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
[1]{ziyao.zeng, hyoungseob.park, daniel.wang.dhw33}@yale.edu
[1]{fengyu.yang, alex.wong}@yale.edu
[2] wuyangchao1997@g.ucla.edu [2]{soatto,lao}@cs.ucla.edu [3]hong@cau.ac.kr

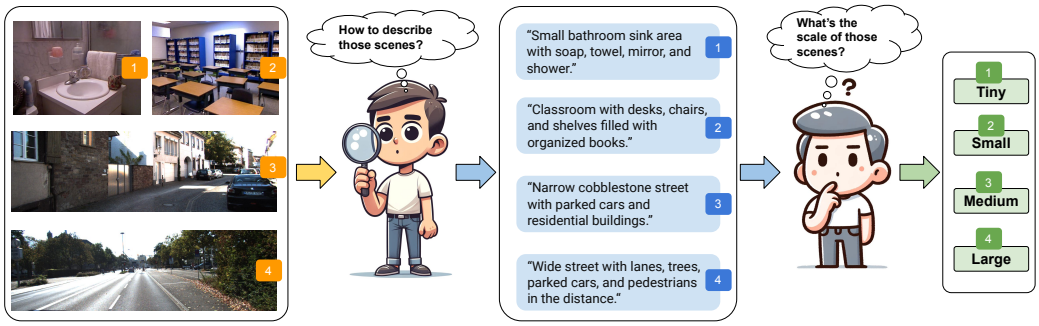

Figure 1: **Can we infer the scale of 3D scenes from their descriptions?** Consider the description above, one may observe that the scale of the 3D scene is closely related to the objects (and their typical sizes) populating it.

Abstract

We propose a method for metric-scale monocular depth estimation. Inferring depth from a single image is an ill-posed problem due to the loss of scale from perspective projection during the image formation process. Any scale chosen is a bias, typically stemming from training on a dataset; hence, existing works have instead opted to use relative (normalized, inverse) depth. Our goal is to recover metric-scaled depth maps through a linear transformation. The crux of our method lies in the observation that certain objects (e.g., cars, trees, street signs) are typically found or associated with certain types of scenes (e.g., outdoor). We explore whether language descriptions can be used to transform relative depth predictions to those in metric scale. Our method, RSA, takes as input a text caption describing objects present in an image and outputs the parameters of a linear transformation which can be applied globally to a relative depth map to yield metric-scaled depth predictions. We demonstrate our method on recent general-purpose monocular depth models on indoors (NYUv2, VOID) and outdoors (KITTI). When trained on multiple datasets, RSA can serve as a general alignment module in zero-shot settings. Our method improves over common practices in aligning relative to metric depth and results in predictions that are comparable to an upper bound of fitting relative depth to ground truth via a linear transformation. Code is available at: https://github.com/Adonis-galaxy/RSA

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-3580

1 Introduction

3-dimensional (3D) reconstruction from images is an ill-posed problem due to the loss of a dimension through perspective projection during the image formation process: Any point along the ray of projection can yield the same image coordinate. This extra degree of freedom is often addressed by using multiple images of the same scene, such as stereo or video. While an additional image (assuming co-visible sufficiently exciting textures across both images) allows one to triangulate unique points in space, a scale ambiguity exists with the absence of camera calibration, measurements from an additional sensor (e.g., range, initial), or a strong prior (e.g. a supervised training set). One may argue that, such additional information should already be available during data collection. Still, modern large-scale training [44, 67] often utilizes data from diverse sources with drastically diverse setups, making resolving the scale ambiguity issue crucial.

When it comes to monocular depth estimation, which predicts a dense depth map from a single image, the problem is also ill-posed in that one cannot measure the distance from the camera from a single view. Hence, to make inference possible, one must rely on the existence of a training set. While one option is to "bake in" an additional bias of scale by training on a number of different datasets (indoors and outdoors) and attributing depth to pixel intensities, these dataset-specific biases come at the cost of generalization, limiting model transfer from one domain (indoor) to another (outdoor), let alone mixing multiple data sources. Existing monocular depth methods resort to predicting relative (normalized, inverse) depth to factor out the scale biases, but leave behind practical utility, as a trade-off, in downstream applications in spatial tasks such as manipulation, planning, and navigation.

We consider whether an additional modality can be used to resolve the scale ambiguity in single-image 3D reconstruction, i.e., transforming scaleless relative depth to metric depth. One might observe that natural (including man-made) scenes do not occur by chance, but rather by design, with the regularity of object-scene co-occurrences [18, 28]: Certain scenes (e.g., outdoors) are composed of certain categories of objects (e.g., cars, trees, buildings) and associated with a certain order of magnitude in scale (e.g. tens of meters). Hence, we hypothesize that language, in the form of text captions or descriptions, can be used to infer the scale of the 3D scene and to transform relative depth to metric depth. The choice of language also has practical value in that it does not require costly data acquisition with an additional synchronized and calibrated sensor (e.g., lidar, time-of-flight). With publicly available pre-trained panoptic segmentation or object detection models, image captioners, and vision-language models, one can automate the data acquisition, training, and inference process. Nonetheless, they are not a necessary part of the work, but may facilitate ease of use, to simulate the language description provided by human users in practical scenarios.

To test the feasibility of our hypothesis, we consider monocular depth estimation, where a strong prior is necessary for inference; this prior may come from an image, or an independent modality such as language. Specifically, we consider monocular depth models belonging to the general-purpose, relative depth estimation paradigm to control for side effects such as the scale learned along with shapes in existing works that focus on predicting metric depth for a specific dataset. To this end, we assume access to a pre-trained scaleless monocular depth estimation model; aside from a set of images, we also assume text captions describing them. As the standard procedure to factor out scale is to normalize through a linear transformation, we propose to learn a parameterized function that predicts the parameters of a linear transformation based on the text description. Applying them to a relative depth map and inverting its values will transform it to a metric-scaled depth map to resolve the scale ambiguity. We term our method RSA, as an acronym for "Resolving Scale Ambiguities".

