TR2M: Transferring Monocular Relative Depth to Metric Depth with Language Descriptions and Dual-Level Scale-Oriented Contrast

Supplementary Material

6. Dataset

As stated in the main paper, our primary datasets for training and evaluation for TR2M are NYU Depth v2 [51], KITTI [19], VOID [58], and C3VD [6]. NYU Depth v2 and VOID are indoor scene datasets; KITTI is an outdoor scene dataset, and C3VD is a surgical dataset that mainly contains scenes of colons. We also evaluate zero-shot performance on five real-world and synthetic datasets: SUN RGB-D [52], iBims-1 [30], HyperSim [48], DIODE Outdoors [53], and SimCol [47]. The first three datasets are indoors; DIODE Outdoors is outdoor scenes, and SimCol is a simulated surgical scene dataset for colons. Text descriptions for NYUv2, KITT,I and VOID are generated by [70], where LLaVAv 1.6 Vicuna and LLaVAv 1.6 Mistral [34] are utilized to generate five text descriptions each. Another five descriptions are formed by listing the main objects detected by MaskDINO [32]. The scenes in C3VD and SimCol are similar among different parts of the colon, and current VLMs can not identify them properly. Therefore, we directly use the name for the colon and the texture type to describe them. An example is: "The image shows a surgical scene of descending colon with texture type two.". We followed RSA [70] to generate text for the other datasets for zero-shot evaluation with LLaVA v1.6 Vicuna [34]. Figure 6 shows some examples of the datasets.

7. Evaluation Metric and Compared Methods

The definitions of the evaluation metrics used in this paper are listed below in 6. $AbsRel$ is the absolute relative error, $SqRel$ is the squared relative error, $RMSE$ is the root mean squared error, $RMSE_{log}$ is the root mean squared logarithmic error, log_{10} is the average (Log_{10}) error, δ_n is the accuracy with threshold, as in [5, 63, 64, 69, 70].

Table 6. Depth evaluation metrics. D and D^* are the predicted and ground truth depth maps.

Metrics Name	Definition
$AbsRel$	$\frac{1}{\lvert\mathbf{D}\rvert}\sum_{d\in\mathbf{D}}\lvert d^*-d\rvert/d^*$
$SqRel$	$\frac{1}{\lvert\mathbf{D}\rvert}\sum_{d\in\mathbf{D}}\lvert d^*-d\rvert^2/d^*$
$RMSE$	$\sqrt{\frac{1}{\lvert\mathbf{D}\rvert}\sum_{d\in\mathbf{D}}\lvert d^*-d\rvert^2}$
$RMSE_{log}$	$\sqrt{\frac{1}{\lvert\mathbf{D}\rvert}\sum_{d\in\mathbf{D}}\lvert logd^*-logd\rvert^2}$
log_{10}	$\frac{1}{\lvert\mathbf{D}\rvert}\sum_{d\in\mathbf{D}}(log_{10}d-log_{10}d^*)^2$
δ_n	$\frac{1}{\lvert\mathbf{D}\rvert}\left\lvert\left\{d\in\mathbf{D}\lvert max(\frac{d^*}{d},\frac{d}{d^*}<(1.25)^n)\right\}\right\rvert\times 100\%$

We compare our method with different types of SOTA methods to obtain metric depth. To be specific, UniK3D [42], Metric3Dv2 [24], UniDepth [41],, ZoeDepth [5] are monocular metric depth estimation methods. DepthAnything [63] (DA) and DepthAnything V2 [64] (DA V2) are monocular relative depth estimation methods fine-tuned with metric heads; DepthCLIP [72], ScaleDepth [77] and WorDepth [69] are SOTA language based metric depth estimation methods; RSA [70] is a SOTA language based relative to metric depth transformation method; We reproduced some methods for fair comparison which are marked with * in the tables.

8. More Details of the Proposed Framework

8.1. Complexity

We perform a detailed analysis of the computational cost of TR2M, presented in Table 7, and compare it to other state-of-the-art methods. The Latency calculates the time it takes for the model to obtain the corresponding metric scale depth map from the input image. To ensure a fair and consistent comparison, we use input sizes that are as similar as possible across all models. Among all the compared models, the inference latency of TR2M is only higher than that of DA V2, which can achieve real-time inference speed. Although the overall parameter amount is higher due to the use of text and image feature extractors, the trainable parameters are only 19.4M, which can be quickly adapted to downstream tasks. Meanwhile, the inference latency of the Rescale Maps estimation module is only 3.08ms, which means that TR2M can be plugged in and increase the metric depth estimation ability of many VLM models with only a small increase in parameters and inference time.

