RSA: Resolving Scale Ambiguities in Monocular Depth Estimators through Language Descriptions

Ziyao Zeng[1] Yangchao Wu[2] Hyoungseob Park[1] Daniel Wang[1] Fengyu Yang[1]
Stefano Soatto[2] Dong Lao[2] Byung-Woo Hong[3] Alex Wong[1]

[1]Yale University [2]University of California, Los Angeles [3]Chung-Ang University

Motivation

- Different language descriptions can specify different scales of scenes.

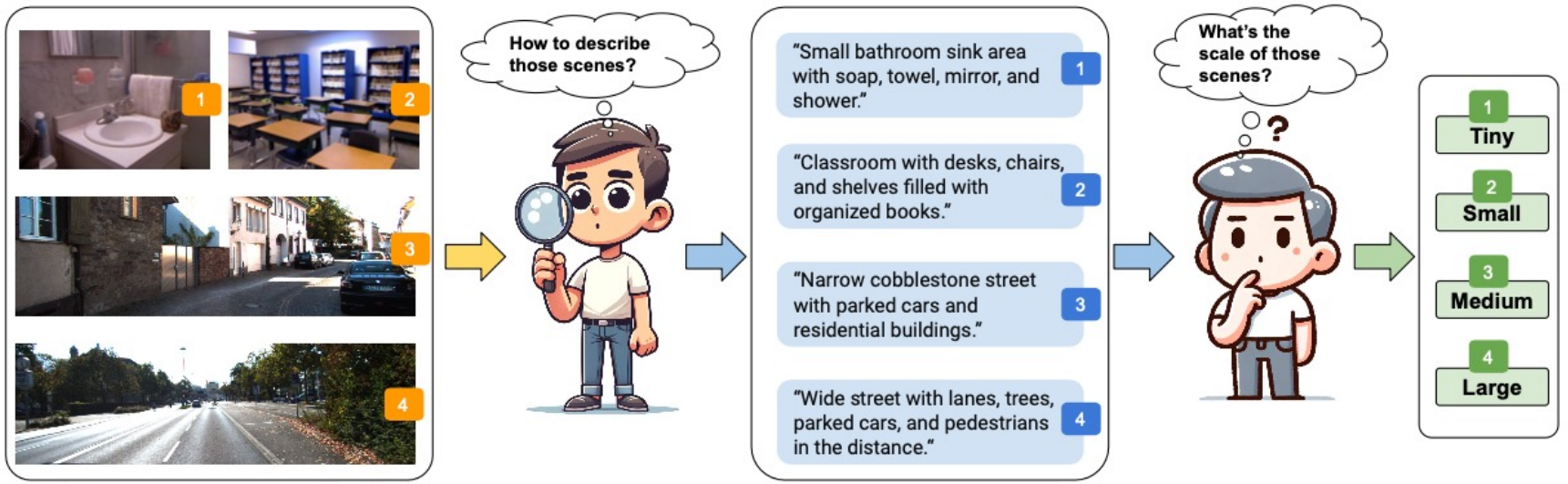


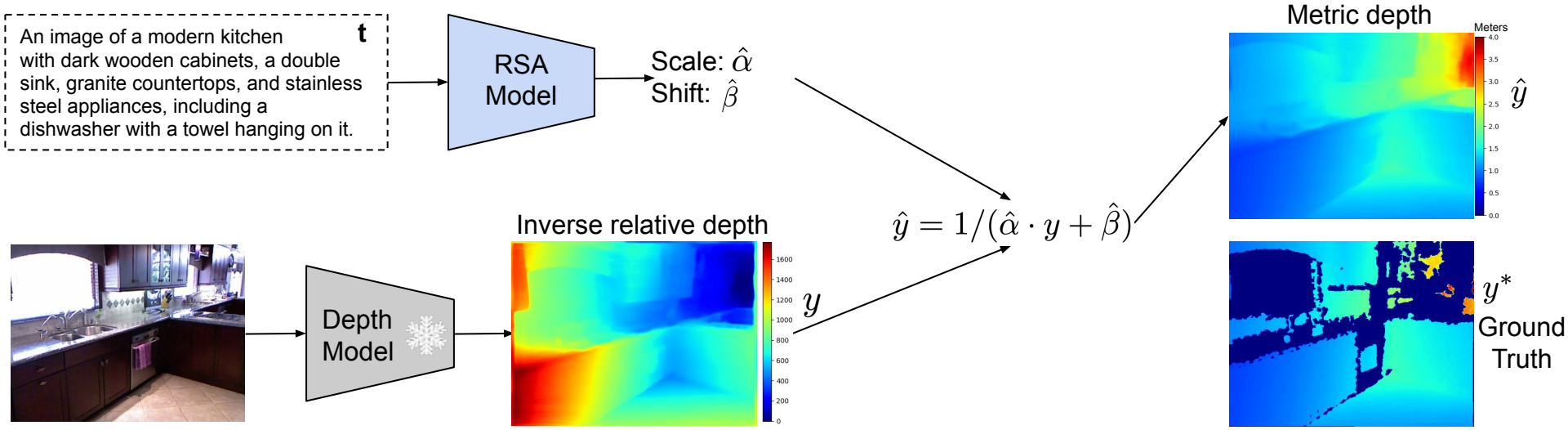
Figure 1: **Can we infer the scale of 3D scenes from their descriptions?** Consider the description above, one may observe that the scale of the 3D scene is closely related to the objects (and their typical sizes) populating it.

