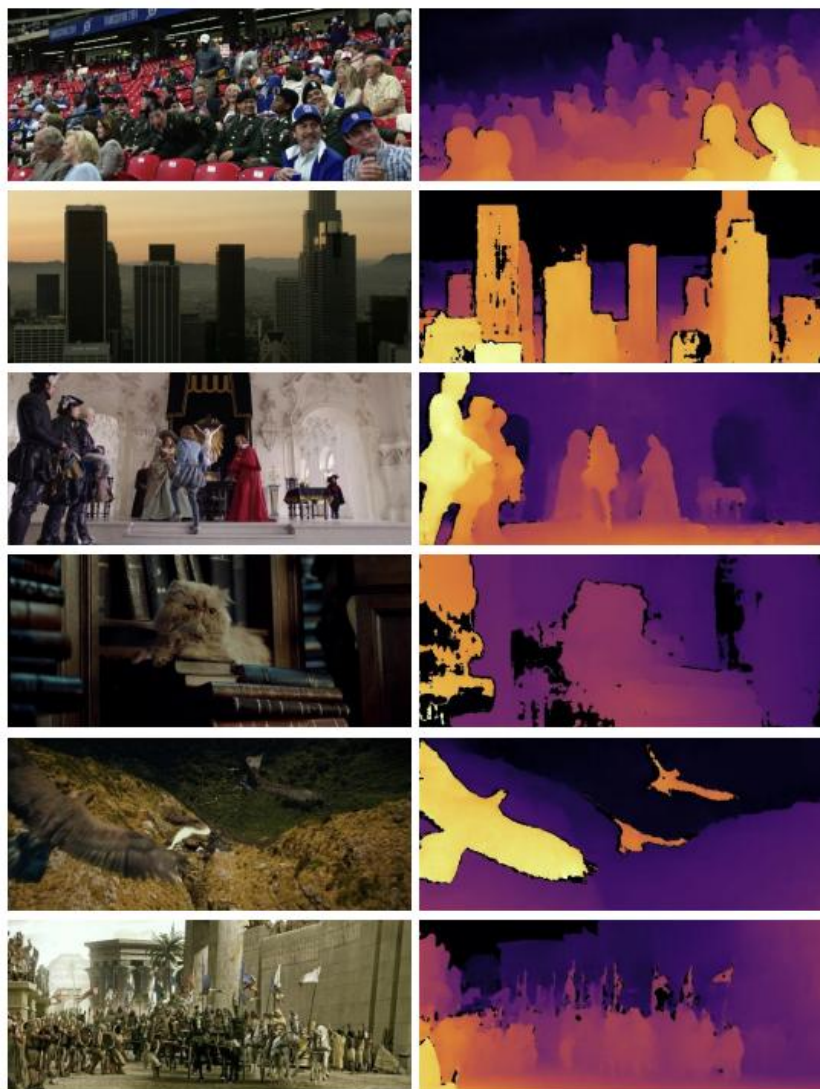
CVPR 2025 Oral

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, Jiaolong Yang

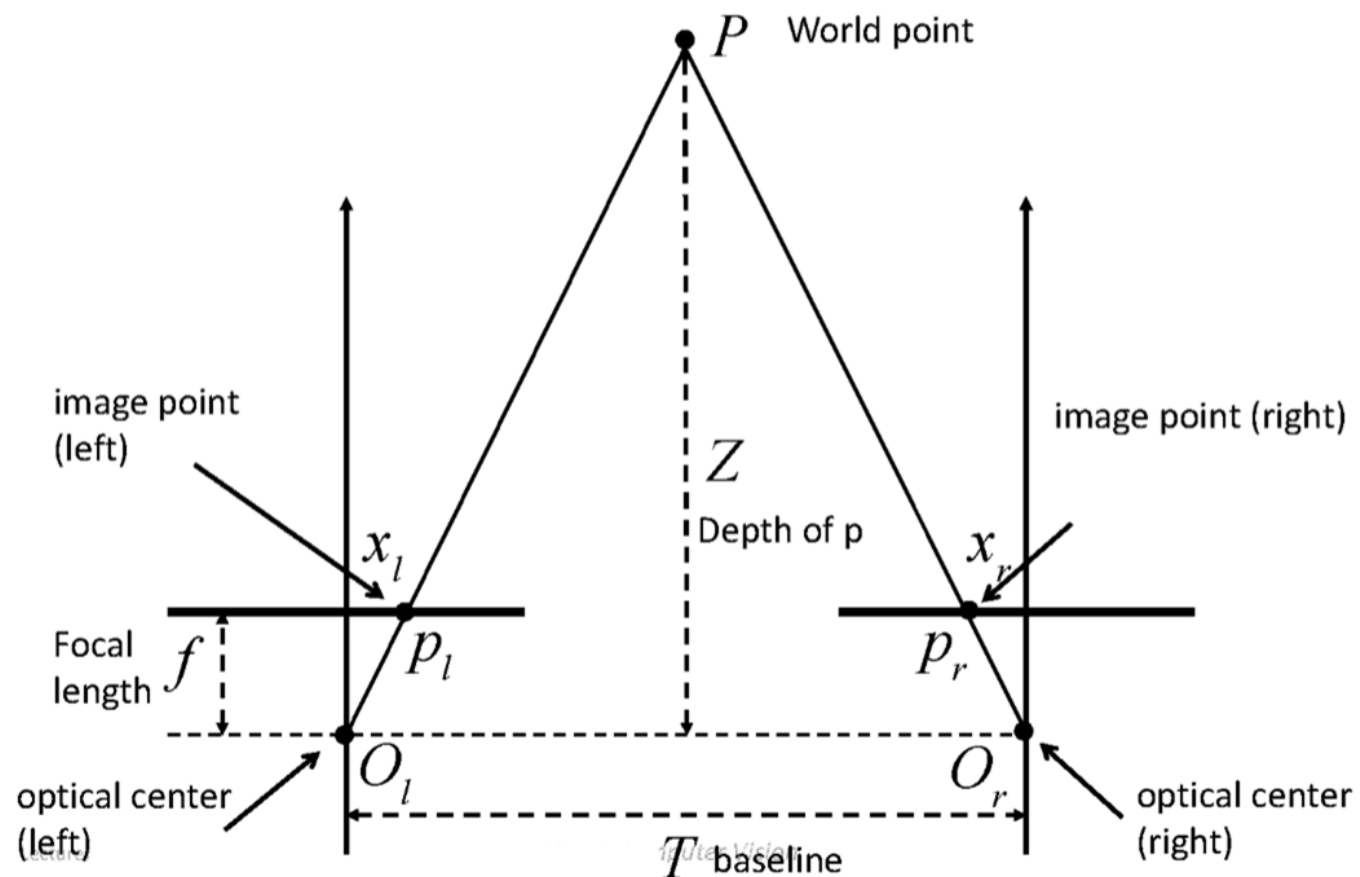
USTC Microsoft Research Harvard Tsinghua University

