WorDepth: Variational Language Prior for Monocular Depth Estimation

Ziyao Zeng[1]   Daniel Wang[1]   Fengyu Yang[1]   Hyoungseob Park[1]
Stefano Soatto[2]   Dong Lao[2]   Alex Wong[1]

[1]Yale University   [2]University of California, Los Angeles

[1]{ziyao.zeng, daniel.wang.dhw33, fengyu.yang, hyoungseob.park, alex.wong}@yale.edu
[2]{soatto,lao}@cs.ucla.edu *

Abstract

Three-dimensional (3D) reconstruction from a single image is an ill-posed problem with inherent ambiguities, i.e. scale. Predicting a 3D scene from text description(s) is similarly ill-posed, i.e. spatial arrangements of objects described. We investigate the question of whether two inherently ambiguous modalities can be used in conjunction to produce metric-scaled reconstructions. To test this, we focus on monocular depth estimation, the problem of predicting a dense depth map from a single image, but with an additional text caption describing the scene. To this end, we begin by encoding the text caption as a mean and standard deviation; using a variational framework, we learn the distribution of the plausible metric reconstructions of 3D scenes corresponding to the text captions as a prior. To "select" a specific reconstruction or depth map, we encode the given image through a conditional sampler that samples from the latent space of the variational text encoder, which is then decoded to the output depth map. Our approach is trained alternatingly between the text and image branches: in one optimization step, we predict the mean and standard deviation from the text description and sample from a standard Gaussian, and in the other, we sample using a (image) conditional sampler. Once trained, we directly predict depth from the encoded text using the conditional sampler. We demonstrate our approach on indoor (NYUv2) and outdoor (KITTI) scenarios, where we show that language can consistently improve performance in both. Code: https://github.com/Adonis-galaxy/WorDepth.

1. Introduction

The process of imaging is a surjection from a 3D scene to the 2D image domain, where infinitely many 3D scenes can map to the same image. Its inverse problem, estimating
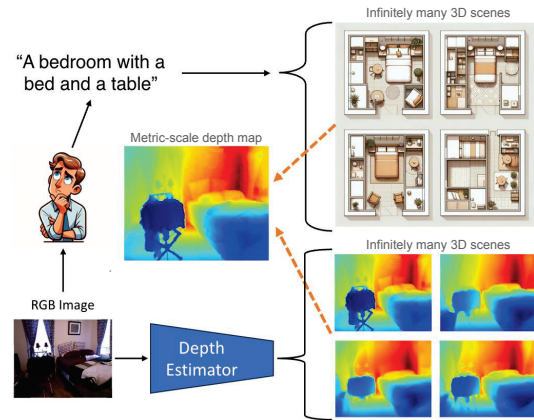

Figure 1. **Language as a prior for depth estimation**. Depth estimation from a single image is an ill-posed problem (i.e., scale), and likewise from text captions (i.e., layout). Can two inherently ambiguous modalities resolve metric-scaled depth estimates?

the 3D scene structure from a single image, i.e., monocular depth estimation, is therefore ill-posed with inherent ambiguity, such as the scale of the reconstruction. Consequently, induction is necessary, and depth estimation becomes drawing a scene with maximum likelihood from the distribution of all possible scenes, conditioned on the image. This conditional scene distribution is learned by a deep neural network on a chosen training set. While an ideal training set should accurately reflect this distribution, practical challenges arise due to the scarcity of well-established large-scale depth datasets. A crucial question arises: Can any priors, other than the training set, be leveraged to calibrate the learned scene distribution to true real-world statistics?

These priors may come in many forms, from generic priors such as local smoothness and connectivity [19, 22, 67, 102] or object orientation [15] that may be imposed as a part of the training objective (regularizer) to specific inductive biases realized as architectural designs (layers) [65] or a collection object shapes [14]. While generic priors are

---

*Due to an oversight, the author list in the published camera ready does not match that of this manuscript. The above is the correct author list.

suitable for a wide variety of scenes, they typically lack specificity, i.e., size or shape of objects within a specific 3D scene. On the other hand, specific network designs may backfire when the assumption motivating the design does not hold, i.e., using specifics about camera parameters for reconstruction. We consider a more flexible source of priors – language – that is closely tied to semantics, and often shape (and functionality) [4, 31, 32]. Consider a text description of "A bedroom with a bed and a table" as in Fig. 1: One can imagine a probable 3D scene containing a bed and a table as the primary objects. In fact, there exist infinitely many 3D scenes compatible with the description, as there are ambiguities in terms of the scene layout and the precise shape of the bed and table. Yet, one may surmise that the scale of the scene is closely related to the objects (and their typical sizes) populating it. This lends to a prior that is specific for a given scene, yet, generic enough without assumptions on the camera used or the shapes within the imaged 3D scene.

Hence, the question at hand becomes whether two inherently ambiguous modalities (camera image and text descriptions) can be exploited for their complementary strengths: In the image, one can observe the layout and object shapes populating the 3D scene; in a text caption, one has strong priors about the scale (and coarse shapes) of the scene. Our work aims to resolve the respective ambiguities of the two modalities by using language to reduce the solution space to yield metric-scaled reconstructions as 2.5D depth maps.

To test the feasibility of this approach, we consider the ill-posed inverse problem of monocular depth estimation, where one predicts a depth map from a single image. Instead of using just an image, we also assume a text description or caption describing the 3D scene captured within the image. Note that we do not make any assumption regarding the source of the description, i.e., it can be dictated by humans or generated by a model. But for practicality, we use an image captioner (ExpansionNet v2 [25]) to generate a brief, concise description of the image.

To exploit the inherent ambiguity of text captions, where a single description can generate infinitely many 3D scenes, we choose to encode the caption using a variational autoencoder (VAE) as a mean and standard deviation of the plausible scene layout distribution. By sampling a noise vector from a standard Gaussian and using the reparameterization trick customary in VAEs, we can draw from the latent distribution and decode it into a metric-scaled depth map. Yet, to choose a particular depth map amongst the many possible, one must rely on the image. This is facilitated by a conditional sampler that predicts the noise vector from the given image in place of the one sampled from a Gaussian to be used in the reparameterization step. Consequently, this substitution enables one to sample the most probable depth map, adhering to the scene arrangement and

object shapes observed in the image, from the learned distribution. This naturally lends to an alternating optimization process between the (text-)VAE and conditional sampler.

In one alternation, one would predict the mean and standard deviation from the text caption and optimize the text-VAE branch for depth by minimizing a loss with respect to ground truth on the depth map sampled using a standard Gaussian (similar to traditional VAEs). In the other alternation, one would still use the mean and standard deviation predicted by the text-VAE, but instead, use the conditional sampler to "select" a specific depth map compatible with the image, and again, minimize a loss on the output depth. Note: that depending on the alternation, either the text-VAE or the conditional sampler is frozen. At test-time, one no longer needs to sample from the Gaussian and may directly predict depth using the text-VAE with the conditional sampler (see Fig. 2). In another mode, one may use the text-VAE alone to generate plausible scenes for a given caption.

Our contributions are as follows: (i) We propose a variational framework that leverages complementary strengths of two inherently ambiguous modalities for monocular depth estimation; we term our approach, WorDepth. (ii) We introduce an image-based conditional sampler that models the use of language as a conditional prior. (iii) We achieve the state-of-the-art on indoor (NYU Depth V2 [58]) and outdoor (KITTI [20]) benchmarks. (iv) To the best of our knowledge, we are the first to treat language as a variational prior for monocular depth estimation.