Depth models predicting relative depth need rescaling

- Learning to predict normalized depth map
- Generalize across different domains (indoors, outdoors)
  - Depth distribution are quite different
- **Needs to rescale relative depth into metric scale**
  - Median scaling
  - Domain-specific depth decoder
- Representative works
  - Depth Anything
  - Midas
  - DPT
  - ……

Method

- Transfer relative depth to metric depth through a linear transformation



Results on Indoors

- Better compare with:
  - Predicting scale using image
  - Global scale.
- Close or even better than oracle.
  - Median scaling, linear fit.
- Comparable or even better than other domain adaptation method.
  - Depending on the depth model we use.

Models	Scaling	Dataset	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	Abs Rel $\downarrow$	$\log_{10} \downarrow$	RMSE $\downarrow$
ZoeDepth	Image	NYUv2	0.951	0.994	0.999	0.077	0.033	0.282
DistDepth	DA	NYUv2	0.706	0.934	-	0.289	-	1.077
DistDepth	DA,Median	NYUv2	0.791	0.942	0.985	0.158	-	0.548
ZeroDepth	DA	-	0.901	0.961	-	0.100	-	0.380
ZeroDepth	DA,Median	-	0.926	0.986	-	0.081	-	0.338
	Median	NYUv2	0.736	0.919	0.981	0.181	0.073	0.912
	Linear Fit	NYUv2	0.926	0.991	0.999	0.094	0.040	0.332
DPT	Global	NYUv2	0.904	0.988	**0.998**	0.109	0.045	0.357
	Image	NYUv2	0.914	**0.990**	**0.998**	**0.097**	**0.042**	0.350
	Image	NYUv2,KITTI	0.911	0.989	**0.998**	0.098	0.043	0.355
	Image	NYUv2,KITTI,VOID	0.903	0.985	0.997	0.100	0.045	0.367
	RSA (Ours)	NYUv2	**0.916**	**0.990**	**0.998**	**0.097**	**0.042**	**0.347**
	RSA (Ours)	NYUv2,KITTI	0.913	0.988	**0.998**	0.099	**0.042**	0.352
	RSA (Ours)	NYUv2,KITTI,VOID	0.912	0.989	**0.998**	0.099	0.043	0.355
	Median	NYUv2	0.449	0.694	0.850	0.411	0.151	2.010
	Linear Fit	NYUv2	0.780	0.970	0.995	0.151	0.069	0.433
MiDas	Global	NYUv2	0.689	0.949	0.992	0.183	0.078	0.600
	Image	NYUv2	0.729	0.958	**0.994**	0.175	0.072	0.563
	Image	NYUv2,KITTI	0.724	0.952	0.992	0.173	0.074	0.579
	Image	NYUv2,KITTI,VOID	0.712	0.948	0.988	0.181	0.075	0.583
	RSA (Ours)	NYUv2	0.731	0.955	0.993	0.171	0.072	0.569
	RSA (Ours)	NYUv2,KITTI	**0.737**	**0.959**	0.993	**0.168**	**0.071**	**0.561**
	RSA (Ours)	NYUv2,KITTI,VOID	0.709	0.944	0.989	0.173	0.076	0.580
	Median	NYUv2	0.480	0.734	0.886	0.353	0.135	1.743
	Linear Fit	NYUv2	0.965	0.993	0.997	0.058	0.025	0.232
DepthAnything	Global	NYUv2	0.630	0.926	0.987	0.199	0.087	0.646
	Image	NYUv2	0.749	0.965	**0.997**	0.169	0.068	0.517
	Image	NYUv2,KITTI	0.710	0.947	0.992	0.181	0.075	0.574
	Image	NYUv2,KITTI,VOID	0.702	0.943	0.990	0.178	0.078	0.583
	RSA (Ours)	NYUv2	0.775	**0.975**	**0.997**	**0.147**	**0.065**	**0.484**
	RSA (Ours)	NYUv2,KITTI	**0.776**	0.974	0.996	0.148	**0.065**	0.498
	RSA (Ours)	NYUv2,KITTI,VOID	0.752	0.964	0.992	0.156	0.071	0.528