2025.10.01
Presented by Gwanyong Kim

MDE(Monocular Depth Estimation)



SDE(Stereo Depth Estimation)



Jin-Hwi Park, Understanding and Application of 3D Depth Estimation, Lecture 3 'Stereo depth Estimation'

MRDE(Monocular Relative Depth Estimation)

MMDE(Monocular Metric Depth Estimation)

= MRDE + Metric Scale

$$\mathbf{D}^m = \frac{b \cdot f}{\mathbf{d}^m}, \mathbf{D}^{\text{rel}} = \text{rel}(\mathbf{D}^m) = \frac{\mathbf{D}^m - t(\mathbf{D}^m)}{s(\mathbf{D}^m)},$$

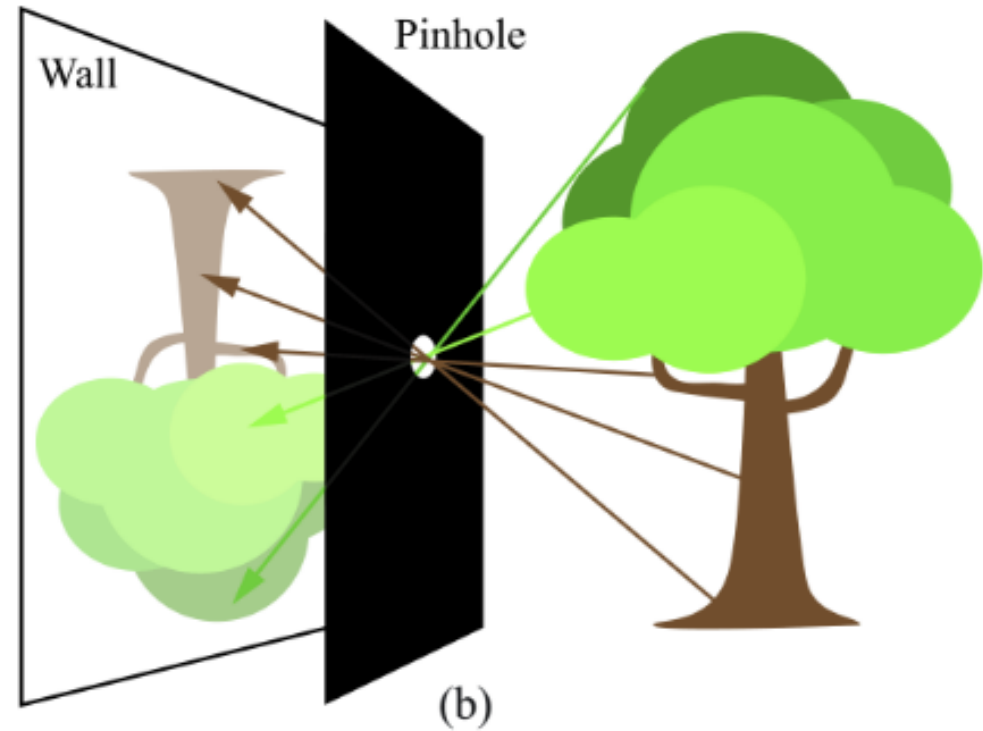
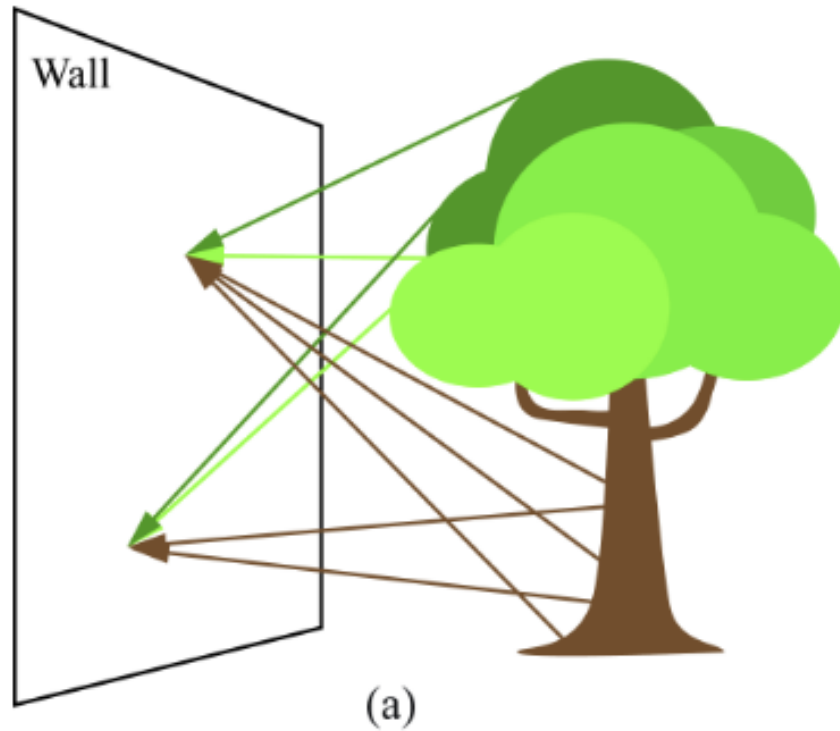
\mathbf{D}^m : Metric Depth
 \mathbf{d}^m : Disparity
 $\text{rel}()$: Relative Depth