2. Related Work

Monocular depth estimation trains by minimizing loss between depth predictions and ground-truth depth maps [2, 7, 17, 35, 46, 52, 54, 61, 66, 78, 80, 84, 86]. Specifically, DORN [16] employs a spacing-increasing discretization strategy for depth estimation as an ordinal regression problem. AdaBins [2] introduces a transformer block that segments the depth range into adaptive bins. ASTransformer [7] incorporates an Attention-based Up-sample Block to enhance detailed texture features. DepthFormer [40] employs hierarchical aggregation and heterogeneous interaction modules for effective feature affinity and modeling. RPSF [47] presents a differentiable model of the aperture mask. However, deriving semantics solely from visual cues is challenging because of scale ambiguity and the limited size of fully annotated training datasets. We use language as a prior to ground predictions to metric scale. When ground-truth depth is not available, self-supervised approaches [3, 15, 27, 36, 51, 62–64, 70, 85, 94, 96, 100] rely on geometric constraints, often established via from various modalities, including lidar [44, 50, 67–69, 72, 79] and radar [59], or through deliberate design. Arising from training, if done at a large scale, is a prior on the scene that can be exploited for semantic tasks [33]. On the other hand,

we consider language as a semantic prior to enhance the effectiveness of monocular depth estimation.

Variational and generative methods focus on the ambiguous nature of monocular depth estimation, many involving Diffusion or VAE models for modeling this ambiguity [5, 10, 41, 56, 57, 73, 83]. DepthGen [56] uses a depth pre-trained diffusion model, which generates depth estimations conditioned on images, and shows that the model is capable of generating multiple plausible depth maps when depth is ambiguous. DDVM [57] uses a similar approach and designed a training pipeline that can produce both depth maps and optical flow outputs with a diffusion model. [73] trained a VAE model that outputs a probability distribution over scene depth given an image, which can then be combined with additional inputs for more accurate depth estimations. VDN [10] models depth as a distribution with its variance interpreted as uncertainty. The CodeSLAM model [5] also employed a VAE conditioned on image intensities for depth estimation. However, although these work explored the idea of uncertainty in depth estimation, and even combined other modalities of inputs [73], none have experimented with language priors, and most VAE-based approaches use images to obtain the mean of the modeled distribution, which is fundamentally different from WorDepth.

Foundation models [6, 21, 23, 37, 38, 48, 49, 53, 77, 98, 104] acquire a comprehensive understanding of languages, images, and other data types through pre-training under substantial and diverse datasets, thus forming an effective baseline for downstream tasks [2, 8, 12, 39, 42, 71, 74–76, 81, 82, 89, 92, 93, 95]. To leverage foundation models for monocular depth estimation, TADP [30] uses captions created by AI to enhance the correlation between text and images in diffusion-based vision models. VPD [97] leverages a diffusion-based pipeline with cross-attention between text and images. Dinov2 [48] trains a ViT [11] with 1B parameters using an automatically built image dataset under contrastive learning objectives. Unlike methods that rely on foundation models for feature extraction, WorDepth is potentially more efficient for industrial applications.

Vision-language models are designed to build connections between visual and language inputs. CLIP [53] conducts contrastive learning between text-image pairs, empowering various tasks like few-shot image classification [18, 87, 88, 101], image segmentation [55, 99], object detection [55, 103], and 3D perception [26, 90, 91, 105]. In light of the powerful emerging ability brought by recent vision-language models, some works have tried to apply the vision-language model for monocular depth estimation. DepthCLIP [91] leverages the semantic depth response of CLIP [53] with a depth projection scheme to conduct zero-shot adaptation from the semantic language response to monocular depth estimation. Furthermore, [26] extends DepthCLIP with learnable prompts and depth codebook to

narrow the depth domain gap among different scenes. Likewise, [1] modifies DepthCLIP [91] using continuous learnable tokens in place of discrete human-language words. Additionally, VPD [97] exploits the high-fidelity embedding of a pre-trained text-to-image diffusion model in monocular depth estimation. However, existing methods using vision-language models rely on implicit modeling. Conversely, WorDepth explicitly models language as a prior for depth estimation and exploits strong priors regarding the size of objects described in text captions to better ground monocular depth (often scaleless) to metric scale.

3. Method

Given an RGB image $x : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^3$, monocular depth estimation aims to infer a dense depth map $y : \Omega \subset \mathbb{R}^2 \to \mathbb{R}_+$ using a parameterized function h realized as a neural network, i.e., $y := h(\cdot)$. We consider a supervised dataset $\mathcal{D} = \{x^{(m)}, \mathbf{t}^{(m)}, y^{*(m)}\}_{m=1}^M$ with M samples, where $y^* : \Omega \subset \mathbb{R}^2 \to \mathbb{R}_+$ denotes the ground-truth depth map, and $\mathbf{t}$ the text caption describing the image.

3.1. Text variational auto-encoder

To incorporate language priors to monocular depth estimation, we first design a variational auto-encoder (VAE) to learn the latent distribution of possible depth maps as described by the text caption. This VAE is comprised of the text encoder from a pre-trained vision-language model, CLIP [53], which by default offers a shared latent space between vision and text embeddings, followed by a multi-layer perceptron (MLP) to estimate the mean $\hat{\mu} \in \mathbb{R}^d$ and standard deviation $\hat{\sigma} \in \mathbb{R}^d$ of the latent distribution of plausible scenes based on the text encoding. Note that the CLIP text encoder is frozen at all times and never updated when training WorDepth. Specifically, given a text caption $\mathbf{t} = \{t_1, t_2, ...\}$, we first encode it using the CLIP text encoder and estimate the mean and standard deviation as $(\hat{\mu}, \hat{\sigma}) = g_\psi(\mathbf{t}) \in \mathbb{R}^{2 \times d}$ using a multi-layer perceptron (MLP). To sample from the distribution parameterized by $\hat{\mu}$ and $\hat{\sigma}$, we first draw a noise vector $\epsilon \in \mathbb{R}^d$ from a standard Gaussian $\epsilon \sim \mathcal{N}(0, 1)$. Then, we use ϵ to sample from the latent distribution via the reparameterization trick [29], $\hat{z} = \hat{\mu} + \epsilon \cdot \hat{\sigma}$. We refer to this module as a text variational auto-encoder (text-VAE). To generate a depth map $\hat{y}$ from the sample $\hat{z} \in \mathbb{R}^d$, we first duplicate $\hat{z}$ along the horizontal and vertical axes to yield a $d \times h \times w$ latent (choice of design to be discussed below in Sec. 3.2) and feed it through a depth decoder to yield $\hat{y} = h_\phi(\hat{z}) \in \mathbb{R}_+^{H \times W}$, where we overload $\hat{z}$ as the spatially duplicated latent, and H and W denote the height and width of the depth map, preset as hyperparameters to match the desired image dimensions.