Table 1: **Quantitative results on NYUv2.** RSA (yellow), especially when trained with multiple datasets, generalizes better than using images to predict the transformation parameters. Global refers to optimizing a single scale and shift for the entire dataset (same scale and shift for every sample). Image denotes predicting scales and shifts using images. Red denotes scaling that uses ground truth. Median indicates scaling using the ratio between median of depth prediction and ground truth. Linear fit denotes optimizing scale and shift to fit to ground truth for each image. DA refers to domain adaptation. ZoeDepth performs per-pixel refinement.

Results on Indoors

- Better compare with:
  - Predicting scale using image
  - Global scale.
- Close or even better than oracle.
  - Median scaling, linear fit.
- Comparable or even better than other domain adaptation method.
  - Depending on the depth model we use.

Models	Scaling	Dataset	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	Abs Rel $\downarrow$	$\log_{10} \downarrow$	RMSE $\downarrow$
DPT	Median	VOID	0.782	0.962	0.990	0.150	0.064	0.340
	Global	NYUv2 (zero-shot)	0.456	0.743	0.912	0.312	0.136	0.896
	Image	NYUv2,KITTI (zero-shot)	0.516	0.812	0.936	0.289	0.112	0.634
	Image	NYUv2,KITTI,VOID	0.534	0.827	0.941	0.266	0.108	0.545
	RSA (Ours)	NYUv2,KITTI (zero-shot)	**0.601**	**0.886**	**0.970**	0.254	**0.096**	**0.444**
	RSA (Ours)	NYUv2,KITTI,VOID	0.598	0.877	0.956	**0.248**	0.100	0.475
MiDas	Median	VOID	0.500	0.781	0.899	0.347	0.130	0.829
	Global	NYUv2 (zero-shot)	0.268	0.597	0.735	0.512	0.193	1.346
	Image	NYUv2,KITTI (zero-shot)	0.304	0.626	0.812	0.487	0.159	0.913
	Image	NYUv2,KITTI,VOID	0.389	0.743	0.911	0.392	0.139	0.652
	RSA (Ours)	NYUv2,KITTI (zero-shot)	0.392	0.696	0.892	0.448	0.148	0.660
	RSA (Ours)	NYUv2,KITTI,VOID	**0.535**	**0.829**	**0.945**	**0.280**	**0.112**	**0.528**
DepthAnything	Median	VOID	0.249	0.465	0.643	0.682	0.254	1.251
	Global	NYUv2 (zero-shot)	0.084	0.194	0.376	1.674	0.389	2.046
	Image	NYUv2,KITTI (zero-shot)	0.093	0.215	0.412	1.497	0.345	1.963
	Image	NYUv2,KITTI,VOID	0.323	0.612	0.768	0.589	0.196	0.956
	RSA (Ours)	NYUv2,KITTI (zero-shot)	0.104	0.262	0.450	1.287	0.323	1.716
	RSA (Ours)	NYUv2,KITTI,VOID	**0.374**	**0.673**	**0.837**	**0.477**	**0.168**	**0.792**

Table 3: **Quantitative results on VOID.** In zero-shot generalization and multi-dataset training (including the target dataset), RSA outperforms image scaling due to the robustness of text, which supports better generalization. Please refer to Table 1 for more details about notations.

Results on Outdoors

- Better compare with:
  - Predicting scale using image
  - Global scale.
- Close or even better than oracle.
  - Median scaling, linear fit.
- Comparable or even better than other domain adaptation method.
  - Depending on the depth model we use.