$t()$: shift
 $s()$: scale

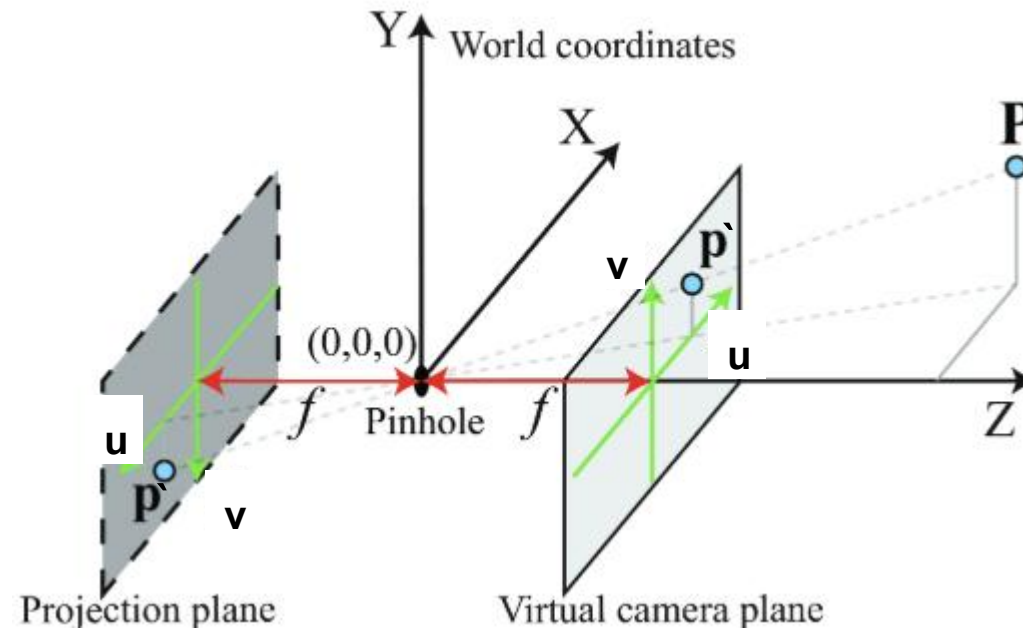
MGE(Monocular Geometry Estimation)



MoGe: Unlocking Accurate Monocular Geometry Estimation for Open-Domain Images with Optimal Training Supervision



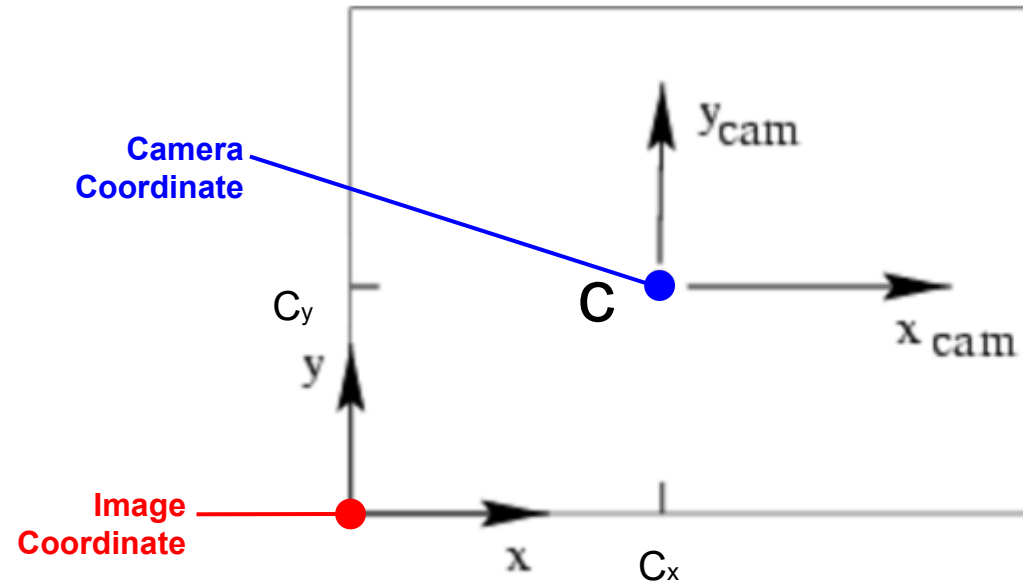
<https://visionbook.mit.edu/imaging.html>



<https://visionbook.mit.edu/imaging.html>

$$u : X = f : Z, \quad v : Y = f : Z$$

$$P(X, Y, Z) \rightarrow p'(u, v) = (f \cdot X / Z, f \cdot Y / Z)$$



principal point: (C_u, C_v)

$$p'(f^*X/Z, f^*Y/Z) \rightarrow p'(f^*X/Z + C_u, f^*Y/Z + C_v)$$

$$\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \Rightarrow \begin{pmatrix} f^*X/Z + C_u \\ f^*Y/Z + C_v \\ 1 \end{pmatrix} \Rightarrow \underbrace{\begin{pmatrix} f_x & C_u \\ & f_y & C_v \\ & & 1 \end{pmatrix}}_{\text{Intrinsics}} \underbrace{\begin{pmatrix} 1 & & 0 \\ & 1 & 0 \\ & & 1 & 0 \end{pmatrix}}_{\text{Projection}} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}$$

Monocular Depth Estimation

MMDE

- 여러 방식에서 특정 센서 데이터에 의존하거나 카메라 파라미터에 의존하는 경우가 있음.

MRDE

- MMDE보다 훨씬 더 다양한 데이터를 활용할 수 있는 능력이 있어 더 향상된 일반화 능력을 가짐.

Monocular point map estimation

LeRes

- affine invariant depth map 예측을 먼저 수행한 뒤 포인트 클라우드 모듈을 통해 shift와 focal length를 복원하는 2단계 파이프라인 (전역 Metric은 모호함)

UniDepth

- 깊이 대신 픽셀별 3D 포인트를 직접 내보내는 구조.
- 카메라 Self-Prompt: 네트워크가 카메라 파라미터를 추정해 Depth feature에 조건으로 주입
- 추가 정보 없이 단일 이미지만으로 Metric 3D point map 생성.

DUS3R

- End to End 모델을 사용하여 두 장의 이미지를 카메라 공간 point map으로 직접 매핑
- 동일한 입력 이미지(단일 이미지)는 같은 이미지를 두 번 사용하는 방식으로 point map 생성
- Scale invariant point map은 초점 거리 모호성을 해결하지 못함

->Depth와 함께 카메라 파라미터를 예측

Camera Intinsics Estimation

- 초기 연구에서는 3차원 형상, 소실점을 활용
- 최근 학습 기반의 예측 연구도 진행되었지만, 만족스러운 결과가 나오진 않음

Large-scale data training for monocular geometry

- **MiDaS** : 다양한 도메인에서 가져온 대규모 데이터셋을 혼합하여 학습
- **Depth Anything** : 정밀한 Synthetic Data + Pseudo Labeled Data