To train our text-VAE and depth decoder, we minimize

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{SI}}(y^*, \hat{y}) + \alpha \cdot \mathcal{L}_{\text{KL}}(\hat{\mu}, \hat{\sigma}) \tag{1}$$
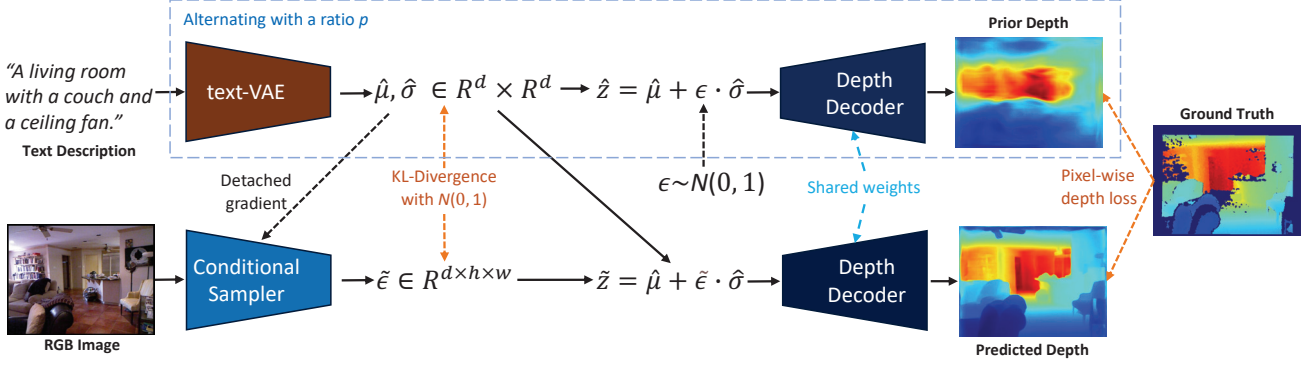
Figure 2. **Training WorDepth.** We begin with optimizing text-VAE by predicting the mean and standard deviation of the latent distribution of depth maps corresponding to a text caption. We then sample $\hat{z}$ from the distribution using the reparameterization trick with $\epsilon \sim \mathcal{N}(0, 1)$ and decode it into a depth map for loss computation. We then optimize a conditional sampler by predicting patch-wise $\tilde{\epsilon}$ from an image to sample $\tilde{z}$ from the latent to yield output depth for the loss computation. The depth decoder is updated in both alternating steps.

with respect to ψ and ϕ, where $\mathcal{L}_{\text{SI}}$ is the scale invariant loss (Eq. (3)), $\mathcal{L}_{\text{KL}}$ the KL divergence loss (Eq. (4)) as detailed in Section 3.3, and α the weight of the KL divergence term.

3.2. Image-based conditional sampler

While our text-VAE can predict plausible metric-scaled depth maps from text captions, we are interested in the depth map corresponding to a specific image. To do so, we treat text-VAE as the latent prior distribution of the plausible scene layouts. Predicting depth $\tilde{y}$ for a specific image x requires *sampling* the latent corresponding to the depth map of the 3D scene layout with the highest likelihood to be compatible with the observed image, i.e., prior conditioned on the image. To this end, we introduce an image-based conditional sampler that will predict the sample $\tilde{\epsilon}$ in place of $\epsilon \sim \mathcal{N}(0, 1)$ drawn from the standard Gaussian. Using the reparameterization trick as before, we will use $\tilde{\epsilon}$ to select the latent vector $\tilde{z}$ to be decoded by the depth decoder.

Specifically, our image-based conditional sampler utilizes a Swin-L transformer backbone to encode an image $x \in \mathbb{R}^{3 \times H \times W}$. We chose this design to exploit the locality of the tokens produced by Swin-L. The tokens are then encoded into $h \times w$ number of local samples $\tilde{\epsilon} \in \mathbb{R}^{d \times h \times w}$ to be used to sample from the latent distribution of our text-VAE; in other words, we perform "patch-wise" selection from latent distribution for more granular predictions. To do so, we additionally include $\hat{\mu}$ and $\hat{\sigma}$ as part of its input. We note that $\hat{\mu}$ and $\hat{\sigma}$ have been detached from the computational graph and treated as input. We refer to this module as our conditional sampler $\tilde{\epsilon} = f_{\varphi}(x, \hat{\mu}, \hat{\sigma})$, which aims to estimate the most probable latent variable of text-VAE. Thus, the scene layout latent vector is now given by $\tilde{z} = \hat{\mu} + \tilde{\epsilon} \cdot \hat{\sigma}$, and the predicted depth $\tilde{y} = h_{\phi}(\tilde{z})$. As an implementation detail, we note that skip connections from the encoder f_{φ} are injected into h_{ϕ} by concatenation; when training text-

VAE (Sec. 3.1), feature maps of skip connections are of the same size, but populated with zeros instead.

To train the conditional sampler, we minimize the same loss (Eq. (1)) as that of text-VAE:

$$\mathcal{L}_{\text{CS}} = \mathcal{L}_{\text{SI}}(y^*, \tilde{y}) + \beta \cdot \mathcal{L}_{\text{KL}}(\tilde{\mu}, \tilde{\sigma}) \qquad (2)$$

with respect to φ and ϕ. With a batch size of b, the number of $\tilde{\epsilon}$ is $b \times h \times w$, while $\tilde{\mu}$ and $\tilde{\sigma}$ are the sample mean and standard deviation of $\tilde{\epsilon}$ over a batch. We impose a KL divergence loss as regularization so that the estimated $\tilde{\epsilon}$ does not drift from the standard Gaussian, which also serves to improve training stability.

3.3. Training Loss

Scale invariant loss. We minimize a supervised loss using ground truth y^*. To improve training stability over diverse scenes, we use the scale-invariant depth loss [13]:

$$\mathcal{L}_{\text{SI}}(y, y^*) = \frac{1}{N_e} \sum_{(i,j) \in \Omega} e(i,j)^2 - \frac{\gamma}{N_e^2} \big(\sum_{(i,j) \in \Omega} e(i,j) \big)^2,$$
$$(3)$$

where $e(i, j) = \log y(i, j) - \log y^*(i, j)$, Ω denotes the image space, N_e the number of pixels, y the predicted depth, and γ the scaling factor to control the sensitivity of the loss.

Kullback-Leibler (KL) divergence loss. Following [29], we employ the KL Divergence loss as a regularizer, which biases the predicted latent distribution (parameterized by mean μ and standard deviation σ) towards a standard Gaussian distribution. We apply the Kullback-Leibler divergence loss to μ and σ as follows:

$$\mathcal{L}_{\text{KL}}(\mu, \sigma) = -\log(\sigma) + \frac{\sigma^2 + \mu^2}{2} - \frac{1}{2}. \qquad (4)$$

3.4. Optimizing Wordepth

Training Wordepth involves optimizing text-VAE with our conditional sampler: One may choose to first train

4

Image | Text | Ours | Ours Error | Adabins | Adabins Error | Ground Truth

A store with chairs and tables in it.

An unmade bed in a bedroom with a window.

A bathroom with a toilet, a sink and a shower curtain.

A man standing in a kitchen next to a counter.

A classroom with a group of desks.