Table 7. Parameters and efficiency comparison. Comparison of performance of methods based on input size, latency, number of total parameters and trainable parameters. Tested on A6000 GPU, 32-bit precision float, and synchronized timers. TR2M's Embeds module are based on ViT-L backbone and the Relative Depth module is based on ViT-S backbone as illustrated in the main paper. All compared models are based on ViT-L backbone.

Methods	Input Size	Latency (ms)	Total Params (M)	Trainable Params (M)
ZoeDepth [5]	434 × 560	45.29	345.9	345.9
DA V2 [64]	434 × 560	21.35	334.7	334.7
Metric3Dv2 [24]	434 × 560	43.68	441.9	441.9
UniDepth [41]	434 × 560	46.83	347.0	347.0
UniK3D [42]	434 × 560	39.33	358.8	358.8
TR2M (Ours)	434 × 560	30.26	471.9	19.0
- Text & Image Embeds	-	18.90	428.1	0.0
- Rescale Maps	-	3.08	19.0	19.0
- Relative Depth	-	8.28	24.8	0.0

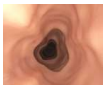
Dataset	Image	Prompt	Description
NYU v2		"Describe the image in one sentence."	"The image shows a small bathroom with a white toilet, sink, and mirror."
KITTI		"Provide a one-sentence description of the image, pay to attention object type."	"A long, empty street with parked cars on one side and train tracks on the other."
VOID		"Capture the essence of the image in a single sentence, pay attention to object relationship. "	"A white table with a blue chair in a room with a tiled floor."
SUN RGB-D		"Condense the image description into one sentence, pay attention to object size."	"The image shows a bedroom with a large wooden bed and a wooden dresser."
iBims-1		"Express the image in just one sentence, pay attention to the overall layout."	"A long, narrow, dimly lit hallway with concrete walls and a ceiling, leading to a doorway at the end."
HyperSim		"Describe the image in one sentence."	"A modern kitchen with a clean, minimalist design, featuring a central island, stainless steel appliances, and a color scheme of white and blue."
DIODE Outdoor		"Provide a one-sentence description of the image, pay to attention object type."	"The image shows a large, historic building with a prominent clock tower, surrounded by bare trees and a well-maintained lawn, with a clear blue sky above."
C3VD		-	"The image shows a medical scene of transcending colon with texture type one."
SimCol		-	"This is a medical scene of a colon."

Figure 6. Examples of the datasets and the supplementary text descriptions which are generated by VLMs [34] with input images and prompts.

8.2. Decoders and Projectors Architecture

We design the scale and shift decoders architecture similar to DPT [45] without the progressive reassemble module for our single image feature input. One convolution projection layer and two spatial resampling layers are first utilized to assemble the feature. The output maps are then obtained by projecting the features into one-dimensional maps. The maps are passed through an exponential function to ensure positive scales and shifts for optimization. The number of parameters of the decoder is about 19.4M. The projectors for depth-oriented contrastive learning are formed with two 1×1 convolutions followed by activation functions and max pooling functions. The number of parameters of the projector is about 0.19M. We utilize a projector to project the feature embedding into higher dimensions following previous works [10, 23] with many advantages. The raw features may contain redundant or low-discriminative information which could cause insufficient sensitivity in similarity calculations. Higher-dimensional spaces provide richer geometric structures, enabling contrastive losses to measure similarity more accurately. The features with higher dimensions with lower resolution also reduce computation costs, which benefit optimization efficiency.