-> 여러 데이터셋을 혼합해서 일반화 능력을 향상시킴

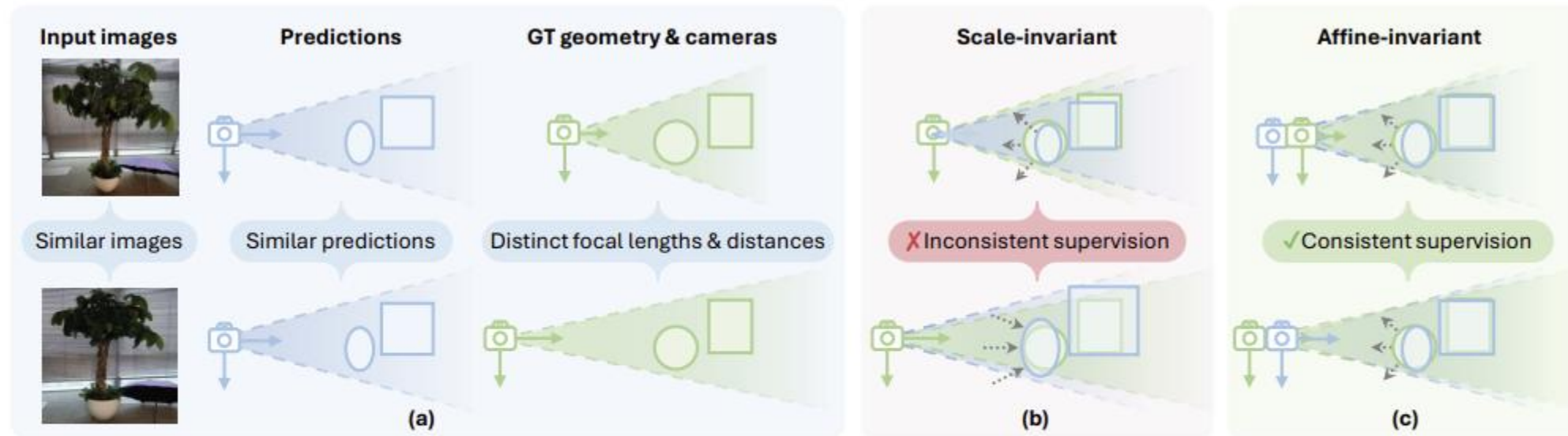
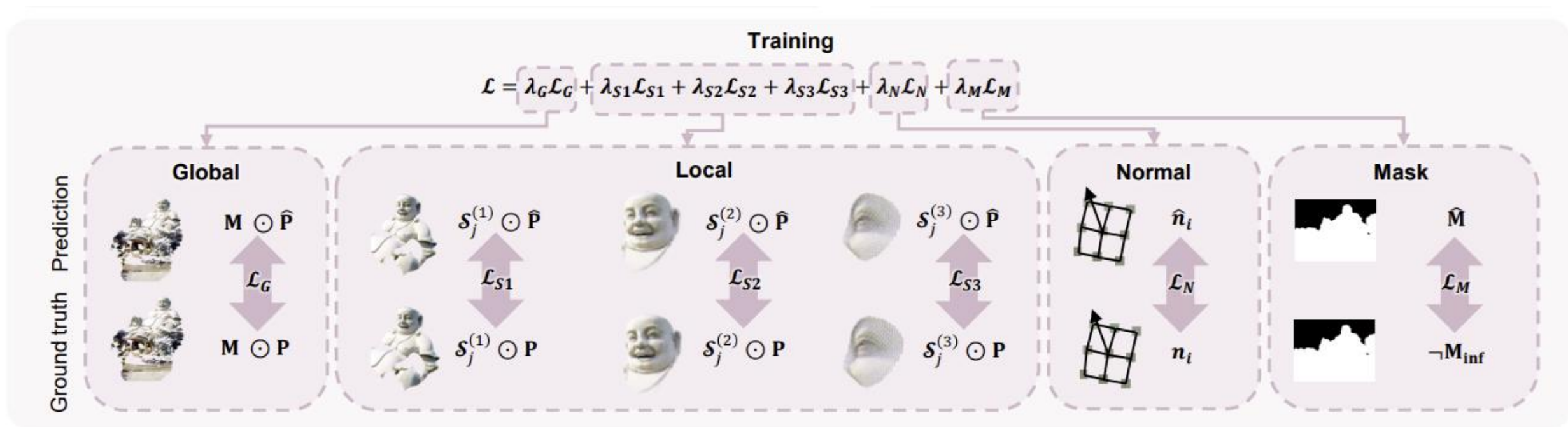
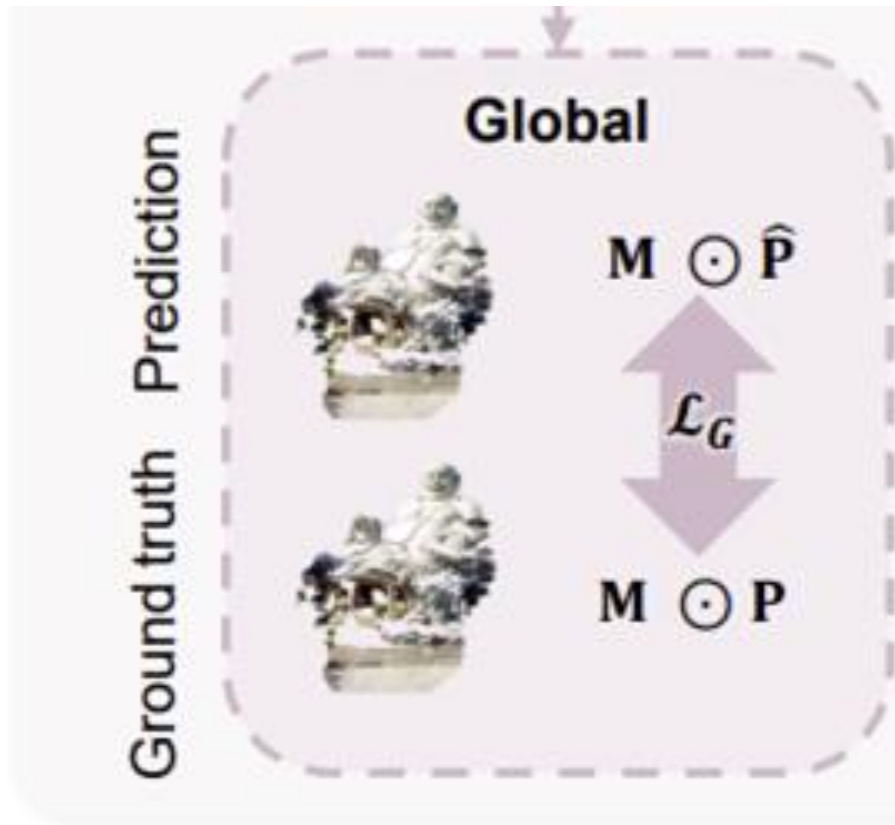


Figure 4. The focal-distance ambiguity and effects of different 3D point representations. (a) For similar images captured with varying camera focal length and distance to the objects, perceiving their true camera setup is challenging and models often produce similar geometries. (b) Inconsistent supervision signals occur with only scale alignment. (c) Consistent geometry supervision with an additional translation alignment.