In one mode, like existing works that also transform relative to metric depth, albeit with images and domain-specific scaling [3], we train specific models of RSA for specific datasets (e.g., NYUv2 [46], KITTI [15], VOID [58]). In another mode, when trained on multiple datasets across different domains, RSA generalizes well and can not only handle images sampled from the datasets it was trained on, but also those from novel datasets in a zero-shot manner. The use of language, which is invariant to illumination, object orientation, occlusion, scene layout, etc., many of the nuisance variables that vision algorithms are sensitive to, demonstrates a promising avenue for general-purpose relative to metric scale recovery to complement the growing works in monocular depth estimation. We evaluate our method on indoor and outdoor benchmarks, where we improve over common practices in aligning relative depth to metric scale. We show that using RSA is comparable to matching relative depth via a linear transformation to the ground truth, or scaling using the median value of the ground truth, both of which are considered oracle relative to metric recovery methods.

Our contributions are as follows: (i) We proposed a novel formulation of the relative to metric depth transfer. (ii) We demonstrate the feasibility of inferring scale from language and as a general alignment module. (iii) We performed extensive experiments to validate performance on indoor and outdoor domains, sensitivity to text caption, and zero-shot generalization. To the best of our knowledge, we are the first to use language as a means of relative to metric depth alignment.

2 Related Work

Monocular depth estimation using metric depth. Metric depth models learn to infer pixel-wise depth in metric scale (i.e. meters) by minimizing loss between depth predictions and ground-truth depth maps [2, 12, 26, 35, 55, 71, 79]. Each model typically applies to only one data domain in which it is trained, with similar camera parameters and object scales. Specifically, DORN [12] leverages a spacing-increasing discretization technique. AdaBins [2] partitions depth ranges into adaptive bins. NeWCRFs [71] uses neural window fully-connected CRFs to compute energy. When ground-truth depth is not available, self-supervised approaches [4, 11, 23, 27, 30, 38, 41, 51, 52, 53, 59, 61, 70, 74, 80, 81, 84, 86] rely on geometric constraints, where scale is attributed through lidar [10, 37, 40, 57, 56, 58, 63, 68], radar [47], binocular images [14, 17, 16], or inertials [20, 54]. However, models that predict metric depth are limited to specific datasets or scenes, and sensors [60]; thus, they do not generalize well. RSA aims to serve as a general alignment module that can predict metric depth based on relative depth across different domains.

Monocular depth estimation using relative depth. Trained across different domains, self-supervised depth estimators trained with multi-view photometric objects produce up-to-scale predictions that are linearly correlated with their absolute depth values across the domain [8, 5, 20, 21, 54, 62, 64], thus requiring scaling of their depth prediction. However, due to the scale-ambiguity of multi-view photometric objective, such models are normally evaluated by aligning predictions to ground-truth at test time (typically median-scaling [16, 17, 86]), at the expense of practicality since ground-truth might not be feasible during real-world application. To enable generalization across different scenes, some (semi-) supervised depth models trained with single images adopt image-level normalization techniques to generate affine-invariant depth representations (i.e. relative depth) [24, 43, 44, 67, 73]. HND [73] hierarchically normalizes the depth representations with spatial information and depth distributions. Depth Anything [67] learns from large-scale automatically annotated data. DPT [43] leverages vision transformers using a scale- and shift-invariant trimmed loss. MiDas [44] mixes multiple datasets with training objectives invariant to depth range and scale. Marigold [24] associates fine-tuning protocol with a diffusion model. However, by definition, relative depth is scaleless, which limits their applications that require metric scaled reconstruction. Our proposed RSA addresses this limitation by grounding relative depth into a metric scale, enabling metric scale 3D reconstruction.

Relative to metric depth transfer. To transform the predicted relative depth into metric depth for evaluation and real-world application, ZoeDepth [3] fine-tunes a metric bins module on each metric depth dataset to output metric depth. Depth Anything [67] follows ZoeDepth [3] using a decoder where a metric bins module computes per-pixel depth bin centers that are linearly combined to output the metric depth. MiDas [44] and Marigold [24] use linear fit to align predictions and ground truth in scale and shift for each image in inverse-depth space based on the least-square criterion before measuring errors. DPT [43] fine-tunes a global scale and shift on metric depth datasets. DistDepth [62] conducts transfer by leveraging left-right stereo consistency to integrate metric scale into a scale-agnostic depth network. ZeroDepth [21] achieves transformation using input-level geometric embedding to learn an indoor scale prior over objects via a variational latent representation. However, metric decoders are limited to specific datasets, and aligning scale and shift of predictions requires ground truth during test time. RSA produces scale using text to transfer relative depth to metric depth across domains and does not require ground truth during test time.

Language model for monocular depth estimation. Vision-Language models [6, 32, 33, 39, 42, 49] acquire a comprehensive understanding of languages and images through pre-training under diverse datasets, thus forming an effective baseline for downstream tasks [34, 65, 69, 76, 78, 79, 88, 66, 50]. Typically, CLIP [42] conducts contrastive learning between text-image pairs, empowering various tasks like few-shot image classification [13, 75, 77, 85], image segmentation [45, 83], object detection [45, 87], and 3D perception [22, 76, 79, 88]. In light of their emerging ability, some works [1, 22, 79, 82] have tried to apply vision-language models for monocular depth estimation. WorDepth [72] learned the distribution 3D scenes from text captions. DepthCLIP [79] leverages the