8.3. Least-squares Criterion to Determine Pseudo Metric Depth

As illustrated in Section 3.2, we construct the pseudo metric depth with two single factors $\tilde{\alpha}, \tilde{\beta} \in \mathbb{R}^1$. A meaningful alignment could be made based on a least-squares criterion as below:

$$(\tilde{\alpha}, \tilde{\beta}) = \arg\min_{\tilde{\alpha},\tilde{\beta}} \sum_{i=1}^{HW} \left(\tilde{\alpha} D_r(i) + \tilde{\beta} - D_m^{gt}(i)\right)^2, \quad (11)$$

The aligned pseudo metric depth can be determined by: $D_m^{pseudo} = \tilde{\alpha} D_r + \tilde{\beta}$, which can be efficiently solved in

closed form by rewriting formula 11 as a least-squares problem as below:

$$\mathbf{h}^{opt} = \arg\min_{\mathbf{h}} \sum_{i=1}^{HW} \left(\overrightarrow{D_r(i)}^{\top} \mathbf{h} - D_m^{gt}(i) \right)^2, \tag{12}$$

where $\overrightarrow{D_r(i)}^{\top} = (D_r(i), 1)^{\top}$ and $\mathbf{h} = (\tilde{\alpha}, \tilde{\beta})^{\top}$. A closed-form solution for the above equation can be written as:

$$\mathbf{h}^{opt} = \left(\sum_{i=1}^{HW} \overrightarrow{D_r(i)} \overrightarrow{D_r(i)}^{\top} \right)^{-1} \left(\sum_{i=1}^{HW} \overrightarrow{D_r(i)} D_m^{gt}(i) \right). \tag{13}$$

We then use the above solution to rescale D_r to D_m^{pseudo} for subsequent evaluation to determine whether it is confident enough to be a pseudo-supervision.

8.4. Selecting Key Samples for Scale-Oriented Contrast

Multiple key samples can be chosen to be the contrastive object for the query sample. Here we choose two samples for efficiency. The first is the query sample itself, which goes through different encoders and projectors, and the scale relation is formed within the same depth map. The other sample is a random image from the datasets that contribute more negative feature samples with different depth distributions. This can be efficiently done in code by rearranging the input images at the batch level.

8.5. Supplementary Discussion

The utilization of pseudo metric depth with threshold ($\mathcal{L}_{\mathrm{tp-si}}$) is not completely beneficial for the model; for example, incorporating $\mathcal{L}_{\mathrm{tp-si}}$ negatively impacts performance on NYUv2. We choose to utilize it based on two reasons. First, using $\mathcal{L}_{\mathrm{tp-si}}$ will improve the zero-shot performance, which is a major advantage of our proposed TR2M. Second, while only supervising the model with depth ground truth obtains better performances, the model tends to estimate poor scale and shift values for distant areas where no supervision were available, resulting in poor overall qualitative performances. We therefore incorporate $\mathcal{L}_{\mathrm{tp-si}}$ with a lower weighting factors than the supervision of ground truth as a compromise.

DepthAnything [63] pointed out that semantic encoders like DINOv2 [37] tend to output similar features for different areas of an object, like the front and back of a car. Therefore, it is not beneficial to exhaustively rely on the features generated from such semantic encoders for depth estimation tasks where depth values can differ drastically among different areas, even adjacent pixels within the same object. DepthAnything chooses to set a tolerance margin to prevent the model from completely aligning their model with DINOv2. We seek to solve this problem by enforcing the model to capture the depth distribution knowledge within the image. The model learns to generate similar features when two objects are near in depth and dissimilar features for different areas within an object that are far in depth. This learning objective will not cause the model to generate similar features for different objects all the time, as the depth distributions of different objects vary across different images in the dataset. The model captures the inherent relations of objects to determine whether they should be close or distant in the depth ranges. We also conduct experiments as shown below, demonstrating that enforcing features to be in line with the variation in depth is beneficial for depth estimation tasks.

This work demonstrates that effective metric depth estimation models with strong generalization capabilities need not require massive training datasets or separate, large-scale architectures when leveraging readily available image or text embeddings—a common feature in multimodal tasks. To capitalize on these pre-existing representations, we freeze the feature extraction component of our framework. Our results show that utilizing such embeddings enables superior performance with an extremely lightweight architecture (TR2M + ViT-S, totaling approximately 43M parameters) compared to significantly larger metric depth models (e.g., UniDepth, UniK3D with 345M+ parameters).

It is important to note that our objective is not to surpass the performance of established metric depth models on benchmarks such as NYU or KITTI. Firstly, as indicated in the table of the main paper section, our approach represents a distinct methodological paradigm. Secondly, our model achieves the results using substantially less training data and a dramatically reduced parameter count. Crucially, despite these differences and constraints, our method achieves the best average performance in zero-shot transfer scenarios compared to other state-of-the-art methods, as referenced in the main paper section.