ROE(Robust, Optimal, Efficient) alignment solver

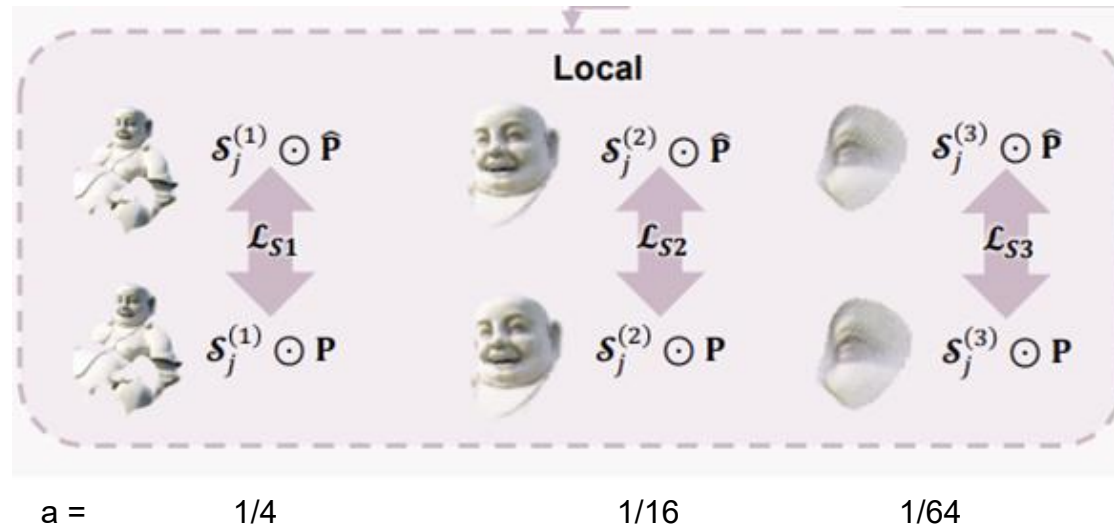
$$(s^*, t^*) = \operatorname{argmin}_{s, t} \sum_{i \in \mathcal{M}} \frac{1}{z_i} \|s \hat{\mathbf{p}}_i + \mathbf{t} - \mathbf{p}_i\|_1$$

s^* : alignment scale factor	$\frac{1}{z_i}$: weighting term(z_i : z-coordi of \mathbf{p}_i)
t^* : alignment shift factor	\mathcal{M}	: valid mask
s : scale factor	$\hat{\mathbf{p}}_i$: predicted 3D point(i th pixel)
t : shift factor	\mathbf{p}_i	: corresponding ground truth(i th pixel)

Global point map loss

$$\mathcal{L}_G = \sum_{i \in \mathcal{M}} \frac{1}{z_i} \|s \hat{\mathbf{p}}_i + \mathbf{t} - \mathbf{p}_i\|_1,$$

s : alignment scale factor
 t : alignment shift factor



Multi-scale geometry loss

$$\mathcal{S}_j = \{i \mid \|\mathbf{p}_i - \mathbf{p}_j\| \leq r_j, i \in \mathcal{M}\}$$

$$r_j = \alpha \cdot z_j \cdot \frac{\sqrt{W^2 + H^2}}{2 \cdot f}$$

$$\mathcal{L}_{S(\alpha)} = \sum_{j \in \mathcal{H}_\alpha} l_{\mathcal{S}_j} = \sum_{j \in \mathcal{H}_\alpha} \sum_{i \in \mathcal{S}_j} \frac{1}{z_i} \|s_j^* \hat{\mathbf{p}}_i + \mathbf{t}_j^* - \mathbf{p}_i\|_1$$

\mathbf{P}_j : ground truth 3D point map

\mathbf{P}_i : 3D point map(i th pixel)

r_j : radius

W, H: image width, height

\mathcal{M} : valid Mask

f: focal length

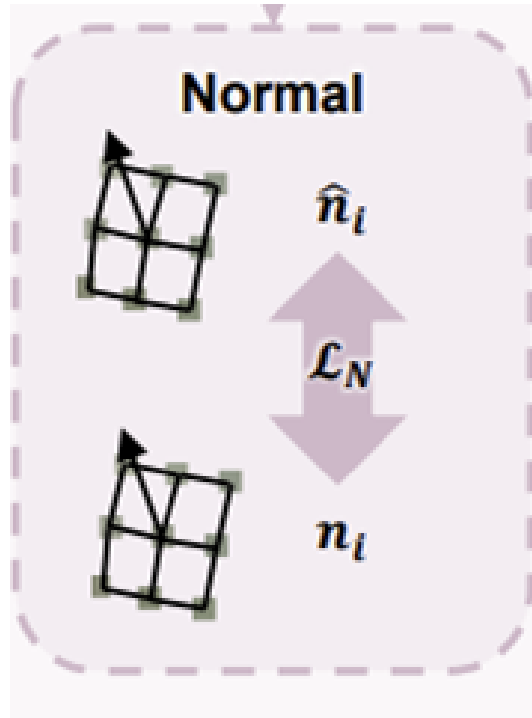
z_j : z-coordinate of pj

α : sphere scale hyper-parameter

s^*, t^* : scale, shift optimal factor(ROE solver)

$\hat{\mathbf{p}}_i$: predicted 3D point(i th pixel)

\mathbf{P}_i : corresponding ground truth(i th pixel)



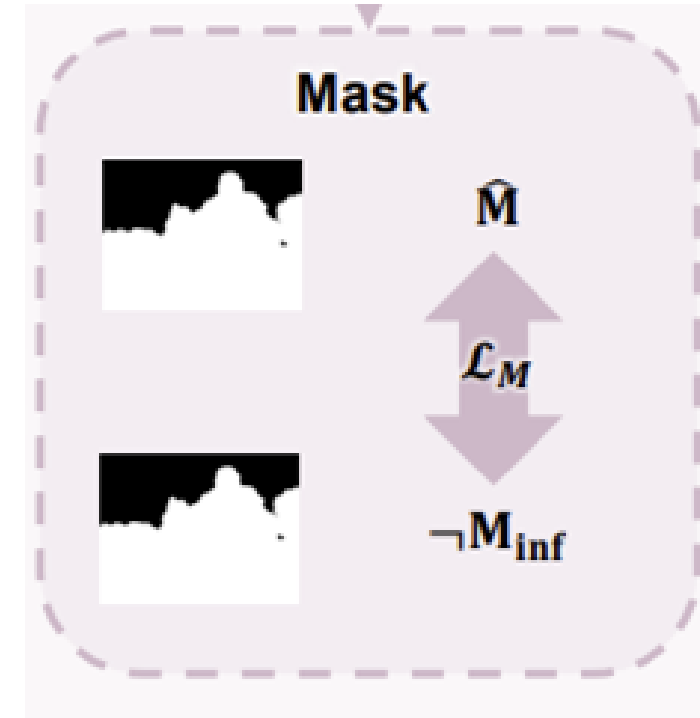
Normal loss

$$\mathcal{L}_N = \sum_{i \in \mathcal{M}} \angle(\hat{n}_i, \mathbf{n}_i)$$

\hat{n}_i : predicted normal vector(i th pixel)