 https://doi.org/10.52202/079017-3580

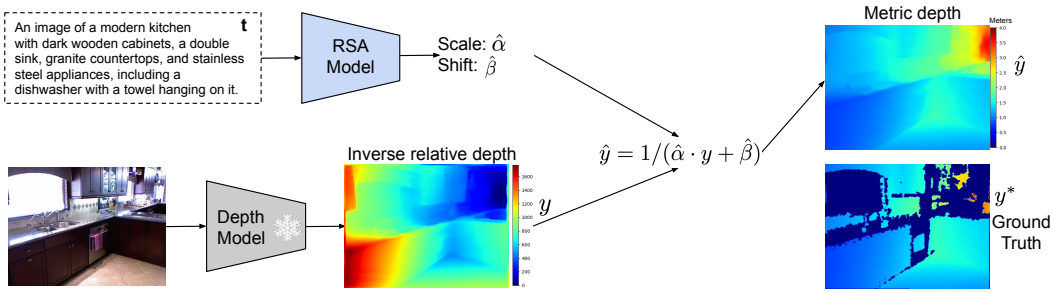

Figure 2: **Overview.** We infer scale and shift from the language description of an image to transform the inverse relative depth from the depth model into metric depth (absolute depth in meters) prediction.

semantic depth response of CLIP [42] with a depth projection scheme to conduct zero-shot monocular depth estimation. Hu et al. [22] extends DepthCLIP with learnable prompts and depth codebook to narrow the depth domain gap. Auty et al. [1] modifies DepthCLIP [79] using continuous learnable tokens in place of discrete human-language words. In contrast, RSA uses language to directly predict scale, which serves as an explicit scaling constraint by transforming relative depth into metric scale.

3 Method

We consider a dataset $\mathcal{D} = \{I^{(n)}, \mathbf{t}^{(n)}, y^{*(n)}\}_{n=1}^{N}$ with N samples of synchronized RGB image, text descriptions, depth maps, where $I : \Omega \subset \mathbb{R}^2 \mapsto \mathbb{R}^3$ denotes an image, $y^* : \Omega \subset \mathbb{R}^2 \mapsto \mathbb{R}_+$ the ground-truth metric depth map, $\mathbf{t}$ a text description of the image, and Ω the image space. We assume access to a pretrained monocular depth estimation model h_θ for the purpose of learning the parameters to predict the transformation between relative and metric depth. Given an RGB image, a monocular depth estimation model predicts inverse relative depth $y : \Omega \subset \mathbb{R}^2 \mapsto \mathbb{R}_+$ using a parameterized function h realized as a neural network, i.e., $y := h_\theta(\cdot)$. To recover metric-scale from (scaleless) inverse relative depth, we consider a global linear transformation through the use of a language description pertaining to the image of the 3D scene. Given the text description $\mathbf{t}$ of an image as input, our method, RSA, predicts a pair of scalars denoting the scale and shift parameters of the transformation: $(\hat{\alpha}, \hat{\beta}) = g_\psi(\mathbf{t}) \in \mathbb{R}^2$. The metric depth prediction is obtained by $\hat{y} = 1/(\hat{\alpha} \cdot y + \hat{\beta})$.

RSA. To infer the parameters of the linear transformation to align relative depth to metric scale, we employ the existing pretrained CLIP text encoder [42] as a feature extractor. Having been trained on a large scale dataset, CLIP offers an latent space suitable to preprocess object-centric text descriptions. We note that CLIP text encoder is frozen within RSA. Given text descriptions $\mathbf{t} = \{t_1, t_2, ...\}$, we first encode them into text embeddings and feed them into a 5-layer shared multi-layer perceptron (MLP) to project them into $k = 256$ hidden dimensions followed by two separate sets of 5-layer MLPs, one serves as the output head $\psi_{\hat{\alpha}} : \mathbb{R}^k \mapsto \mathbb{R}_+$ for scale parameter $\hat{\alpha}$ and the other as the output head $\psi_{\hat{\beta}} : \mathbb{R}^k \mapsto \mathbb{R}_+$ for shift $\hat{\beta}$ parameter. Scale and shift are assumed to be positive in favor of optimization. For ease of notation, we refer to ψ as the parameters for the shared MLP as well as the output heads, where $(\hat{\alpha}, \hat{\beta}) = g_\psi(\mathbf{t})$.

Optimizing RSA involves minimizing a supervised loss with respect to ψ, which requires a forward pass of a given training image $I^{(n)}$ through the monocular depth model to yield $y^{(n)} = h_\theta(I^{(n)})$. To ensure that the monocular depth model does not drift and update its parameters during the training of RSA, we freeze θ while optimizing for ψ. Hence, training entails minimizing an L1 loss:

$$\psi^* = \arg\min_{\psi} \sum_{n=1}^{N} \frac{1}{|M^{(n)}|} \sum_{x \in \Omega} M^{(n)}(x) |\hat{y}^{(n)}(x) - y^{*(n)}(x)|, \tag{1}$$

where $\hat{y}^{(n)} = 1/(\hat{\alpha}^{(n)} \cdot y^{(n)} + \hat{\beta}^{(n)})$ denotes the predicted metric-scale depth aligned from relative depth $y^{(n)}$, $x \in \Omega$ denotes an image coordinate, and $M : \Omega \mapsto \{0, 1\}$ denotes a binary mask indicating valid coordinates in the ground truth depth y^* with values greater than zero.