9. More Experiments and Analysis

Experiment Results on VOID and C3VD. Table 8 presents the results on VOID and C3VD. We can notice that without being pre-trained on surgical datasets, DepthAnything can not handle such big domain gaps without fine-tuning. Our method obtains the best evaluation metric except for $AbsRel$ and δ_1 compared to EndoDAC, which is an SOTA surgical depth estimation network.

Language resolves scale ambiguity. To demonstrate the effects of language to resolve scale ambiguity, we conduct an additional experiment with toy-model scenes. During training, we randomly modify part of the supervision by

Table 8. Quantitative results on VOID and C3VD.

Method	VOID			C3VD		
	$AbsRel \downarrow$	$RMSE \downarrow$	$\delta_1 \uparrow$	$AbsRel \downarrow$	$RMSE \downarrow$	$\delta_1 \uparrow$
DA	0.059	0.160	0.969	0.183	10.236	0.728
RSA	0.477	0.792	0.374	0.172	7.686	0.736
EndoDAC*	-	-	-	**0.083**	4.655	**0.949**
TR2M (Ours)	**0.103**	**0.205**	**0.899**	0.092	**2.952**	0.918

rewriting some descriptions to explicitly indicate a miniature scale (e.g., "a house" $\rightarrow$ "a house toy model") and proportionally downscaling the ground-truth depth to match the expected toy scale. We further collect a small toy-model dataset with ground-truth object length annotations for evaluation. As shown in Table 9 and Figure 7, TR2M with language guidance better captures the correct global scale, whereas vision-only baselines tend to interpret miniature scenes as real indoor environments. Moreover, removing the text component from TR2M also leads to ambiguous scale estimates. These results highlight the importance of language cues in resolving scale ambiguity for metric depth prediction.

Table 9. Zero-shot results on toy model dataset.

Method	UniK3D (T) [40]	UniK3D (P) [40]	TR2M (Only Vision)	TR2M (Ours)
$AbsRel \downarrow$	7.694	9.896	10.582	**0.276**
$RMSE \downarrow$	0.775	0.796	0.804	**0.023**

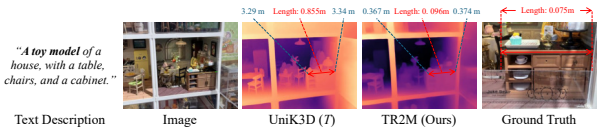


Figure 7. Qualitative examples on toy model dataset. TR2M predicts a scale close to the GT, while the image-only model misestimates it.

Incorporating Scale-Oriented Contrast to Depth Model. To verify the effectiveness and universality of the proposed Scale-Oriented Contrastive learning (SOC), we evaluate on directly implementing the proposed SOC on an existing monocular depth estimation method. We reproduce NeWCRFs [68] with and without SOC as an additional module while maintaining all the other settings. The results are shown in Table 10 in which the model obtains improvement in all metrics with the implementation of SOC. The improvement for zero-shot evaluation on iBims-1 is more significant, e.g. $AbsRel$ 0.223 $\rightarrow$ 0.210. The results demonstrate that enforcing feature embeddings aligned with depth distribution is beneficial for depth estimation tasks. The proposed SOC can be implemented as a plug-in module to existing depth estimation frameworks to promote overall performance.

Table 10. Quantitative results on implementing Scale-Oriented Contrastive on monocular depth estimation method.

Method	NYUv2			iBims-1		
	$AbsRel \downarrow$	$RMSE \downarrow$	$\delta_1 \uparrow$	$AbsRel \downarrow$	$RMSE \downarrow$	$\delta_1 \uparrow$
NeWCRFs	0.090	0.324	0.929	0.223	0.891	0.541
NeWCRFs + SOC	**0.089**	**0.318**	**0.932**	**0.210**	**0.882**	**0.550**

Impact of Relative Depth Model. We conduct an ablation study on the choice of backbone relative depth estimation model. We compare on three different frameworks: MiDas [46], DPT [35] and DepthAnything (DA) [63]. We follow RSA [70] to choose MiDas 3.1 Swin2_large-384 with 213M parameters for MiDas, DPT-Hybrid with 123M parameters for DPT and DA-Small with 25M parameters for DepthAnything. As shown in Table 11, DA obtains the best performance for all metrics of the three datasets. The performances do not drop too much from MiDas and DPT to DA because the pixel-wise rescaling method of TR2M bridges the depth perception gap between different depth estimation networks. Nevertheless, a great relative depth estimation model could still benefit the overall performance, leading us to choose DepthAnything as our backbone model.