Depth Value (m)
8
7
6
5
4
3
2
1
0

Error (Abs Rel)
0.200
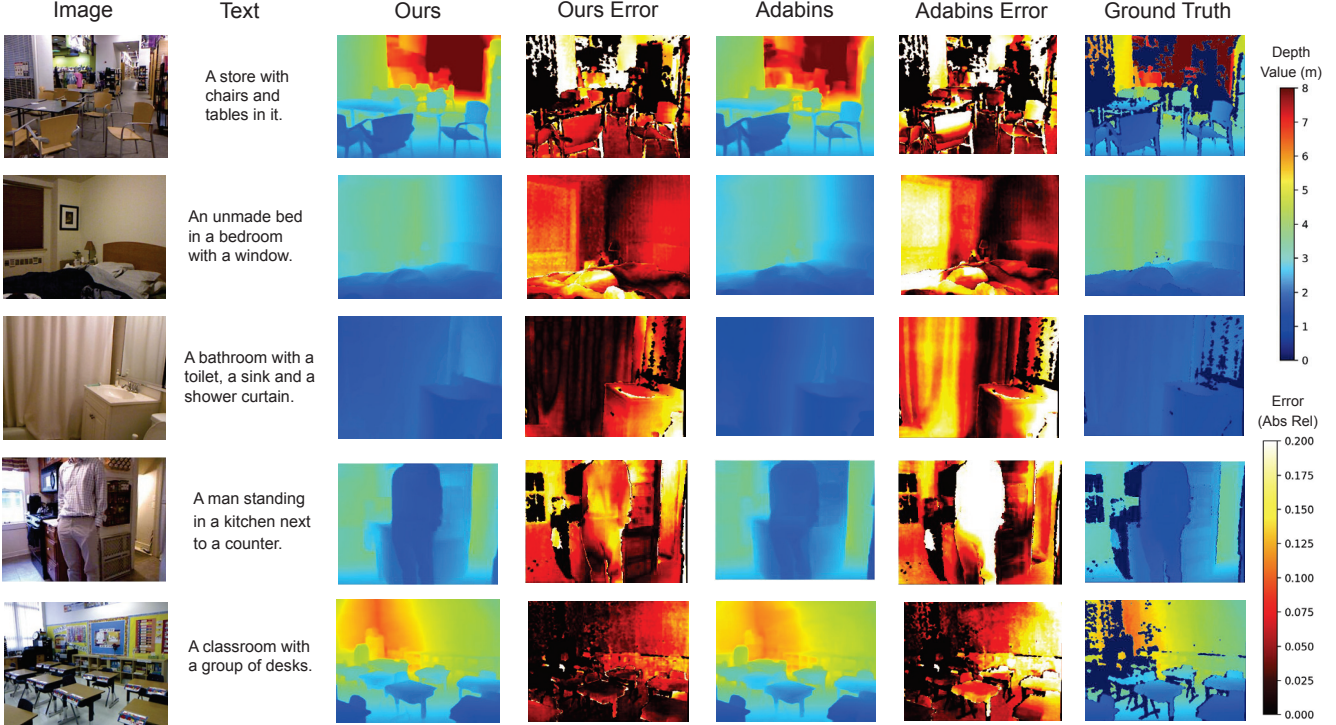0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000

Figure 3. **Qualitative results on NYU Depth V2.** We compare WorDepth with AdaBins [2]. Text descriptions are generated using ExpansionNet v2 [25]. Overall, WorDepth improves uniformly across the image (darker in error map), implying better scale. WorDepth also predicts more accurate depth in regions corresponding to "chairs", "window", "shower curtain", "man", and "desks", which are objects specified by text descriptions. Note: Zeros in the ground truth depth map indicate the absence of valid depth values.

text-VAE until convergence (i.e., optimize for ψ^*, ϕ^*), then freeze ψ^*, ϕ^*, and finally train the image-based conditional sample (i.e., optimize for φ^*). However, we find that doing so often results in the conditional sampler being trapped in a suboptimal local minimum. Moreover, this introduces the inconvenience of an extra stage of training. Instead, we propose an alternating optimization scheme to train the text-VAE with conditional sampler. In one alternating step, we freeze the conditional sampler and train the text-VAE and depth decoder following the procedure in Sec. 3.1, i.e., predicting $\hat{\mu}$ and $\hat{\sigma}$ from text caption **t** and using the reparameterization trick with an ϵ drawn from a standard Gaussian to sample the latent vector. In the next alternating step, we freeze text-VAE and train the conditional sampler with the depth decoder following Sec. 3.2, i.e., predicting $\hat{\mu}$ and $\hat{\sigma}$ using the frozen text-VAE and sample from the latent distribution using $\tilde{\epsilon}$ predicted from the image. These alternating steps are repeated with a ratio of p (for optimizing text-VAE) to $1 - p$ (for optimizing the conditional sampler).

Inference. Once trained, we no longer require drawing ϵ from a standard Gaussian. Instead, at test time, the inference step simply follows Sec. 3.2. In another mode, if one wants to generate depth maps from text captions, one can discard the conditional sampler branch and directly sample from a standard Gaussian instead.

4. Experiments

Datasets. We evaluate our method on indoor (NYU Depth V2 [58]) and outdoor (KITTI [20]) scenarios. NYU Depth V2 contains 480×640 images with depth values from 1×10^{-3} to 10 meters. We follow [34] for the dataset partition, which contains 24,231 train images and 654 test images. KITTI contains 352×1216 images where depth values from 1×10^{-3} to 80 meters. We adopt the Eigen Split [13] consisting of 23,488 training images and 697 testing images. Following [2, 86], after cleaning out samples without valid ground truth, we have 652 valid images for testing.

Network Architecture. We use the ResNet-50 [24] version of CLIP [53] text encoder to extract text features. We use ExpansionNet-v2 [25] for captioning images for efficiency. We set the dimension d of the latent space of the text-VAE and image-based conditional sampler to be 128. As for the image-based conditional sampler, we use a Swin-L Transformer backbone [45] pre-trained on ImageNet [9]. For the text-VAE, given CLIP features of size 1024, we use a 3-layer MLP with hidden dimensions of 512, 256, and 128 to encode text features. For the depth decoder, there are 3 convolutional up-sampling and refinement layers. For depth

Method	Backbone	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	Abs Rel $\downarrow$	$\log_{10} \downarrow$	RMSE $\downarrow$
DepthCLIP [91]	CLIP (zero-shot)	0.394	0.683	0.851	0.388	0.156	1.167
CLIPMDE [1]	CLIP	0.465	0.776	0.922	0.319	0.139	0.970
GeoNet [52]	ResNet-50	0.834	0.960	0.990	0.128	0.057	0.569
DORN [16]	ResNet-101	0.828	0.965	0.992	0.115	0.051	0.509
Yin et al. [80]	ResNeXt-101	0.875	0.976	0.994	0.108	0.048	0.416
TransDepth [78]	ViT-B	0.900	0.983	0.996	0.106	0.045	0.365
ASN [46]	HRNet-48	0.890	0.982	0.996	0.101	0.044	0.377
Big to Small [35]	DenseNet-161	0.885	0.978	0.994	0.110	0.047	0.392
DPT-Hybird [54]	ViT-B	0.904	0.988	**0.998**	0.110	0.045	0.357
ASTransformer [7]	ViT-B	0.902	0.985	0.997	0.103	0.044	0.374
AdaBins [2]	EffNet-B5 + ViT-mini	0.903	0.984	0.997	0.103	0.044	0.364
NeWCRFs [86]	Swin-L	0.922	**0.992**	**0.998**	0.095	0.041	0.331
Yu et al. [84]	Swin-L	0.921	0.990	**0.998**	0.093	0.040	0.331
DepthFormer [40]	Swin-L	0.923	0.989	0.997	0.094	0.040	0.329
Baseline	Swin-L	0.910	0.990	**0.998**	0.098	0.043	0.351
WorDepth	Swin-L	**0.932**	**0.992**	**0.998**	**0.088**	**0.038**	**0.317**
%Improvement	-	+2.42%	+0.02%	+0.00%	-10.20%	-11.63%	-9.69%