Text prompt design. To test our hypothesis, we require text descriptions to be paired with images. As standard benchmarks do not provide the text description of each image, we extend existing datasets by associating each image with several text descriptions. To achieve this, we propose to use off-the-shelf models to generate different kinds of text. First, we considered structured text, which adheres to a

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
DPT	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
	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
MiDas	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
	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
DepthAnything	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
	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.

certain template. We use a panoptic segmentation model MaskDINO [31] to extract the significant objects and background in the image. For an input image I, the segmentation model returns a set of B object and background instances $\{\mathbf{n}^{(i)}, \mathbf{c}^{(i)}\}_{i=1}^{B}$, where $\mathbf{c}^{(i)}$ denotes the class of the object or background, and $\mathbf{n}^{(i)}$ denotes the number of instances of the object or background. Using the set of instances, a structured caption for an image I can be obtained: " An image with $\mathbf{n}^{(1)}$ $\mathbf{c}^{(1)}$, $\mathbf{n}^{(2)}$ $\mathbf{c}^{(2)}$, ..., $\mathbf{n}^{(B)}$ $\mathbf{c}^{(B)}$. " We will shuffle the order of instances to produce 5 different structured captions for each image. Then we consider the natural text, where the text doesn't adhere to certain templates and is closer to human descriptions, we use two visual question-answering models LLaVA v1.6 Vicuna and LLaVA v1.6 Mistral [36]. For each model, we prompt it with the input image and a prompt, asking the model to describe the image. For each model, we provide 5 different prompts to produce different natural captions. During training, in each iteration, for a given image, we randomly select one caption from those 15 captions to predict scale and shift.

4 Experiments

Datasets. We present our main result on three datasets: NYUv2 [46] and VOID [58] for indoor scenes, and KITTI [15] for outdoor scenes. NYUv2 contains images with a resolution of 480×640 where depth values from 1×10^{-3} to 10 meters. We follow [29, 35, 79] for the dataset partition, which contains 24,231 train images and 654 test images. VOID contains images with a resolution of 480×640 where depth values from 0.2 to 5 meters. It contains 48,248 train images and 800 test images following the official splits [58]. KITTI contains images with a resolution of 352×1216 where depth values from 1×10^{-3} to 80 meters. We adopt the Eigen Split [9] consisting of 23,488 training images and 697 testing images. Following [2, 71], we remove samples without valid ground truth, leaving 652 valid images for testing. We also report zero-shot generalization results on SUN-RGBD [48], which contains 5050 testing images, and DDAD [19], which contains 3950 validation images.

https://doi.org/10.52202/079017-3580

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.

Depth models. For DPT [43], we use DPT-Hybrid fine-tuned for NYUv2 and KITTI respectively, with 123M parameters. We used the one fine-tuned on NYUv2 for VOID. For MiDas [44], we use MiDaS 3.1 Swin2_large-384 with 213M parameters. For DepthAnything [67], we use Depth-Anything-Small with 24.8M parameters. Different from our setting, DepthAnything evaluates using ZoeDepth [3]'s depth decoder that is separately fine-tuned on NYUv2 and KITTI to produce a pixel-wise scale, and MiDas evaluates by aligning prediction with ground truth in scales and shifts. We re-implement several baselines aligning with the setting of predicting a global scale and shift for scaling relative depth, provided in Table 1 and 2.

Hyperparameters. We use the Adam [25] optimizer without weight decay. The learning rate is reduced from 3×10^{-5} to 1×10^{-5} by a cosine learning rate scheduler. The model is trained for 50 epochs under this scheduler. We run our experiment on GeForce RTX 3090 GPUs, with 24GB memory. For reference, if using a single GPU, the training time for RSA with DepthAnything on jointly NYUv2, KITTI, and VOID for 50 epochs takes 57 hours..

Evaluation metrics. We follow [7, 35, 79] to evaluate using mean absolute relative error (Abs Rel), root mean square error (RMSE), absolute error in log space $(\log_{10})$, logarithmic root mean square error ($\text{RMSE}_{\log}$) and threshold accuracy (δ_i).

Quantitative results. We show results on NYUv2 in Table 1, KITTI in Table 2, and VOID in Table 3, where we improve over baselines across all evaluation metrics and approach the performance of using ground truth for scaling. Following DPT, we optimize the scale and shift for the "Global" scaling baseline over the training set and use it for evaluation (i.e., same scale and shift for all test samples). We obtain the "Image" baseline by substituting CLIP text features with CLIP image features. Following [44], we perform a linear regression to find the scale and shift that minimizes the least-square error between the ground truth metric depth and the predicted metric depth. Additionally, we also test median scaling [17], a common practice for evaluation, which uses the ratio between the median of depth prediction and ground truth as the scaling factor. Both linear fitting and median scaling are shown to demonstrate what is achievable if one were to directly fit to ground truth. We train separate RSA models for each dataset, as well as a unified RSA model combining both KITTI and NYUv2, or all KITTI, NYUv2, and VOID.

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.

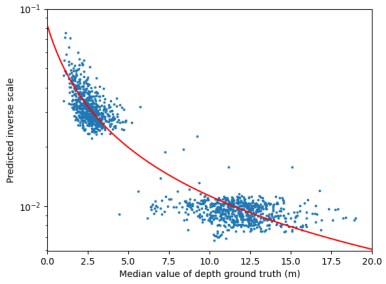

Figure 3: **Left: Predicted inverse scale w.r.t. median depth ground truth.** Larger scenes tend to have larger median ground truth depth. For RSA trained on combined KITTI and NYUv2 with Depth Anything model, we fit an inverse proportional function for the predicted inverse scale in the test set (each point is an image), to verify that the scale is proportional to the median depth, that larger scenes are predicted with larger scales.

Although RSA trained on a single dataset may achieve slightly better performance than the unified model, the difference is minimal. In some metrics, the unified model even outperforms, demonstrating the generalizability of using language as input, given the narrow domain gap for language. Additionally, RSA models consistently outperform "Image" baselines in both in-domain and cross-domain scenarios and are comparable with existing methods under various settings. Our method significantly narrows the performance gap to that of using ground truth for scaling, validating the effectiveness of using language instead of the input image to predict scale.