\mathbf{n}_i : normal vector GT(i th pixel)

$\angle(\cdot, \cdot)$: the angle difference between the two vectors



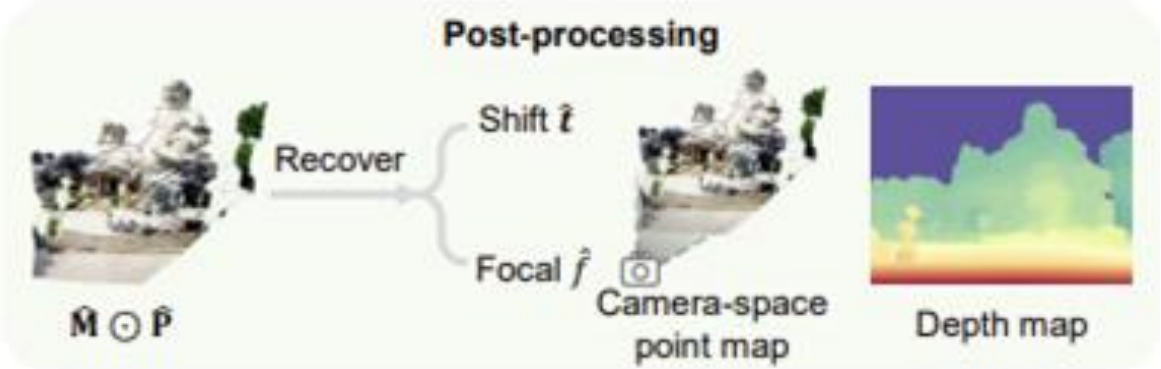
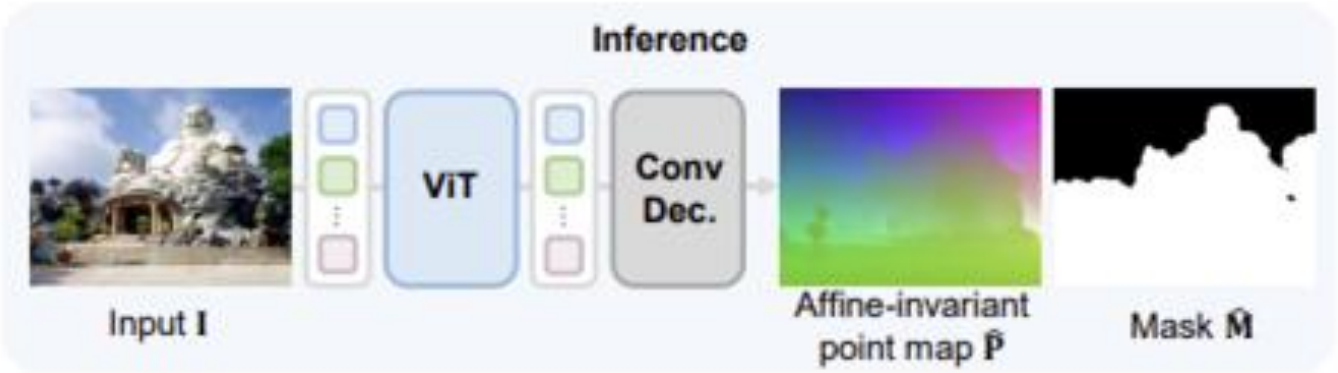
Mask loss

$$\mathcal{L}_M = \|\hat{M} - (1 - M_{\text{inf}})\|_2^2$$

M_{inf} : Infinity(invalid) mask label(ex. sky, background..)

$(1 - M_{\text{inf}})$: valid mask

\hat{M} : predicted valid mask



Recovery camera focal and shift

$$\min_{f, t'_z} \sum_{i=1}^N \left(\frac{f x_i}{z_i + t'_z} - u_i \right)^2 + \left(\frac{f y_i}{z_i + t'_z} - v_i \right)^2$$

(u_i, v_i) : 2D Image pixel coordinates f : Predicted focal length

(x_i, y_i, z_i) : 3D Unprojection affine invariant point coordinates t'_z : Z-axis shift

Method	NYUv2		KITTI		ETH3D		iBims-1		GSO		Sintel		DDAD		DIODE		Average		
	Rel ^P ↓	δ_1^P ↑	Rel ^P ↓	δ_1^P ↑	Rel ^P ↓	δ_1^P ↑	Rel ^P ↓	δ_1^P ↑	Rel ^P ↓	δ_1^P ↑	Rel ^P ↓	δ_1^P ↑	Rel ^P ↓	δ_1^P ↑	Rel ^P ↓	δ_1^P ↑	Rel ^P ↓	δ_1^P ↑	Rank↓
Scale-invariant point map																			
LeReS	16.9	76.0	31.6	28.4	17.1	75.8	18.5	72.2	14.7	76.0	38.6	30.6	32.0	39.4	27.6	46.4	24.6	55.6	3.94
DUST3R	5.53	97.1	15.2	87.9	<u>10.7</u>	<u>90.6</u>	6.18	95.4	<u>4.54</u>	99.3	34.8	<u>50.3</u>	21.4	70.1	12.4	86.7	13.8	84.7	2.75
UniDepth	<u>5.33</u>	98.4	<u>5.96</u>	98.5	18.5	77.6	<u>5.29</u>	97.4	6.58	<u>99.6</u>	<u>33.0</u>	48.9	11.4	<u>90.2</u>	<u>12.3</u>	<u>91.0</u>	<u>12.3</u>	<u>87.7</u>	<u>2.09</u>
Ours	4.86	98.4	5.47	<u>97.4</u>	4.58	98.9	4.63	<u>97.1</u>	2.58	100	22.3	69.5	<u>12.3</u>	90.3	6.58	94.5	7.91	93.3	1.22
Affine-invariant point map																			
LeReS	9.51	91.4	26.1	49.1	14.7	79.6	11.0	88.6	8.91	95.2	29.7	55.5	29.4	46.7	15.1	80.1	18.1	73.3	3.94
DUST3R	4.45	97.4	12.7	83.3	<u>7.27</u>	<u>95.0</u>	5.04	96.0	3.07	99.6	30.3	56.6	19.7	71.2	8.97	88.7	11.4	86.0	2.94
UniDepth	<u>3.93</u>	98.4	4.29	98.6	12.2	89.6	<u>4.65</u>	98.0	<u>2.99</u>	<u>99.8</u>	<u>28.5</u>	<u>58.4</u>	10.3	<u>90.5</u>	<u>8.56</u>	<u>90.9</u>	<u>9.43</u>	<u>90.5</u>	<u>1.81</u>
Ours	3.68	<u>98.3</u>	<u>4.86</u>	<u>97.2</u>	3.57	99.0	3.61	<u>97.3</u>	1.14	100	16.8	77.8	<u>10.5</u>	91.4	<u>4.37</u>	96.4	6.07	94.7	1.31
Local point map																			
LeReS	-	-	-	-	9.32	91.9	8.57	93.2	-	-	13.3	84.8	10.7	88.9	11.6	88.2	10.7	89.4	3.80
DUST3R	-	-	-	-	<u>6.05</u>	<u>94.8</u>	<u>5.44</u>	95.9	-	-	<u>11.8</u>	<u>87.0</u>	9.24	90.8	<u>7.32</u>	<u>93.1</u>	<u>7.97</u>	<u>92.3</u>	<u>2.30</u>
UniDepth	-	-	-	-	8.61	92.6	5.92	<u>96.0</u>	-	-	13.4	84.3	<u>8.18</u>	<u>92.0</u>	9.95	90.0	9.21	91.0	2.90
Ours	-	-	-	-	3.21	98.1	4.16	96.8	-	-	8.63	92.7	6.74	94.3	4.78	96.3	5.50	95.6	1.00