Table 1. **Quantitative results on NYU Depth V2.** The baseline method is to directly train a Swin-L image encoder and the depth decoder without the help of language prior. Improvement refers to the performance enhancement relative to the Baseline.

prediction, we attach 3 skip connections from the conditional sampler to the depth decoder between corresponding layers. When optimizing for text-VAE by our alternating optimization scheme (Sec. 3.4), we sample $\epsilon \sim \mathcal{N}(0, 1)$ from a standard Gaussian; as an implementation detail, all values passed from the skip connections are set to be zero.

Hyperparameters. We use the Adam [28] 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 for both KITTI and NYU Depth V2 under this scheduler. γ for scale-invariant loss is set to 0.85, and the weights α and β for KL-Divergence are set to 1×10^{-3}. We set the probability p to optimizing text-VAE branch to 1%. Data augmentation includes random gamma within $[0.9, 1.1]$, random brightness within $[0.75, 1.25]$ for NYU Depth V2 [58] and $[0.9, 1.1]$ for KITTI [20], random color intensity within $[0.9, 1.1]$ for each channel, random horizontal flipping with 50% probability, and random rotations within $[-2.5, 2.5]$ degrees.

Evaluation metrics. Following [7, 43], we evaluate WorDepth and baseline methods quantitatively using mean absolute relative error (Abs Rel), root mean square error (RMSE), absolute error in log space ($\log_{10}$), logarithmic root mean square error (RMSE$_{\log}$) and threshold accuracy (δ_i). The evaluation metrics are summarized in the Supp. Mat. For qualitative results and comparisons, see Fig. 3 and 4, where the error map shows the absolute relative error.

Quantitative results. We show results on NYU Depth V2 in Tab. 1, where we improve over the baseline and existing works across all evaluation metrics. We want to highlight that WorDepth significantly excels in terms of the threshold accuracy $\delta < 1.25$, which measures the proportion of predictions deviating from the ground truth within a

specific range. We note that while existing methods often produce high fidelity shapes (i.e., ordinal relationships of points) in depth maps, the scale tends to be off – leading to lower $\delta < 1.25$. Our gain in the $\delta < 1.25$ metric indicates that a greater proportion of depth estimations align closely with the ground truth, thanks to better scale estimated based on objects that populate the scene, thereby yielding depth values in ranges closer to that of ground truth.

Tab. 2 shows the results on the KITTI dataset, using the Eigen Split [13] partition. WorDepth also achieves state-of-the-art performance. Like NYU Depth V2, WorDepth improves the threshold accuracy $\delta < 1.25$, however, the relative performance gain on this metric is not as pronounced as on NYU Depth V2. This difference can be due to the wider range of object sizes and shapes that may populate an outdoor scene that are attributed to the same equivalence class of an object. For example, the term "car" may refer to a sedan, a coupe, or a hatchback – all exhibit different sizes (coupes are smaller than sedans) and shapes (hatchbacks have an elevated and connect trunk). While text captions give flexibility between generality and specificity as a prior, in cases where captions tend to be vague, the explicit reliance (by modeling as a conditional prior) on them may backfire, leading to incorrect shapes and sizes. Nonetheless, conditioning on the image resolve such cases to a degree and usage of the prior leads to more benefits than harm.

Qualitative comparisons. We show representative visual examples comparing WorDepth with a baseline method AdaBins [2] on the NYU Depth V2 dataset in Fig. 3, to highlight the benefit of the language prior.

From the error map where brighter regions indicate larger errors, it is evident that WorDepth predicts more accurate depth for objects mentioned in the text description,
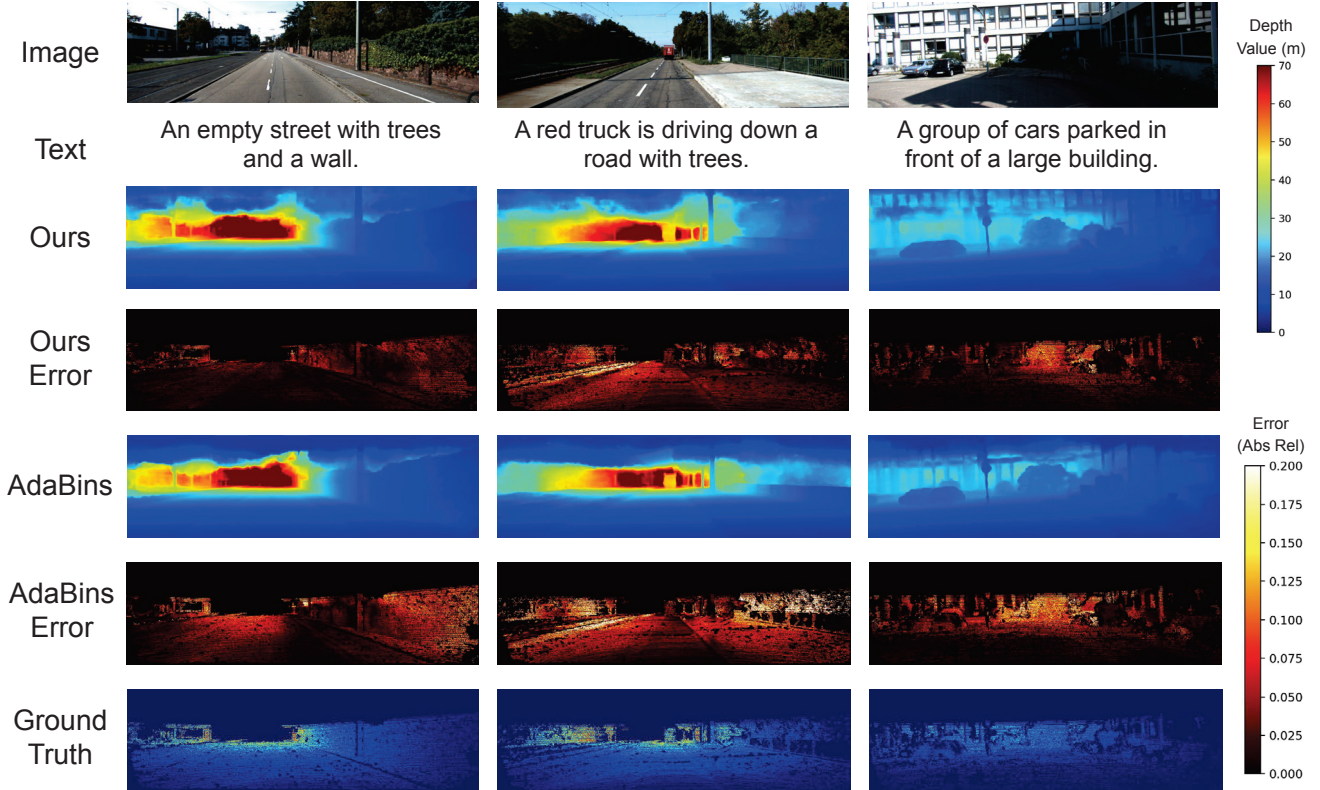
Figure 4. **Visualization of depth estimations on KITTI.** Compared with AdaBins [2], WorDepth improves uniformly across the image (darker in error map), implying better scale. WorDepth also predicts more accurate depth in regions corresponding to "wall", "trees", "building", which are objects specified by text descriptions. Note: Zeros in ground truth depth indicate the absence of valid depth values.