To examine our scale predictions in more detail, Figure 3 shows a curve fitting plotted for our predicted (inverse) scale against the median of the ground truth. The trend line shows that the scale inferred from text descriptions matches well with median scaling, which is a robust estimator.

Qualitative comparisons. We present representative visual examples comparing RSA with the baseline method on the NYUv2 and KITTI datasets in Figure 4 and Figure 5, respectively, to highlight the benefits of RSA. The error maps illustrate the absolute relative error. Unlike the original DPT, which uses a fixed scale and shift, RSA enhances accuracy uniformly across the image without altering the structure or fine details of the depth map. This improvement is evidenced by the darker areas in the error maps, indicating better scaling and reduced errors.

Zero-shot Generalization. Considering the smaller domain gap in language descriptions across various scenes, we perform a zero-shot transfer experiment to demonstrate RSA's generalization ability. We evaluate the models on the Sun-RGBD [48] and DDAD [19] without fine-tuning. As shown in Table 4 and Table 5, RSA achieves superior results compared to baselines, existing methods, and ground truth scaling. This suggests that language descriptions offer a viable option for relative to metric transfer when generalizing across diverse data domains. Note that a single global scale and shift are ineffective for both indoor and outdoor settings. Therefore, for the "Global" model, we fit it only to NYUv2 to obtain a reasonable global scale and shift for the zero-shot Sun-RGBD evaluation, and fit it to KITTI for DDAD evaluation.

Prompt design for input text. In Table 6, we investigate different designs of RSA text prompts in training and how they affect the performance. Here, to make the experiment more controllable, we use only structured text and only produce one caption for each image to train each model.

https://doi.org/10.52202/079017-3580

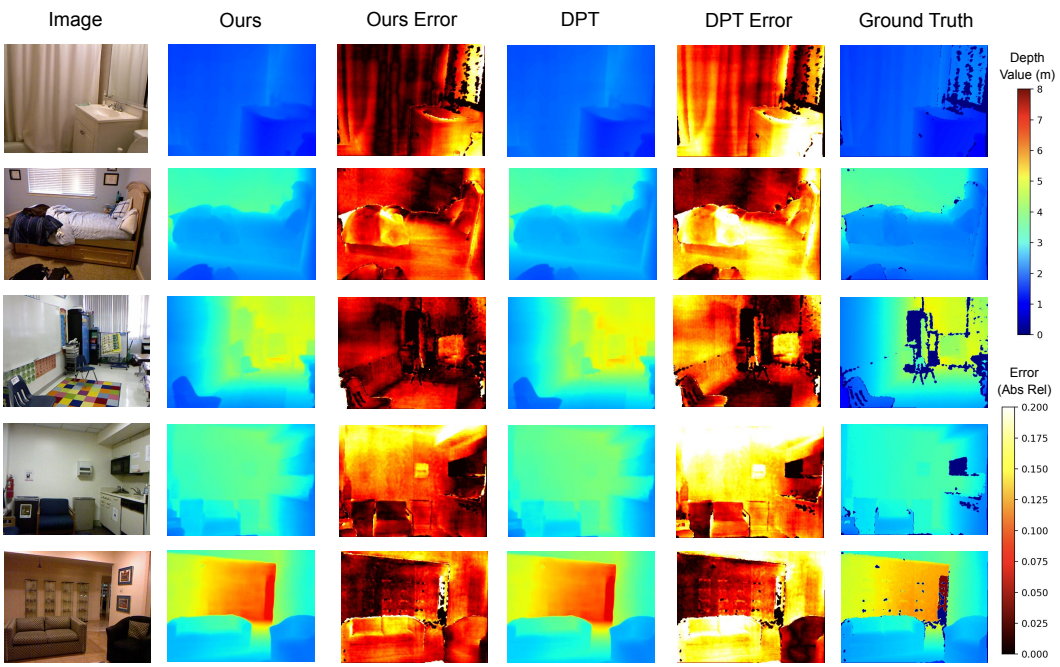

Figure 4: **Visualization of depth estimations on NYUv2.** Building upon DPT, while a better scale factor does not change the structure of the depth prediction, leading to visually similar depth maps, it significantly reduces the overall error (darker in the error map). Note: Zeros in ground truth indicate the absence of valid depth values.

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
DPT	Linear Fit	SUN-RGBD	0.812	0.967	0.993	0.139	0.059	0.412
	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**
MiDas	Linear Fit	SUN-RGBD	0.632	0.912	0.971	0.241	0.102	1.132
	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**
DepthAnything	Linear Fit	SUN-RGBD	0.878	0.979	0.995	0.113	0.054	0.332
	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.

In the 1st and 2nd rows of Table 6, we use an object detector Detic [87] to produce only foreground objects in images and form input text using only foreground objects. We observe that if the input text only specifies the types of objects and their numbers, RSA can still accurately predict the scale for indoor scenes, as these spaces are typically filled with various pieces of furniture. However, this approach performs poorly for outdoor scenes, which tend to be more open and sparse. For instance, parking lots of different sizes may contain varying numbers of cars: a small lot may be crowded, while a large lot may appear empty. In the 3rd and 4th rows of Table 6, we use text formed using segmentation results; we observe that after including background classes, the model works better

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
DPT	Linear Fit	DDAD	0.802	0.954	0.990	0.163	0.254	10.342
	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**
MiDas	Linear Fit	DDAD	0.664	0.912	0.973	0.209	0.301	18.341
	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**
DepthAnything	Linear Fit	DDAD	0.673	0.932	0.983	0.182	0.286	18.423
	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.