Table 11. Ablation study on the relative depth model.

Backbone	NYUv2		iBims-1		DIODE Outdoor	
	$AbsRel \downarrow$	$\delta_1 \uparrow$	$AbsRel \downarrow$	$\delta_1 \uparrow$	$AbsRel \downarrow$	$\delta_1 \uparrow$
MiDas	0.089	0.934	0.168	0.707	0.690	0.258
DPT	0.088	0.942	0.164	0.715	0.682	0.262
DA (Ours)	0.082	0.954	0.154	0.736	0.673	0.274

Impact of Different Text Description. In order to validate the impact of text description on the results, we evaluate our model given the same image input with different language inputs. We test on three settings: 1. The correct text description; 2. The incorrect description but within the same domain (Indoor, Outdoor or Surgical); 3. Incorrect description from a different domain. The results are presented in Table 12. Performance degradations can be noticed for all datasets when an incorrect description from the same domain is used. It is worth noting that, the model exhibits varying capabilities in scale perception when incorrect descriptions from different domains are utilized. Metrics for C3VD with surgical image and text description from an incorrect domain as inputs drops significantly for both $AbsRel$ and δ_1. We believe that the low diversification in colon text descriptions leads the model to overfit on certain descriptions. This promotes future research on enhancing the diversity and robustness of text descriptions.

Impact of Text Description over Different Approaches. We study how the granularity of text descriptions affects

Table 12. Ablation study on text description.

Text Description Type	NYUv2		iBims-1		C3VD	
	AbsRel ↓	δ_1 ↑	*AbsRel* ↓	δ_1 ↑	*AbsRel* ↓	δ_1 ↑
Incorrect Domain	0.096	0.935	0.164	0.718	0.221	0.445
Correct Domain, Incorrect Scene	0.091	0.940	0.159	0.727	0.108	0.901
Correct Domain, Correct Scene	0.082	0.954	0.154	0.736	0.092	0.918

performance by comparing our original text prompts with three simpler alternatives (Table 13). Specifically, we consider: (1) **scene category only** (e.g., "indoor", "outdoor"); (2) **functional category** (e.g., "studying", "driving"); and (3) **object-level statistics** (e.g., "[4, 10]" indicating 4 categories and 10 objects). Results on NYUv2, iBims-1, and DIODE Outdoor show a consistent trend: all simplified texts degrade performance, while the **original text** yields the best results across all datasets, achieving the lowest Abs Rel and highest δ_1. This suggests that richer natural-language descriptions provide more informative scale cues than coarse scene/functional labels or simple object-count statistics.

Table 13. Ablation study on text over simpler approaches.

Type of Text	NYUv2		iBims-1		DIODE Outdoor	
	AbsRel ↓	δ_1 ↑	*AbsRel* ↓	δ_1 ↑	*AbsRel* ↓	δ_1 ↑
Approach 1	0.086	0.945	0.161	0.705	0.688	0.262
Approach 2	0.084	0.947	0.159	0.709	0.685	0.264
Approach 3	0.091	0.930	0.183	0.654	0.705	0.245
Origin text	**0.080**	**0.956**	**0.150**	**0.740**	**0.668**	**0.279**

Impact of Scale and Shift Map. We propose to use two maps: scale and shift maps, to rescale the relative depth map to the metric depth map, considering it as an affine transformation. We conduct an ablation study on the impact of two maps as shown in Table 14. Using only a single map results in distinct degradation in performance in all metrics of the three datasets. The performance of utilizing only the scale map shows minor degradation, for example, 3.7% drop in δ_1 of NYUv2 and 19.5% increase in *AbsRel* of iBims-1. By contrast, using the shift map only leads to a significant performance decrease, for example, 12.8% drop in δ_1 of NYUv2 and 79.0% increase in *AbsRel* of iBims-1. The reason is that the production operation between the scale map and the depth map enables the value to alter arbitrarily, satisfying the value gap between relative depth and metric depth. Shift map cannot change the overall scale of the depth map. It can only enlarge the depth value, which limits the transformation from relative to metric. As a result, TR2M exhibits both scale and shift maps to conduct the transformation, which guarantees flexible and generalizable rescaling.