like "chairs and tables" in the first row, "a window" in the second row, "a shower curtain" in the third row, "a man" in the fourth row, and "a group of desks" in the last row. Note that errors maps of WorDepth shows improvement uniformly across the image regions; this implies that our method estimates a better scale than existing ones, thanks to priors about object size (and coarse shapes) from text captions. Knowing that a certain object exists within an image reduces the problem to "placing" the object in the 3D scene based on its shape and location in the image. We showed that scale can be inferred from language, which can narrow down the solution space of depth prediction, leading to improved accuracy.

A similar pattern is also evident in KITTI (Fig. 4). Examples include improved accuracy for "a wall" shown in the first column, "trees" in the second column, and "a group of cars" alongside "a large building" in the last column. This observation is intriguing because, for example, the text "a wall" is ambiguous by itself, especially in outdoor scenes, where the wall could be any size or distance away from the camera, 1 or 100 meters. However, the text description of a scene, either from a human annotator or a deep neural network, inherently carries biases that emphasize "a wall" when its size (tall or wide enough) or depth falls within a

specific range while ignoring it when it falls within another range. The resulting prior embedded in the text description may convey more scale information than initially apparent.

Optimizing with different alternation ratios. As a sensitivity study, we investigate how different ratios of alternating optimization steps between text-VAE and conditional sampler have an effect on the performance of WorDepth. We find that optimizing text-VAE with a lower ratio will lead to a more deterministic model, which is anticipated. On the other hand, optimizing text-VAE more frequently enables the model to learn a better variational prior on the depth maps corresponding to text captions, which facilitates the generation of diverse prior depth maps. However, this comes at the cost of training time as the conditional sampler takes more steps to converge and, given a fixed number of steps, results in more blurry predictions. We identify the ratio at 1% in updating text-VAE to be the best empirically (Tab. 3). Ratios exceeding 10% notably degrades performance given a fixed training length because of fewer updates to the sampler. Note that at 100%, where we do not condition the image, caption to depth generation still yields reasonable results as the text captions produce plausible statistics that match the ground truth depth. On the other hand, without the modeling language as a variational prior

Method	Backbone	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	Abs Rel $\downarrow$	$RMSE_{log} \downarrow$	RMSE $\downarrow$
CLIPMDE [1]	CLIP	0.550	0.830	0.938	0.303	0.119	6.322
DORN [16]	ResNet-101	0.932	0.984	0.995	0.072	0.120	2.727
Yin et al. [80]	ResNeXt-101	0.938	0.990	0.998	0.072	0.117	3.258
TransDepth [78]	ViT-B	0.956	0.994	**0.999**	0.064	0.098	2.755
Big to Small [35]	DenseNet-161	0.955	0.993	0.998	0.060	0.096	2.798
DPT-Hybird [54]	ViT-B	0.959	0.995	**0.999**	0.062	0.092	2.573
ASTransformer [7]	ViT-B	0.963	0.995	**0.999**	0.058	0.089	2.685
AdaBins [2]	EffNet-B5+ViT-mini	0.964	0.995	**0.999**	0.058	0.089	2.360
NeWCRFs [86]	Swin-L	0.974	0.997	**0.999**	0.052	0.079	2.129
Yu et al. [84]	Swin-L	0.972	0.996	**0.999**	0.054	0.081	2.134
DepthFormer [40]	Swin-L	0.975	0.997	**0.999**	0.052	0.079	2.143
Baseline	Swin-L	0.969	0.996	**0.999**	0.054	0.085	2.343
WorDepth	Swin-L	**0.979**	**0.998**	0.999	**0.049**	**0.074**	**2.039**
% Improvement	-	+1.03%	+0.20%	+0.00%	-9.26%	-12.94%	-12.97%

Table 2. **Quantitative results on KITTI Eigen Split.** The baseline method is to directly train a Swin-L image encoder and the depth decoder without the help of language prior. Improvement is the relative performance gain compared with the Baseline.

p	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	AbsRel $\downarrow$	$\log_{10} \downarrow$	RMSE $\downarrow$
0%	0.929	0.990	**0.998**	0.091	0.039	0.323
1%	**0.932**	**0.992**	**0.998**	**0.088**	**0.038**	**0.317**
10%	0.930	0.991	**0.998**	0.090	0.039	0.322
50%	0.763	0.942	0.987	0.163	0.068	0.527
90%	0.642	0.906	0.975	0.211	0.089	0.687
100%	0.590	0.889	0.973	0.225	0.097	0.746

Table 3. **Sensitivity to different ratios of alternating optimization steps between text-VAE and conditional sampler on NYU Depth V2.** p denotes probability of optimizing text-VAE. While more steps spent on text-VAE will yield better generative results, it comes at the cost of slower convergence for the sampler.

Method	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	AbsRel $\downarrow$	$\log_{10} \downarrow$	RMSE $\downarrow$
Adabins	0.771	0.944	0.983	0.159	0.068	0.476
DepthFormer	0.815	0.970	0.993	0.137	0.059	0.408
Baseline	0.803	0.965	0.990	0.141	0.062	0.427
WorDepth	**0.833**	**0.976**	**0.994**	**0.123**	**0.054**	**0.376**

Table 4. **Zero-shot generalization to SUN-RGBD.** The models are trained on the NYU Depth V2 and testing on the Sun-RGBD without any fine-tuning.

(at 0%, where we train both text-VAE and conditional optimizer jointly as a direct map from single image and caption to depth), performance degrade to do the lack of the prior.

Zero-shot Generalization. Given the smaller domain gap in language descriptions across different scenes compared to images, we conduct a zero-shot transfer experiment to highlight our improved generalization ability. We train the model on the NYU Depth V2 [58] and test it on the Sun-RGBD [60] without fine-tuning. As shown in Tab. 4, WorDepth outperforms baseline methods by a substantial margin, indicating the transferability of language priors which underscores the robustness of text descriptions in handling scene variability. This suggests that language descriptions may offer a more stable basis for generalization across diverse data domains than direct visual signals.

5. Discussion

In this study, we seek to answer the question of whether language can be used to calibrate the learned scene distribution to true real-world statistics. The answer is yes, which is valuable for circumventing the long-standing problem of scale ambiguity in monocular depth or structure-from-motion problems. The approach is a first in leveraging complementary properties of two modalities with inherent ambiguities for the 3D reconstruction, to address the deficits in one another. We show that by exploiting the layout/scene ambiguity in language as a strength via our variational approach, we can ground predictions to metric scale. This opens up new avenue in how one can address the issue of scale in 3D reconstruction as well as provide a direct framework to extending the many works that currently are limited to relative or scaleless depth predictions.