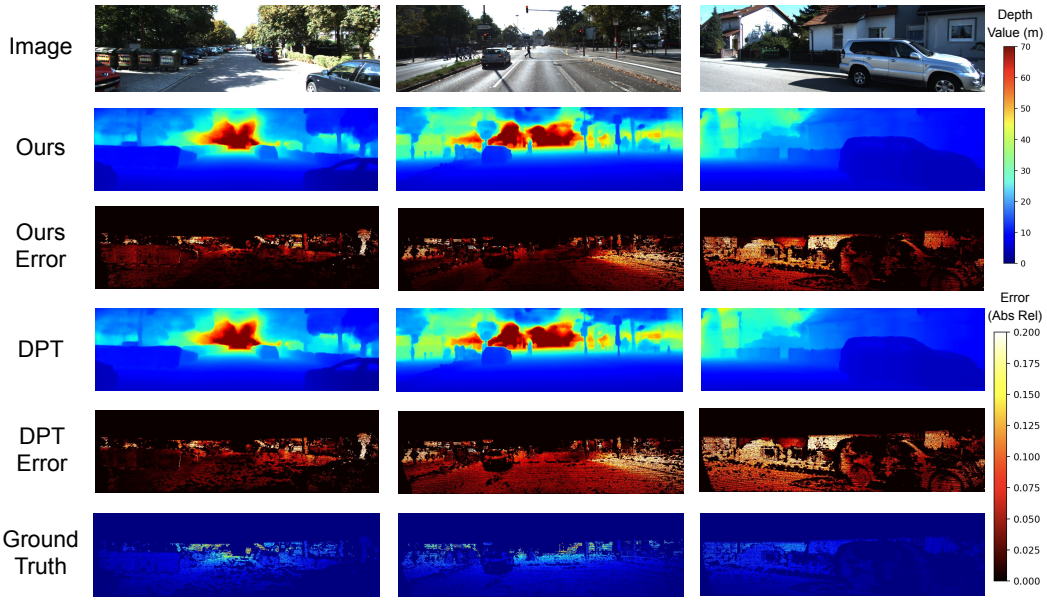

Figure 5: **Visualization of depth estimations on KITTI.** Building upon DPT, while a better scale factor does not change the structure of the depth prediction, it significantly reduces the overall error (darker in the error map). Note: Zeros in ground truth depth indicate the absence of valid depth values.

especially for outdoor scenes, since different backgrounds (like wall, sidewalk, highway, road, sky, house, building) can reflect different scales.

Sensitivity to different text input. To demonstrate the impact of various text inputs on scale prediction, we utilize a trained RSA model based on the DepthAnything model. By altering the text input, we observe changes in scale and shift predictions. We provide the experiment results in Table 7. From top to bottom, we provide text descriptions of scenes from small to large, and our model can properly predict the corresponding scale and shift. This shows promise that our method is able to manipulate the scale of 3D scenes with ease.

https://doi.org/10.52202/079017-3580

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.

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.

5 Discussion

Conclusion. We present the first study exploring whether language, as an additional input modality, can resolve the scale ambiguity in monocular depth estimation, an issue particularly relevant in the context of the recent trend towards large-scale mixed dataset training. We propose a framework, RSA, which learns to convert a scaleless (relative) depth map to metric depth using language descriptions as input. RSA utilizes the pre-trained CLIP encoder, and maps a language description of the image to scale and shift factors that transform the relative depth to metric depth. RSA is validated through extensive experiments on three benchmark datasets and three pre-trained relative depth models, demonstrating significant promise by drastically closing the gap between depth estimation accuracy and its upper bound (that relies on ground truth), validating the hypothesis that language carries valuable scale information that could be used to enhance depth estimation. Moreover, we demonstrate that a unified model trained on both indoor and outdoor datasets with diverse scene compositions generalizes across both scenarios, highlighting the robustness of language information in inferring scale. Finally, a zero-shot transfer experiment shows that the minimal domain gap of language description across scenes further generalizes to unseen data domains, without additional training. The generalizability of RSA stems of our choice of modality, language, which is invariant to nuisance variability that are present in images from lighting conditions and occlusions to specular reflectance and object deformations. Our results demonstrate that RSA is a viable choice to support relative to metric scale alignment for general-purpose monocular depth estimators.

Limitations and future work. We assume that the estimated 3D scene is up to an unknown scale. Although a simple global scale has proven effective, it may not always be sufficiently expressive for converting relative depth to absolute depth, especially when the relative depth is inaccurate. In such cases, global scaling may not adequately recover a high-fidelity metric-scaled depth map due to the presence of outliers. To address this, one may need to refine the relative depth outputted by general-purpose monocular depth estimators. Future research may include extending RSA to handle finer adjustments to also refine depth estimates and investigate the potential of inferring region-wise or even pixel-wise scales using language input. Lastly, while language boasts high ease of use, RSA is also vulnerable to malicious users who may choose captions to steer predictions incorrectly.

Acknowledgements. This work was supported by NSF 2112562 Athena AI Institute, IITP-2021-0-01341, and NRF-RS-2023-00251366.

References

[1] Dylan Auty and Krystian Mikolajczyk. Learning to prompt clip for monocular depth estimation: Exploring the limits of human language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2039–2047, 2023.

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021.

[3] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

[4] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Zhichao Li, Le Zhang, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth learning from video. *International Journal of Computer Vision*, 129(9):2548–2564, 2021.

[5] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Advances in neural information processing systems*, 32, 2019.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[7] Wenjie Chang, Yueyi Zhang, and Zhiwei Xiong. Transformer-based monocular depth estimation with attention supervision. In *32nd British Machine Vision Conference (BMVC 2021)*, 2021.

[8] Alexandra Dana, Nadav Carmel, Amit Shomer, Ofer Manela, and Tomer Peleg. One scalar is all you need–absolute depth estimation using monocular self-supervision. *arXiv preprint arXiv:2303.07662*, 2023.

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

[10] Vadim Ezhov, Hyoungseob Park, Zhaoyang Zhang, Rishi Upadhyay, Howard Zhang, Chethan Chinder Chandrappa, Achuta Kadambi, Yunhao Ba, Julie Dorsey, and Alex Wong. All-day depth completion. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024.