Class Number of Depth. The proposed scale-oriented contrast learning first classifies the depth values based on their distribution and enforces attraction on features with the same distribution and repulsion on features with different distributions. Therefore, we conduct an ablation study on the corresponding class number $|\mathcal{C}|$ of the depth value as shown in Table 15. The framework obtains the best performance when classifying the depth value into 20 classes. When the class number is too small, the depth range of a class is too wide leading little differences between features with a large gap in depth values; when the class number is too large, the depth range of a class is too small leading too much differences between features with a small gap in depth values. Correctly selecting the appropriate number of class is necessary to ensure consistency between features and depth distribution, thereby enhancing the model's zero-shot capability.

Table 14. Ablation study on utilizing different approaches to rescale depth.

Scale	Shift	NYUv2		iBims-1		DIODE Outdoor	
		AbsRel ↓	δ_1 ↑	*AbsRel* ↓	δ_1 ↑	*AbsRel* ↓	δ_1 ↑
Factor	Factor	0.115	0.912	0.189	0.636	0.740	0.231
Map	None	0.102	0.923	0.177	0.663	0.727	0.238
None	Map	0.256	0.836	0.356	0.453	0.786	0.205
Map	Map	0.082	0.954	0.154	0.736	0.673	0.274

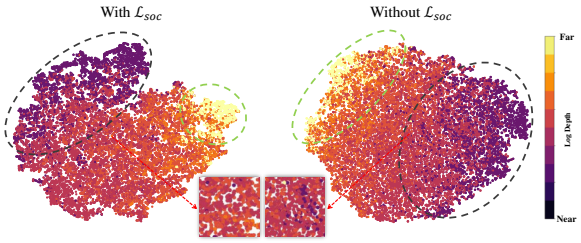


Figure 8. Visualization of embedding space with t-SNE. Features with the same depth ranges are clustered more closely (Compare the circles with the same color) with $\mathcal{L}_{soc}$. The feature distribution is more in line with the variation in depth, which is beneficial for the robustness of depth estimation.

Visualization of embedding space with t-SNE. Figure 8 shows the visualization of embedding space with t-SNE. Features with the same depth ranges are clustered more closely (Compare the circles with the same color) with $\mathcal{L}_{soc}$. The feature distribution is more in line with the variation in depth, which is beneficial for the robustness of depth estimation.

Table 15. Ablation study on the number of classes of depth map.

Class Number of Depth	NYUv2		iBims-1		DIODE Outdoor	
	AbsRel ↓	δ_1 ↑	*AbsRel* ↓	δ_1 ↑	*AbsRel* ↓	δ_1 ↑
5	0.086	0.948	0.157	0.728	0.682	0.264
10	0.083	0.953	0.155	0.730	0.679	0.268
20	0.082	0.954	0.154	0.736	0.673	0.274
40	0.083	0.952	0.156	0.731	0.681	0.269

10. Limitation and Future Work

While pixel-wise rescaling maps could make the rescaling process more precise and correct erroneous regions generated by the relative depth model, it also tends to make the model overfit to the scale of certain objects and cause wrong alignment compared to rescaling factors. Also, the current light-weight architecture is not powerful enough to handle every situation, resulting in unclear edges and details of the transferred depth maps in some circumstances. Additionally, our approach enables flexible input descriptions for transferring relative depth to metric depth. While this grants users precise control over 3D reconstruction results, it also introduces risks that bad users could choose the wrong descriptions, causing incorrect predictions. Future research may include extending TR2M to more precise and higher resolution transferring and investigating the model's robustness to erroneous text descriptions.

11. More Qualitative Results

Please refer to the following pages for additional qualitative results on the datasets utilized in the paper. We compare our model with the DepthAnything-Large [63] model fine-tuned on separate datasets for metric depth estimation. As discussed above, the light-weight architecture may cause unclear edges in some circumstances, but our method is capable of capturing some objects lost in previous methods, and our method is more accurate in the actual metric scale.