Models	Scaling	Dataset	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	Abs Rel $\downarrow$	$\text{RMSE}_{\log} \downarrow$	RMSE $\downarrow$
ZoeDepth	Image	KITTI	0.971	0.996	0.999	0.054	0.082	2.281
Monodepth2	Median	KITTI	0.877	0.959	0.981	0.115	0.193	4.863
ZeroDepth	DA	-	0.892	0.961	0.977	0.102	0.196	4.378
ZeroDepth	DA,Median	-	0.886	0.965	0.984	0.105	0.178	4.194
DPT	Median	KITTI	0.950	0.994	0.999	0.069	0.100	3.365
	Linear fit	KITTI	0.974	0.997	0.999	0.052	0.080	2.198
	Global	KITTI	0.959	**0.995**	**0.999**	0.062	0.092	2.575
	Image	KITTI	0.961	**0.995**	**0.999**	0.064	0.092	2.379
	Image	NYUv2,KITTI	0.956	0.989	0.993	0.066	0.098	2.477
	Image	NYUv2,KITTI,VOID	0.952	0.987	0.993	0.068	0.098	2.568
	RSA (Ours)	KITTI	**0.963**	**0.995**	**0.999**	0.061	0.090	2.354
	RSA (Ours)	NYUv2,KITTI	0.962	0.994	0.998	**0.060**	**0.089**	2.342
	RSA (Ours)	NYUv2,KITTI,VOID	0.961	0.994	**0.999**	0.064	0.091	**2.335**
MiDas	Median	KITTI	0.856	0.959	0.988	0.138	0.204	6.372
	Linear fit	KITTI	0.824	0.952	0.989	0.154	0.174	3.833
	Global	KITTI	0.729	0.939	0.978	0.192	0.212	4.811
	Image	KITTI	0.749	0.949	0.982	0.164	0.199	4.254
	Image	NYUv2,KITTI	0.718	0.943	0.979	0.171	0.211	4.456
	Image	NYUv2,KITTI,VOID	0.683	0.931	0.972	0.165	0.232	4.862
	RSA (Ours)	KITTI	**0.798**	0.948	0.981	0.163	0.185	4.082
	RSA (Ours)	NYUv2,KITTI	0.782	0.946	0.980	0.160	0.194	4.232
	RSA (Ours)	NYUv2,KITTI,VOID	0.794	**0.960**	**0.992**	**0.155**	**0.179**	**3.989**
DepthAnything	Median	KITTI	0.925	0.986	0.996	0.091	0.129	3.648
	Linear fit	KITTI	0.824	0.896	0.922	0.149	0.224	3.595
	Global	KITTI	0.663	0.932	0.981	0.191	0.228	5.273
	Image	KITTI	0.768	0.951	0.983	0.162	0.195	4.483
	Image	NYUv2,KITTI	0.697	0.933	0.977	0.181	0.218	4.824
	Image	NYUv2,KITTI,VOID	0.678	0.924	0.974	0.186	0.243	5.021
	RSA (Ours)	KITTI	0.780	0.958	0.988	0.160	0.189	4.437
	RSA (Ours)	NYUv2,KITTI	0.756	0.956	0.987	0.158	0.191	4.457
	RSA (Ours)	NYUv2,KITTI,VOID	**0.786**	**0.967**	**0.995**	**0.147**	**0.179**	**4.143**

Table 2: **Quantitative results on KITTI Eigen Split.** RSA (yellow), especially when trained with multiple datasets, generalizes better than using images to predict the transformation parameters. Please refer to Table 1 for more details about notations.

Zero-shot Generalization

- Due to the robustness of language description:
- We achieve a better generalization ability compared with using image feature.

Models	Scaling	Dataset	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	Abs Rel $\downarrow$	$\log_{10} \downarrow$	RMSE $\downarrow$
Adabins	-	NYUv2	0.771	0.944	0.983	0.159	0.068	0.476
DepthFormer	-	NYUv2	0.815	0.970	0.993	0.137	0.059	0.408
ZoeDepth-X	Image	NYUv2	0.857	-	-	0.124	-	0.363
ZoeDepth-M12	Image	NYUv2	0.864	-	-	0.119	-	0.346
ZoeDepth-M12	Image	NYUv2, KITTI	0.856	-	-	0.123	-	0.356
	Linear Fit	SUN-RGBD	0.812	0.967	0.993	0.139	0.059	0.412
DPT	Global	NYUv2	0.773	0.945	0.984	0.154	0.071	0.482
	Image	NYUv2,KITTI	0.778	**0.953**	0.984	0.153	0.068	0.478
	RSA (Ours)	NYUv2,KITTI	0.781	**0.953**	**0.986**	0.152	0.066	0.463
	RSA (Ours)	NYUv2,KITTI,VOID	**0.788**	**0.953**	**0.986**	**0.150**	**0.065**	**0.458**
	Linear Fit	SUN-RGBD	0.632	0.912	0.971	0.241	0.102	1.132
MiDas	Global	NYUv2	0.572	0.889	0.956	0.297	0.132	1.464
	Image	NYUv2,KITTI	0.594	0.895	0.962	0.275	0.125	1.374
	RSA (Ours)	NYUv2,KITTI	0.612	0.903	0.964	0.268	0.122	1.302
	RSA (Ours)	NYUv2,KITTI,VOID	**0.623**	**0.908**	**0.968**	**0.253**	**0.116**	**1.223**
	Linear Fit	SUN-RGBD	0.878	0.979	0.995	0.113	0.054	0.332
DepthAnything	Global	NYUv2	0.534	0.872	0.951	0.313	0.138	1.692
	Image	NYUv2,KITTI,VOID	0.588	0.892	0.963	0.279	0.126	1.392
	RSA (Ours)	NYUv2,KITTI	0.621	0.915	0.970	0.238	0.099	**1.024**
	RSA (Ours)	NYUv2,KITTI,VOID	**0.645**	**0.927**	**0.978**	**0.203**	**0.095**	1.137