[11] Xiaohan Fei, Alex Wong, and Stefano Soatto. Geo-supervised visual depth prediction. *IEEE Robotics and Automation Letters*, 4(2):1661–1668, 2019.

[12] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018.

[13] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

[14] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 740–756. Springer, 2016.

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[16] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.

 https://doi.org/10.52202/079017-3580

[17] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019.

[18] Michelle R Greene. Statistics of high-level scene context. *Frontiers in psychology*, 4:54269, 2013.

[19] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[20] Vitor Guizilini, Igor Vasiljevic, Rares Ambrus, Greg Shakhnarovich, and Adrien Gaidon. Full surround monodepth from multiple cameras. *IEEE Robotics and Automation Letters*, 7(2):5397–5404, 2022.

[21] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9233–9243, 2023.

[22] Xueting Hu, Ce Zhang, Yi Zhang, Bowen Hai, Ke Yu, and Zhihai He. Learning to adapt clip for few-shot monocular depth estimation. *arXiv preprint arXiv:2311.01034*, 2023.

[23] Pan Ji, Runze Li, Bir Bhanu, and Yi Xu. Monoindoor: Towards good practice of self-supervised monocular depth estimation for indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12787–12796, 2021.

[24] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. *arXiv preprint arXiv:2312.02145*, 2023.

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[26] Dong Lao, Yangchao Wu, Tian Yu Liu, Alex Wong, and Stefano Soatto. Sub-token vit embedding via stochastic resonance transformers. In *International Conference on Machine Learning*. PMLR, 2024.

[27] Dong Lao, Fengyu Yang, Daniel Wang, Hyoungseob Park, Samuel Lu, Alex Wong, and Stefano Soatto. On the viability of monocular depth pre-training for semantic segmentation. In *European Conference on Computer Vision*. Springer, 2024.

[28] Tim Lauer, Tim HW Cornelissen, Dejan Draschkow, Verena Willenbockel, and Melissa L-H Võ. The role of scene summary statistics in object recognition. *Scientific reports*, 8(1):14666, 2018.

[29] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[30] Boying Li, Yuan Huang, Zeyu Liu, Danping Zou, and Wenxian Yu. Structdepth: Leveraging the structural regularities for self-supervised indoor depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12663–12673, 2021.

[31] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023.

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[34] Panfeng Li, Qikai Yang, Xieming Geng, Wenjing Zhou, Zhicheng Ding, and Yi Nian. Exploring diverse methods in visual question answering. In *2024 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI)*, pages 681–685. IEEE, 2024.

[35] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, and Luc Van Gool. Va-depthnet: A variational approach to single image depth prediction. *arXiv preprint arXiv:2302.06556*, 2023.

[36] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[37] Tian Yu Liu, Parth Agrawal, Allison Chen, Byung-Woo Hong, and Alex Wong. Monitored distillation for positive congruent depth completion. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 35–53. Springer, 2022.

[38] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5667–5675, 2018.

[39] 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.

[40] Hyoungseob Park, Anjali Gupta, and Alex Wong. Test-time adaptation for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20519–20529, 2024.

[41] Rui Peng, Ronggang Wang, Yawen Lai, Luyang Tang, and Yangang Cai. Excavating the potential capacity of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15560–15569, 2021.

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[43] 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.

[44] 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.

[45] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. *arXiv preprint arXiv:2112.01518*, 2021.

[46] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012.

[47] Akash Deep Singh, Yunhao Ba, Ankur Sarker, Howard Zhang, Achuta Kadambi, Stefano Soatto, Mani Srivastava, and Alex Wong. Depth estimation from camera image and mmwave radar point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9275–9285, 2023.

[48] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[49] Yiyi Tao, Zhuoyue Wang, Hang Zhang, and Lun Wang. Nevlp: Noise-robust framework for efficient vision-language pre-training. *arXiv preprint arXiv:2409.09582*, 2024.

[50] Jiahang Tu, Hao Fu, Fengyu Yang, Hanbin Zhao, Chao Zhang, and Hui Qian. Texttoucher: Fine-grained text-to-touch generation. *arXiv preprint arXiv:2409.05427*, 2024.

[51] Rishi Upadhyay, Howard Zhang, Yunhao Ba, Ethan Yang, Blake Gella, Sicheng Jiang, Alex Wong, and Achuta Kadambi. Enhancing diffusion models with 3d perspective geometry constraints. *ACM Transactions on Graphics (TOG)*, 42(6):1–15, 2023.

[52] Ruoyu Wang, Zehao Yu, and Shenghua Gao. Planedepth: Self-supervised depth estimation via orthogonal planes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21425–21434, 2023.

 https://doi.org/10.52202/079017-3580

[53] Youhong Wang, Yunji Liang, Hao Xu, Shaohui Jiao, and Hongkai Yu. Sqldepth: Generalizable self-supervised fine-structured monocular depth estimation. *arXiv preprint arXiv:2309.00526*, 2023.

[54] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Yongming Rao, Guan Huang, Jiwen Lu, and Jie Zhou. Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In *Conference on Robot Learning*, pages 539–549. PMLR, 2023.

[55] Alex Wong, Safa Cicek, and Stefano Soatto. Targeted adversarial perturbations for monocular depth prediction. *Advances in neural information processing systems*, 33:8486–8497, 2020.

[56] Alex Wong, Safa Cicek, and Stefano Soatto. Learning topology from synthetic data for unsupervised depth completion. *IEEE Robotics and Automation Letters*, 6(2):1495–1502, 2021.

[57] Alex Wong, Xiaohan Fei, Byung-Woo Hong, and Stefano Soatto. An adaptive framework for learning unsupervised depth completion. *IEEE Robotics and Automation Letters*, 6(2):3120–3127, 2021.