Table 1. Quantitative results for point map estimation. Rel^P and δ_1^P are in percentage. The best values are highlighted in **bold**, and the second-best ones are underlined. Local point map accuracy is evaluated on affine-invariant point maps within local object regions

Method	NYUv2		KITTI		ETH3D		iBims-1		GSO		Sintel		DDAD		DIODE		Average		
	Rel ^d ↓	δ_1^d ↑	Rel.↓	δ_1^d ↑	Rel ^d ↓	δ_1^d ↑	Rel.↓	δ_1^d ↑	Rel ^d ↓	δ_1^d ↑	Rel ^d ↓	δ_1^d ↑	Rel ^d ↓	δ_1^d ↑	Rel ^d ↓	δ_1^d ↑	Rel ^d ↓	δ_1^d ↑	Rank↓
Scale-invariant depth map																			
LeReS	12.1	82.6	19.2	64.8	14.2	78.4	14.0	78.8	13.6	77.9	30.5	52.1	26.5	52.0	18.2	69.6	18.5	69.5	7.31
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	11.2	86.1	5.50
DUSt3R	4.40	97.1	7.81	90.6	6.04	95.7	4.98	95.8	3.27	99.5	31.1	57.2	18.6	73.3	8.91	88.8	10.6	87.2	5.00
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	6.33	94.3	2.07
UniDepth	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	8.43	91.6	3.00
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	10.3	87.9	5.67
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	9.09	89.6	4.06
Ours	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	6.17	93.8	1.62
Affine-invariant depth map																			
Marigold	4.63	97.3	7.29	93.8	6.08	96.3	4.35	97.2	2.78	99.9	21.2	75.0	14.6	80.5	6.34	94.3	8.41	91.8	2.25
GeoWizard	4.69	97.4	8.14	92.5	6.90	94.0	4.50	97.1	2.00	99.9	17.8	76.2	16.5	75.7	7.03	92.7	8.44	90.7	2.69
Ours	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	4.72	95.8	1.00
Affine-invariant disparity map																			
MiDaS V3.1	4.58	98.1	6.25	94.7	5.77	96.8	4.73	97.4	1.86	100	21.3	73.1	14.5	82.6	6.05	94.9	8.13	92.2	3.69
DA V1	4.20	98.4	5.40	97.0	4.68	98.2	4.18	97.6	1.54	100	20.1	77.6	12.7	86.9	5.69	95.7	7.31	93.9	2.31
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	7.37	93.3	2.56
Ours	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	5.76	95.2	1.06

Table 2. Quantitative results for depth map estimation. Gray numbers denote models trained on respective benchmarks and thus excluded from ranking. A more extensive comparison can be found in the *suppl. materials*.

Method	NYUv2		ETH3D		iBims-1		Average		
	Mean↓	Med.↓	Mean	Med.↓	Mean↓	Med.↓	Mean↓	Med.↓	Rank↓
Perspective	5.38	4.39	13.6	11.9	10.6	9.30	9.86	8.53	5.00
WildCam	3.82	<u>3.20</u>	7.70	5.81	9.48	9.08	7.00	6.03	3.00
LeReS	19.4	19.6	8.26	7.19	18.4	17.5	15.4	14.8	5.53
DUSt3R	2.57	1.86	<u>5.77</u>	<u>3.60</u>	<u>3.83</u>	<u>2.53</u>	<u>4.06</u>	<u>2.66</u>	<u>1.67</u>
UniDepth	7.56	4.31	10.7	9.96	11.9	5.96	10.1	6.74	4.50
Ours	<u>3.41</u>	<u>3.21</u>	2.50	1.54	2.81	1.89	2.91	2.21	1.50

Table 3. Evaluation results for camera FOV in degrees.

Ablation	Point						Depth				Disparity	
	Scale-inv.		Affine-inv.		Local		Scale-inv.		Affine-inv.		Affine-inv.	
	RelP \downarrow	$\delta_1^P\uparrow$	RelP \downarrow	$\delta_1^P\uparrow$	RelP \downarrow	$\delta_1^P\uparrow$	Rel $^d\downarrow$	$\delta_1^d\uparrow$	Rel $^d\downarrow$	$\delta_1^d\uparrow$	Rel $^d\downarrow$	$\delta_1^d\uparrow$
SI-Log depth	11.2	88.7	9.09	90.6	9.19	91.2	8.94	90.1	7.27	92.6	8.23	92.1
Affine-inv. depth	29.9	51.4	29.0	52.7	12.2	86.0	28.9	52.7	6.18	93.9	15.9	76.6
ROE scale-inv.	10.3	89.8	8.34	91.6	8.59	91.9	8.27	90.9	6.73	93.2	7.90	92.6
L2 affine-inv.	13.5	84.2	10.3	88.2	9.48	91.0	11.1	85.7	8.03	91.2	9.37	90.5
Med. affine-inv.	10.9	89.0	8.97	90.7	9.44	90.7	9.10	89.8	7.50	92.4	8.74	91.8
ROE affine-inv.	9.84	90.3	7.88	92.1	7.62	93.3	7.91	91.2	6.29	93.7	7.43	93.2
Full w/o trunc.	9.81	90.5	7.91	91.7	7.12	93.8	7.92	91.3	6.31	93.5	7.45	93.1
Full w/o \mathcal{L}_S	9.98	90.3	7.94	92.1	7.47	93.4	7.94	91.2	6.30	93.6	7.47	93.2
Full	9.78	90.6	7.83	92.1	7.16	93.8	7.82	91.3	6.20	93.7	7.30	93.3

Table 4. Quantitative ablation study results. All experiments are conducted with a ViT-Base encoder. The first six rows are trained with global loss only.