Table 4: **Zero-shot generalization to SUN-RGBD.** With more training datasets for scale prediction, RSA model generalizes better due to the robustness of text, but predicting scale using images suffers from domain gaps among training images. The models are tested on the Sun-RGBD without any fine-tuning. For ZoeDepth, X indicates no pre-training, and M12 indicates 12 datasets for pre-training. ZoeDepth results were taken from their original manuscripts, using a depth decoder for scaling. Please refer to Table 1 for detailed notations.

Models	Scaling	Dataset	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	Abs Rel $\downarrow$	$\text{RMSE}_{\log} \downarrow$	RMSE $\downarrow$
Adabins	-	KITTI	0.790	-	-	0.154	-	8.560
NeWCRFs	-	KITTI	0.874	-	-	0.119	-	6.183
ZoeDepth-X	Image	KITTI	0.790	-	-	0.137	-	7.734
ZoeDepth-M12	Image	KITTI	0.835	-	-	0.129	-	7.108
ZoeDepth-M12	Image	NYUv2, KITTI	0.824	-	-	0.138	-	7.225
	Linear Fit	DDAD	0.802	0.954	0.990	0.163	0.254	10.342
DPT	Global	KITTI	0.752	0.925	0.969	0.183	0.312	15.967
	Image	NYUv2,KITTI	0.763	0.931	0.975	0.179	0.308	14.468
	Image	NYUv2,KITTI,VOID	0.731	0.910	0.962	0.191	0.324	16.132
	RSA (Ours)	NYUv2,KITTI	**0.777**	0.938	0.981	0.171	0.284	13.539
	RSA (Ours)	NYUv2,KITTI,VOID	0.768	**0.942**	**0.983**	**0.165**	**0.276**	**12.437**
	Linear Fit	DDAD	0.664	0.912	0.973	0.209	0.301	18.341
MiDas	Global	KITTI	0.603	0.864	0.925	0.253	0.336	20.594
	Image	NYUv2,KITTI	0.616	0.883	0.934	0.231	0.331	20.034
	Image	NYUv2,KITTI,VOID	0.564	0.862	0.925	0.243	0.352	22.689
	RSA (Ours)	NYUv2,KITTI	0.631	0.903	**0.966**	0.223	**0.325**	**19.342**
	RSA (Ours)	NYUv2,KITTI,VOID	**0.642**	**0.908**	**0.966**	**0.218**	0.331	**18.293**
	Linear Fit	DDAD	0.673	0.932	0.983	0.182	0.286	18.423
DepthAnything	Global	KITTI	0.612	0.883	0.963	0.221	0.323	21.345
	Image	NYUv2,KITTI	0.623	0.890	0.968	0.217	0.316	20.834
	Image	NYUv2,KITTI,VOID	0.586	0.874	0.956	0.243	0.348	22.351
	RSA (Ours)	NYUv2,KITTI	0.642	0.903	**0.976**	0.207	0.303	19.715
	RSA (Ours)	NYUv2,KITTI,VOID	**0.648**	**0.905**	0.975	**0.198**	**0.297**	**18.984**

Table 5: **Zero-shot generalization to DDAD.** With more training datasets for scale prediction, RSA model achieves a better generalization due to the robustness of text, but predicting scale using images suffers from domain gaps among training images. Models are tested on the DDAD without any fine-tuning. Please refer to Table 1 and Table 4 for more details about notations.

Indoor Visualization

- RSA reduce overall error (darker in the error map) comparing with baseline (DPT with global scale).