[58] Alex Wong, Xiaohan Fei, Stephanie Tsuei, and Stefano Soatto. Unsupervised depth completion from visual inertial odometry. *IEEE Robotics and Automation Letters*, 5(2):1899–1906, 2020.

[59] Alex Wong and Stefano Soatto. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5644–5653, 2019.

[60] Alex Wong and Stefano Soatto. Unsupervised depth completion with calibrated backprojection layers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12747–12756, 2021.

[61] Cho-Ying Wu, Jialiang Wang, Michael Hall, Ulrich Neumann, and Shuochen Su. Toward practical monocular indoor depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3814–3824, 2022.

[62] Cho-Ying Wu, Jialiang Wang, Michael Hall, Ulrich Neumann, and Shuochen Su. Toward practical monocular indoor depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3814–3824, 2022.

[63] Yangchao Wu, Tian Yu Liu, Hyoungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. Augundo: Scaling up augmentations for monocular depth completion and estimation. In *European Conference on Computer Vision*. Springer, 2024.

[64] Feng Xue, Guirong Zhuo, Ziyuan Huang, Wufei Fu, Zhuoyue Wu, and Marcelo H Ang. Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2330–2337. IEEE, 2020.

[65] Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, and Alex Wong. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[66] Fengyu Yang, Chao Feng, Daniel Wang, Tianye Wang, Ziyao Zeng, Zhiyang Xu, Hyoungseob Park, Pengliang Ji, Hanbin Zhao, Yuanning Li, et al. Neurobind: Towards unified multimodal representations for neural signals. *arXiv preprint arXiv:2407.14020*, 2024.

[67] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024.

[68] Yanchao Yang, Alex Wong, and Stefano Soatto. Dense depth posterior (ddp) from single image and sparse range. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3353–3362, 2019.

[69] Chenyu You, Yifei Mint, Weicheng Dai, Jasjeet S Sekhon, Lawrence Staib, and James S Duncan. Calibrating multi-modal representations: A pursuit of group robustness without annotations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26140–26150. IEEE, 2024.

[70] Zehao Yu, Lei Jin, and Shenghua Gao. P 2 net: Patch-match and plane-regularization for unsupervised indoor depth estimation. In *European Conference on Computer Vision*, pages 206–222. Springer, 2020.

[71] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3916–3925, 2022.

[72] Ziyao Zeng, Daniel Wang, Fengyu Yang, Hyoungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. Wordepth: Variational language prior for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9708–9719, 2024.

[73] Chi Zhang, Wei Yin, Billzb Wang, Gang Yu, Bin Fu, and Chunhua Shen. Hierarchical normalization for robust monocular depth estimation. *Advances in Neural Information Processing Systems*, 35:14128–14139, 2022.

[74] Mingliang Zhang, Xinchen Ye, Xin Fan, and Wei Zhong. Unsupervised depth estimation from monocular videos with hybrid geometric-refined loss and contextual attention. *Neurocomputing*, 379:250–261, 2020.

[75] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

[76] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022.

[77] Renrui Zhang, Longtian Qiu, Wei Zhang, and Ziyao Zeng. Vt-clip: Enhancing vision-language models with visual-guided texts. *arXiv preprint arXiv:2112.02399*, 2021.

[78] Renrui Zhang, Ziyao Zeng, Ziyu Guo, Xinben Gao, Kexue Fu, and Jianbo Shi. Dspoint: Dual-scale point cloud recognition with high-frequency fusion. *arXiv preprint arXiv:2111.10332*, 2021.

[79] Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. Can language understand depth? In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6868–6874, 2022.

[80] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *2022 International Conference on 3D Vision (3DV)*, pages 668–678. IEEE, 2022.

[81] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. Towards better generalization: Joint depth-pose learning without posenet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9151–9161, 2020.

[82] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *arXiv preprint arXiv:2303.02153*, 2023.

[83] Chong Zhou, Chen Change Loy, and Bo Dai. Denseclip: Extract free dense labels from clip. *arXiv preprint arXiv:2112.01071*, 2021.

[84] Junsheng Zhou, Yuwang Wang, Kaihuai Qin, and Wenjun Zeng. Moving indoor: Unsupervised video depth learning in challenging environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8618–8627, 2019.

[85] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021.

[86] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017.

[87] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022.

[88] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2639–2650, 2023.

 https://doi.org/10.52202/079017-3580

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim in the abstract and introduction that we proposed a novel formulation of the relative to metric depth transfer, and verify the feasibility of inferring scale from language and as a general alignment module.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We create a separate section to discuss the potential limitations, including lack of exploring human-generated descriptions, lack of certain factors in current descriptions, and may not be expressive enough especially when the relative depth is inaccurate.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We don't have theoretical result in this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have provided enough details about model architectures, datasets, depth model setup, evaluation metrics, and hyperparameters to reproduce all necessary experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

https://doi.org/10.52202/079017-3580

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release our code and checkpoints, including training and evaluation scripts. All datasets used in the paper are open-sourced.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have provided enough details about training and testing, including data splits, hyperparameters, type of optimizer, ablation study, and analysis to understand the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported because it would be too computationally expensive.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have stated sufficient information for computer resources in the hyperparameter section, including the type of compute worker, memory, and time of execution.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have carefully reviewed the NeurIPS Code of Ethics and will obey it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed that our method could potentially enhance safety in autonomous systems and benefit augmented reality and virtual reality, but might lead to potential job losses in sectors like transportation.

    Guidelines:

https://doi.org/10.52202/079017-3580

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: As a model for Monocular Depth Estimation, our model doesn't have a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have properly cited all existing assets used in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release and document our code well.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper doesn't involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper doesn't involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

https://doi.org/10.52202/079017-3580

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.