Limitations. Generic regularizers typically yield little gains, but do little harm; specific regularizers can provide larger boosts but are limited in their applications. While using language as a prior gives flexibility between the two, specificity in the caption controls the degree of regularization imposed. Naturally, vague captions give little to no information on object shape or size, so there is little to be gained; specific, but incorrect captions may misfire, barring any malicious intent. As WorDepth relies on the quality of the caption, it is susceptible to inaccuracies stemming from descriptions provided by the image captioner. Its ease of use also opens up vulnerabilities from malicious users who may choose captions to steer predictions incorrectly.

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. 3, 6, 8

[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. 2, 3, 5, 6, 7, 8, 1

[3] 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. 2

[4] Irving Biederman and Ginny Ju. Surface versus edge-based determinants of visual recognition. *Cognitive psychology*, 20(1):38–64, 1988. 2

[5] Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J Davison. Codeslam—learning a compact, optimisable representation for dense visual slam. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2560–2568, 2018. 3

[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. 3

[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. 2, 6, 8, 1

[8] Jiaben Chen, Renrui Zhang, Dongze Lian, Jiaqi Yang, Ziyao Zeng, and Jianbo Shi. iquery: Instruments as queries for audio-visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14675–14686, 2023. 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[10] Georgi Dikov and Joris van Vugt. Variational depth networks: Uncertainty-aware monocular self-supervised depth estimation. In *European Conference on Computer Vision*, pages 43–60. Springer, 2022. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[12] Yiming Dou, Fengyu Yang, Yi Liu, Antonio Loquercio, and Andrew Owens. Tactile-augmented radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[13] 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. 4, 5, 6, 1

[14] Xiaohan Fei and Stefano Soatto. Visual-inertial object detection and mapping. In *Proceedings of the European conference on computer vision (ECCV)*, pages 301–317, 2018. 1

[15] Xiaohan Fei, Alex Wong, and Stefano Soatto. Geo-supervised visual depth prediction. *IEEE Robotics and Automation Letters*, 4(2):1661–1668, 2019. 1, 2

[16] 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. 2, 6, 8

[17] 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. 2

[18] 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. 3

[19] 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. 1

[20] 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. 2, 5, 6, 1

[21] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 3

[22] 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. 1

[23] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023. 3

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

9

[25] Jia Cheng Hu, Roberto Cavicchioli, and Alessandro Capotondi. Expansionnet v2: Block static expansion in fast end to end training for image captioning. *arXiv preprint arXiv:2208.06551*, 2022. 2, 5

[26] 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. 3

[27] 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. 2

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3, 4

[30] Neehar Kondapaneni, Markus Marks, Manuel Knott, Rogério Guimarães, and Pietro Perona. Text-image alignment for diffusion-based perception. *arXiv preprint arXiv:2310.00031*, 2023. 3

[31] Barbara Landau, Linda B Smith, and Susan S Jones. The importance of shape in early lexical learning. *Cognitive development*, 3(3):299–321, 1988. 2

[32] Barbara Landau, Linda Smith, and Susan Jones. Object shape, object function, and object name. *Journal of memory and language*, 38(1):1–27, 1998. 2

[33] 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. *arXiv preprint arXiv:2203.13987*, 2022. 2

[34] 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. 5

[35] 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. 2, 6, 8

[36] 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. 2

[37] 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. 3

[38] 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. 3

[39] Yong-Lu Li, Xiaoqian Wu, Xinpeng Liu, Yiming Dou, Yikun Ji, Junyi Zhang, Yixing Li, Jingru Tan, Xudong Lu, and Cewu Lu. From isolated islands to pangea: Unifying semantic space for human action understanding. *arXiv preprint arXiv:2304.00553*, 2023. 3

[40] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*, 2022. 2, 6, 8

[41] Yiqing Liang, Numair Khan, Zhengqin Li, Thu Nguyen-Phuoc, Douglas Lanman, James Tompkin, and Lei Xiao. Gaufre: Gaussian deformation fields for real-time dynamic novel view synthesis, 2023. 3

[42] Yiqing Liang, Eliot Laidlaw, Alexander Meyerowitz, Srinath Sridhar, and James Tompkin. Semantic attention flow fields for monocular dynamic scene decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[43] 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. 6, 1

[44] Tian Yu Liu, Parth Agrawal, Allison Chen, Byung-Woo Hong, and Alex Wong. Monitored distillation for positive congruent depth completion. In *European Conference on Computer Vision*, pages 35–53. Springer, 2022. 2

[45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 5

[46] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Wei Li, Christian Theobalt, Ruigang Yang, and Wenping Wang. Adaptive surface normal constraint for depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12849–12858, 2021. 2, 6

[47] Mazen Mel, Muhammad Siddiqui, and Pietro Zanuttigh. End-to-end learning for joint depth and image reconstruction from diffracted rotation. *arXiv preprint arXiv:2204.07076*, 2022. 2

[48] 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. 3

[49] Zixuan Pan, Zihao Wei, and Andrew Owens. Efficient vision-language pre-training by cluster masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[50] 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*, 2024. 2

[51] 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. 2

[52] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint

depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018. 2, 6

[53] 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. 3, 5

[54] 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. 2, 6, 8

[55] 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. 3

[56] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023. 3

[57] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[58] 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. 2, 5, 6, 8, 1

[59] 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. 2

[60] 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. 8

[61] 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. 2

[62] 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. 2

[63] 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.

[64] 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. 2

[65] 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. 1

[66] Alex Wong, Safa Cicek, and Stefano Soatto. Targeted adversarial perturbations for monocular depth prediction. *Advances in neural information processing systems*, 33:8486–8497, 2020. 2

[67] 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. 1, 2

[68] 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.

[69] 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. 2

[70] 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. 2

[71] Shaokai Wu and Fengyu Yang. Boosting detection in crowd analysis via underutilized output features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15609–15618, 2023. 3

[72] Yangchao Wu, Tian Yu Liu, Hyoungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. Augundo: Scaling up augmentations for unsupervised depth completion. *arXiv preprint arXiv:2310.09739*, 2023. 2

[73] Zhihao Xia, Patrick Sullivan, and Ayan Chakrabarti. Generating and exploiting probabilistic monocular depth estimates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 65–74, 2020. 3

[74] Fengyu Yang and Chenyang Ma. Sparse and complete latent organization for geospatial semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1809–1818, 2022. 3

[75] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *Neural Information Processing Systems (NeurIPS) - Datasets and Benchmarks Track*, 2022.