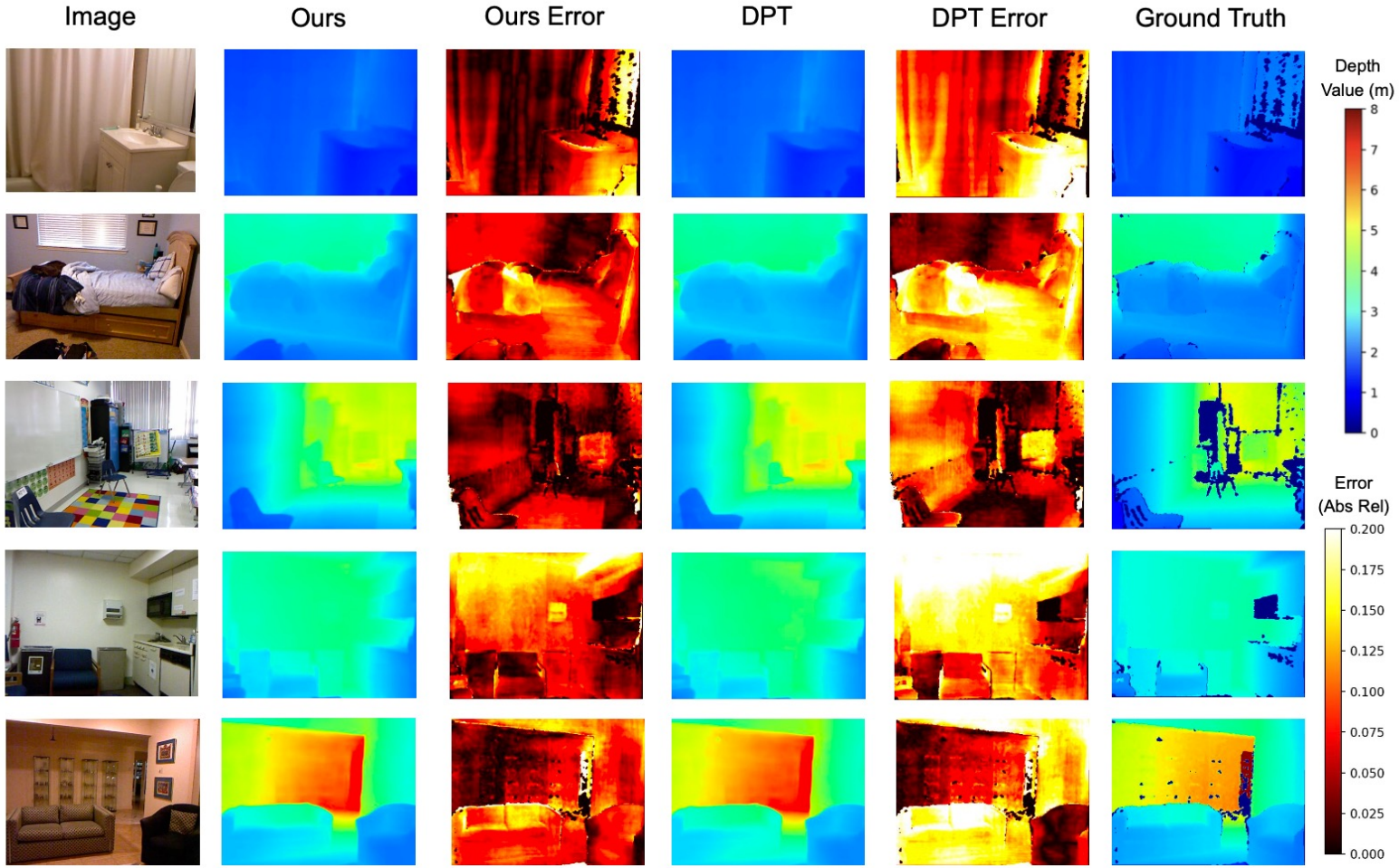


Figure 3: **Visualization of depth estimations on NYUv2.** Building upon DPT, while a better scale factor does not change the structure of the depth prediction, leading to visually similar depth maps, it significantly reduces the overall error (darker in the error map). Note: Zeros in ground truth indicate the absence of valid depth values.

Outdoor Visualization

- RSA reduce overall error (darker in the error map) comparing with baseline (DPT with global scale).