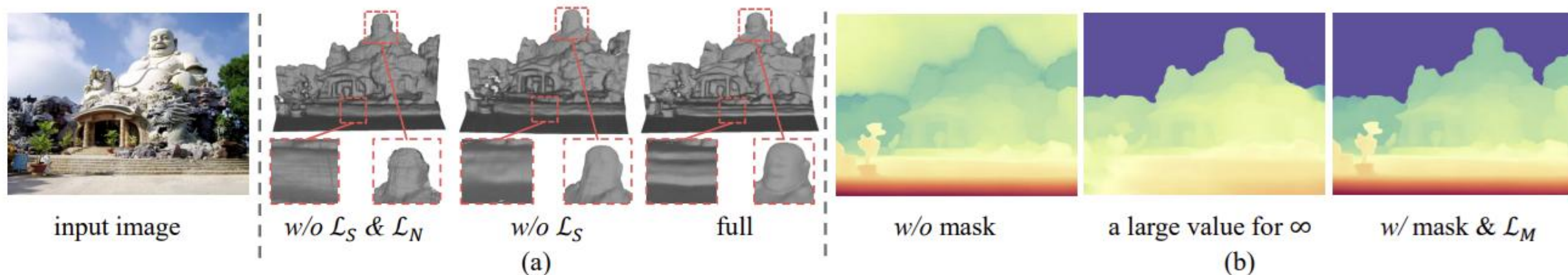
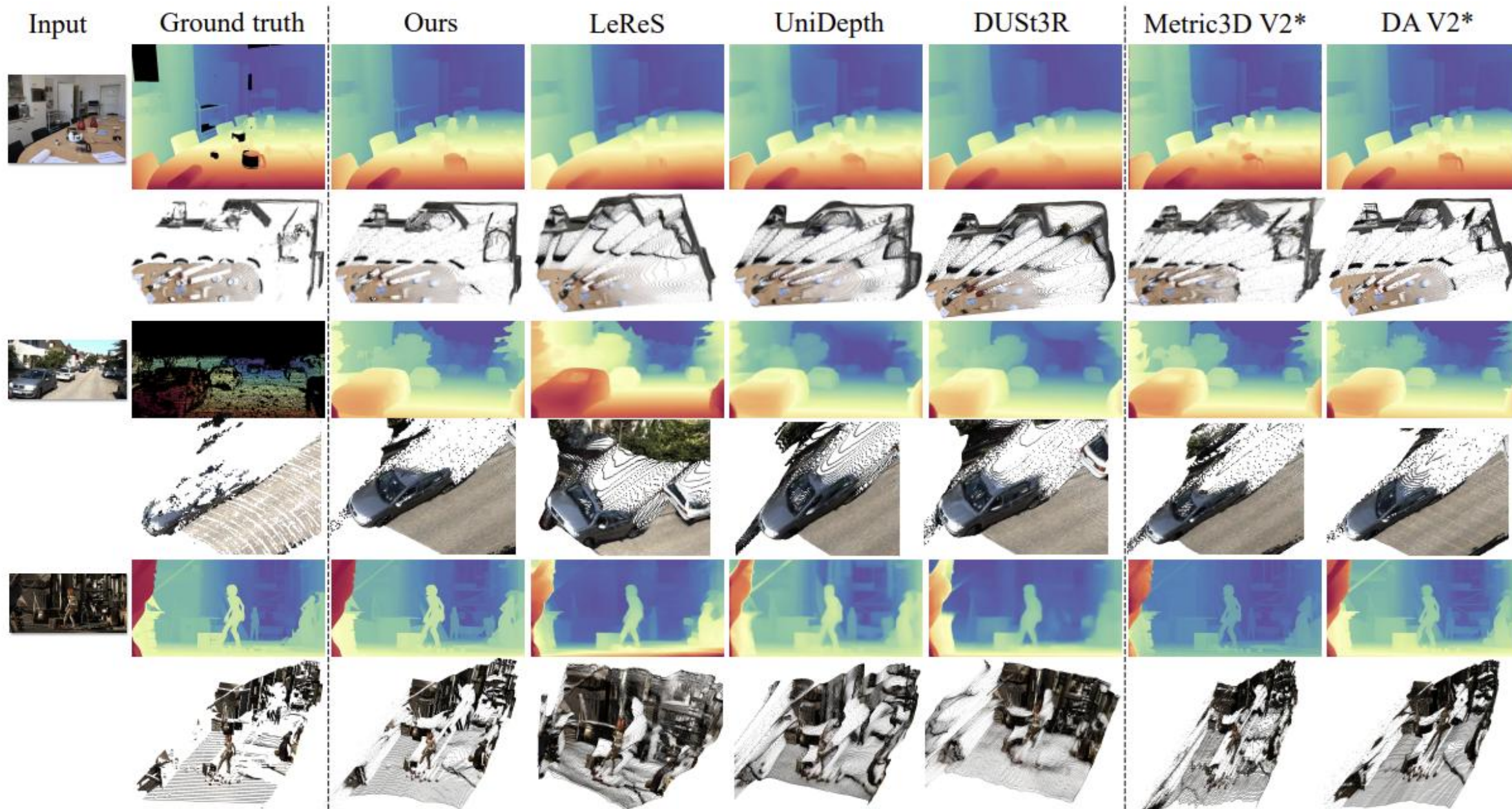




























Figure 6. Qualitative ablation study results. All experiments use the ViT-Base backbone for the encoder. **(a)** Surface visualization of the impact of \mathcal{L}_S and \mathcal{L}_N . Removing either leads to noisy surfaces and poor geometry. **(b)** Depth visualization for the ablation of valid region mask prediction. Our method correctly predicts the sky regions, for which the predicted point values will be erroneous if the masks are removed. Supervising infinity regions by assigning a large distance label can negatively affect the foreground prediction accuracy.



Input	Ground truth	Ours	LeReS	UniDepth	DUST3R	Metric3D V2*	DA V2*
	Not available						
	Not available						
	Not available						
	Not available						

Affine invariant point map

ROE alignment solver

Multi-Scale Geometry loss

Rich implicit knowledge in LLM



Metric scale priors