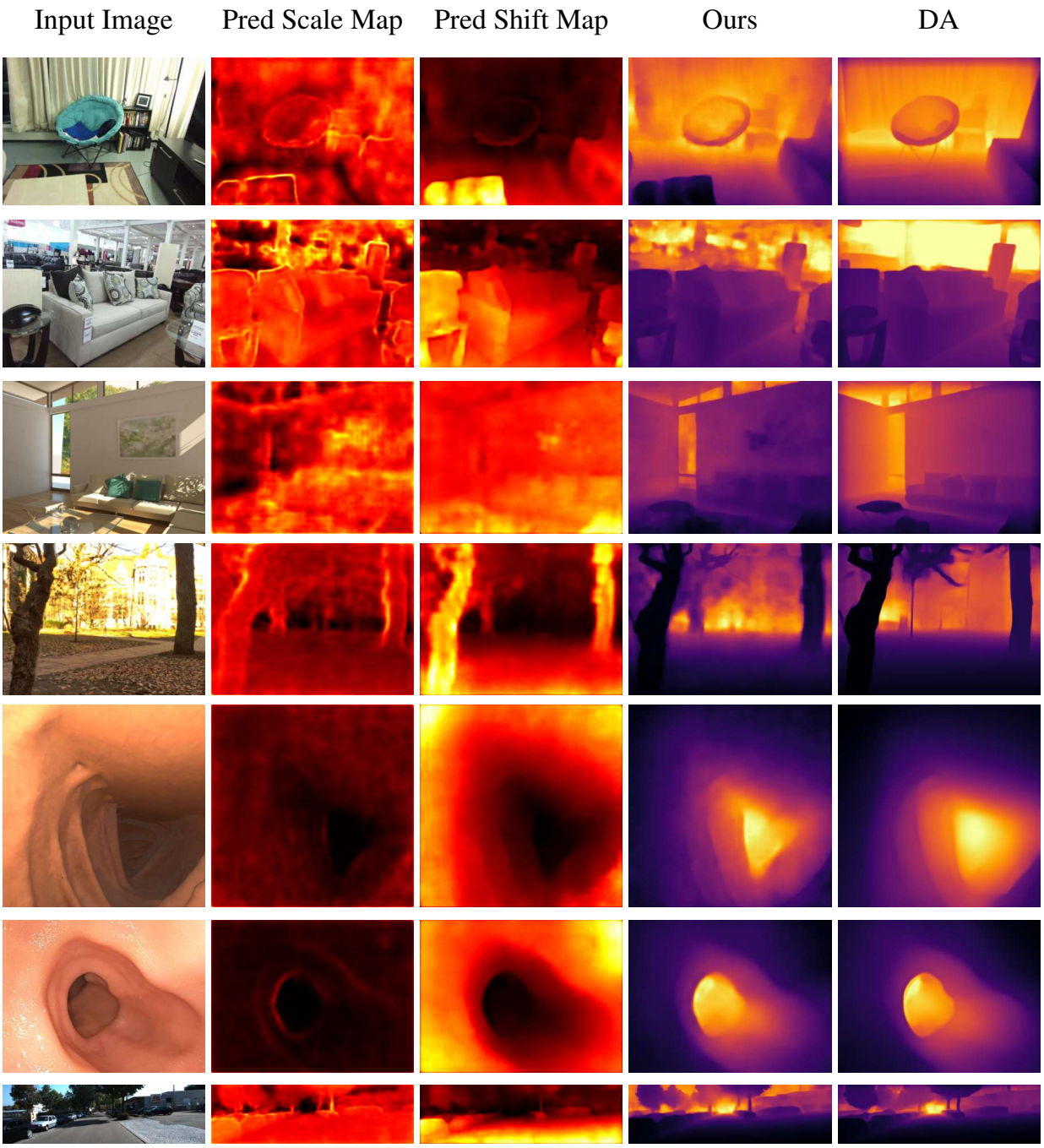


Figure 9. Example qualitative results on the datasets utilized in the paper. We compared our method with the DepthAnything-Large [63] model fine-tuned on separate datasets for metric depth estimation. The darker color denotes the closer distance.