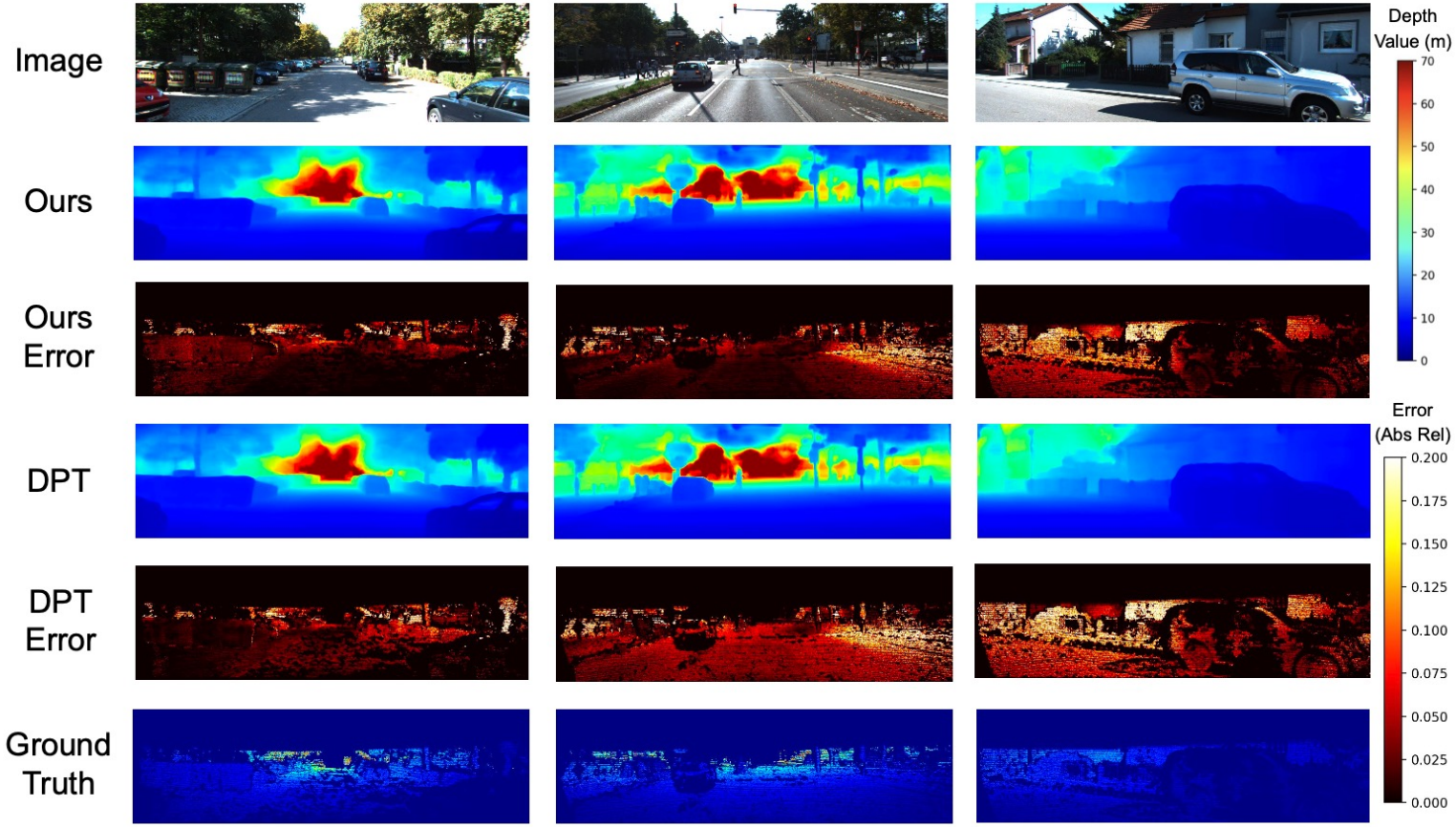


Figure 4: **Visualization of depth estimations on KITTI.** Building upon DPT, while a better scale factor does not change the structure of the depth prediction, it significantly reduces the overall error (darker in the error map). Note: Zeros in ground truth depth indicate the absence of valid depth values.

Structural text: Prompt design for RSA

- By including background in text description, scale predication is improved especially for outdoors.

Prompt	NYUv2	KITTI
"An image with $\mathbf{c}^{(1)}$, $\mathbf{c}^{(2)}$, $\mathbf{c}^{(3)}$, ... " with Object Detection Results	0.106	0.070
"An image with $K^{(1)}$ $\mathbf{c}^{(1)}$, $K^{(2)}$ $\mathbf{c}^{(2)}$, ... " with Object Detection Results	0.101	0.068
"An image with $\mathbf{c}^{(1)}$, $\mathbf{c}^{(2)}$, $\mathbf{c}^{(3)}$, ... " with Panoptic Segmentation Results	0.102	0.063
"An image with $K^{(1)}$ $\mathbf{c}^{(1)}$, $K^{(2)}$ $\mathbf{c}^{(2)}$, ... " with Panoptic Segmentation Results	**0.100**	**0.061**

Table 6: **Different prompt design for RSA.** Absolute relative errors (Abs Rel) reported. RSA models are trained using cross-datasets with the DPT model. For one given image, $\mathbf{c}^{(i)}$ is the class of a detected or segmented instance, $K^{(i)}$ is the number of all instances belonging to $\mathbf{c}^{(i)}$. By using segmentation results, the text includes background, which improves scale predication, especially for outdoors.

Natural text: Example of LLaVA generated natural text

- The image shows a bathroom with a white sink, a white towel hanging on a rack, and a soap dispenser on the sink.
- A bathroom with a white sink and a white towel.
- A bathroom with a white sink, a white towel, and a soap dispenser.
- A small bathroom with a white sink, a white towel, a small soap dispenser, a small bottle of soap, and a small bag of toiletries.
- The image shows a bathroom with a white sink, a white towel hanging on a rack, a soap dispenser, a toothbrush holder, and a small bag on the sink.

Control the scale of a scene using language

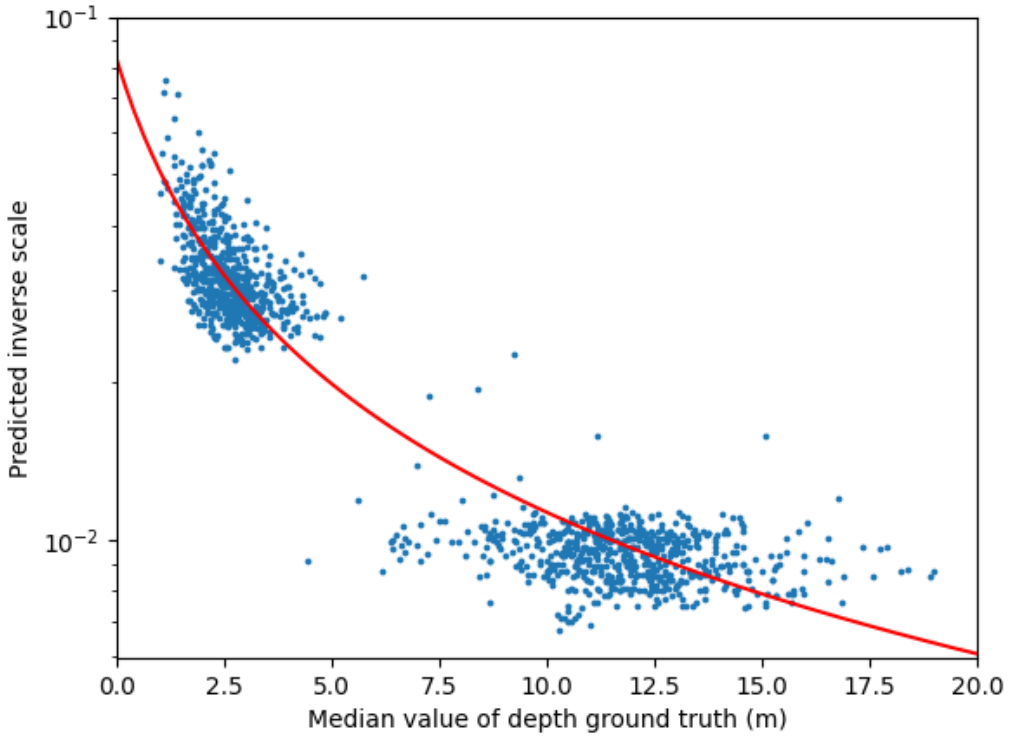
- We can control the scale of a scene by controlling its language description, to manipulate the scene in a controllable manner.

Input Text	Inv scale	Inv shift
A room with a refrigerator, a table, and a shelf.	0.0387	0.2286
A black office chair in a bedroom, next to a white door and a clothes rack.	0.0354	0.2437
The image shows a store with a variety of items for sale.	0.0276	0.1812
The image shows a classroom with desks and chairs, a bulletin board, and a clock.	0.0254	0.1633
A group of people walking down a city street.	0.0102	0.0063
A bustling city street with a white van driving down it.	0.0096	0.0053
A busy highway filled with cars, with a blue and white sign on the right side.	0.0067	0.0045

Table 7: **Sensitivity study to different text input.** We show the inverse scale and shift here; a smaller value indicates a larger scene. From top to bottom, we describe scenes from small to large scale. Results show that we could control the scale of a scene by providing different text descriptions, to better manipulate a 3D scene.

Predicted scale w.r.t true scale

- The predicted scale is proportional to the median depth (reflecting true scale of a scene), which aligned with our hypothesis.



Take Home Message

- Language can infer the scale of 3D scenes.
  - Manipulate 3D scenes in a controllable manner.
- Groud relative depth predication into metric scale with language.
  - Enpower application of sota depth models.