[76] Fengyu Yang, Jiacheng Zhang, and Andrew Owens. Generating visual scenes from touch. *International Conference on Computer Vision (ICCV)*, 2023. 3

[77] 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 tac-

tile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[78] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *Proceedings of the IEEE/CVF International Conference on Computer vision*, pages 16269–16279, 2021. 2, 6, 8

[79] 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. 2

[80] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5684–5693, 2019. 2, 6, 8

[81] Chenyu You, Weicheng Dai, Fenglin Liu, Yifei Min, Haoran Su, Xiaoran Zhang, Xiaoxiao Li, David A Clifton, Lawrence Staib, and James S Duncan. Mine your own anatomy: Revisiting medical image segmentation with extremely limited labels. *arXiv preprint arXiv:2209.13476*, 2022. 3

[82] Chenyu You, Weicheng Dai, Yifei Min, Lawrence Staib, and James S Duncan. Implicit anatomical rendering for medical image segmentation with stochastic experts. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 561–571. Springer, 2023. 3

[83] Chenyu You, Weicheng Dai, Yifei Min, Fenglin Liu, David Clifton, S Kevin Zhou, Lawrence Staib, and James Duncan. Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[84] Shangbin Yu, Renyan Zhang, Shuaiye Ma, and Xinfang Jiang. Monocular depth estimation network based on swin transformer. In *Journal of Physics: Conference Series*, page 012019. IOP Publishing, 2023. 2, 6, 8

[85] 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. 2

[86] 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. 2, 5, 6, 8

[87] 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. 3

[88] 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. 3

[89] Renrui Zhang, Ziyao Zeng, Ziyu Guo, Xinben Gao, Kexue Fu, and Jianbo Shi. Dspoint: Dual-scale point cloud

[90] 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. 3

[91] 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. 3, 6

[92] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *ICLR 2024*, 2023. 3

[93] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024. 3

[94] 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. 2

[95] Hanbin Zhao, Fengyu Yang, Xinghe Fu, and Xi Li. Rbc: Rectifying the biased context in continual semantic segmentation. *ECCV*, 2022. 3

[96] 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. 2

[97] 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. 3

[98] Chenhao Zheng, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. Iterated learning improves compositionality in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[99] Chong Zhou, Chen Change Loy, and Bo Dai. Denseclip: Extract free dense labels from clip. *arXiv preprint arXiv:2112.01071*, 2021. 3

[100] 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. 2

[101] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021. 3

[102] 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. 1

[103] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. *arXiv preprint arXiv:2201.02605*, 2022. 3

[104] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023. 3

[105] 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. 3

WorDepth: Variational Language Prior for Monocular Depth Estimation

Supplementary Material

A. Evaluation metrics.

Drawing on [7, 43], our evaluation of WorDepth alongside comparison methods involves a quantitative assessment through several metrics. These include 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). The evaluation metrics are summarized in Table 5 for details.

B. Ablation on Model Architecture

We evaluated varying hidden variables d of text-VAE using the NYU Depth V2 dataset[58], shown in Table 6. A key consideration was ensuring the hidden space was sufficiently large to encode the necessary structural and geometric features for reconstructing depth maps. This size requirement arises from the need to preserve essential features about the scene's objects and layout derived from text features encoded by text-VAE.

However, it's equally crucial to avoid excessively large hidden variables. A relatively constrained dimensionality acts as a form of regularization, compelling the text-VAE to focus on extracting features crucial for depth decoding. Additionally, a limited hidden dimension prompts the model to learn not just the distribution mean but also its variance. This aspect is particularly important when mapping a text description to multiple scenes, such scenes' text features are encoded with identical distribution means but exhibit significant variance.

We established hidden variables d of 32, 64, 128, 256, 512, and 1024 for training WorDepth. It was observed that the optimal hidden dimension is 128, striking a balance between capturing sufficient geometric features of scenes while maintaining effective regularization. Deviating from this optimal size, either too small or too large, adversely impacts performance.

Metric	Formulation
Abs Rel	$\frac{1}{N_e}\sum_{(i,j)\in\Omega}\frac{|y^*(i,j)-y(i,j)|}{y^*(i,j)}$
RMSE	$\sqrt{\frac{1}{N_e}\sum_{(i,j)\in\Omega}(y^*(i,j)-y(i,j))^2}$
$\log_{10}$	$\frac{1}{N_e}\sum_{(i,j)\in\Omega}|\log_{10}(y^*(i,j))-\log_{10}(y(i,j))|$
$\text{RMSE}_{\log}$	$\sqrt{\frac{1}{N_e}\sum_{(i,j)\in\Omega}(\ln(y^*(i,j))-\ln(y(i,j)))^2}$
δ	% of $y(i,j)$ s.t. $\max(\frac{y(i,j)}{y^*(i,j)},\frac{y^*(i,j)}{y(i,j)}) < thr \in [1.25, 1.25^2, 1.25^3]$

Table 5. **Evaluation metric for monocular depth estimation.** y denotes predictions and y^* denotes ground truth.

Method	$\delta < 1.25\uparrow$	$\delta < 1.25^2\uparrow$	$\delta < 1.25^3\uparrow$	Abs Rel↓	$\log_{10}$↓	RMSE↓
$d = 32$	0.925	0.990	**0.998**	0.093	0.039	0.327
$d = 64$	0.928	0.990	**0.998**	0.090	0.039	0.325
$d = 128$	**0.932**	**0.992**	**0.998**	**0.088**	**0.038**	**0.317**
$d = 256$	0.930	0.991	**0.998**	0.089	0.039	0.323
$d = 512$	0.929	0.990	**0.998**	0.089	0.039	0.324
$d = 1024$	0.926	0.989	**0.998**	0.091	0.039	0.325

Table 6. **Sensitivity to different numbers of hidden variables d.** Experiments are conducted on NYU Depth V2. d is the number of hidden variables d of the text-VAE.

C. Additional Visualization on NYU Depth V2

In this section, as illustrated in Figure 5, We present additional visualizations comparing WorDepth with a baseline method AdaBins [2] on the NYU Depth V2 [58] dataset, emphasizing the advantages gained from integrating the language prior. Compared with AdaBins, the error map, with its brighter regions highlighting larger errors, clearly demonstrates that WorDepth achieves more precise depth predictions for objects identified in the text description. For instance: "a sink and a bath tub" in the first row, "a white bath tub" in the second row, "a wooden dresser" in the third row, "a bed" in the fourth row, "a bunk bed" in the fifth row, "an unmade bed with clothes on top of it" in the sixth row, "a couch and a table" in the seventh row, "a table and chairs" in the eighth row, "a blender on a counter" in the ninth row, "chairs" in the tenth row, and "machine on top of a wooden table" in the last row.

D. Additional Visualization on KITTI

This section, depicted in Figure 6, showcases visualizations of Monocular Depth Estimation in outdoor scenarios with the KITTI dataset [20] using Eigen Split [13], comparing with Adabins [2]. Due to the limited variety of objects in outdoor scenes, our method captures fewer objects compared to indoor scenes. However, when salient objects and scenes are present outdoors, our method gains a preliminary understanding of their scale. This understanding aids in enhancing monocular depth estimation for these objects. The error map's brighter regions, which emphasize greater absolute relative errors, unequivocally show that WorDepth outperforms AdaBins in making more accurate depth predictions for objects and scenes mentioned in the text description. For instance: "two white trucks" in the upper right, "a woman riding a scooter" in the lower left, "buildings" in the lower middle, and "forest with tree" in the lower left.

Image	Text	Ours	Ours Error	Adabins	Adabins Error	Ground Truth

A bathroom with a sink and a bath tub.

A white bath tub sitting in a bathroom.

A television sitting on top of a wooden dresser.

A man standing next to a bed in a bedroom.

A bunk bed with a table in a room.

An unmade bed with clothes on top of it.

A living room with a couch and a table.

A dining room with a table and chairs.

A kitchen with a blender on a counter.

A living room with chairs and a table.

A machine sitting on top of a wooden table.
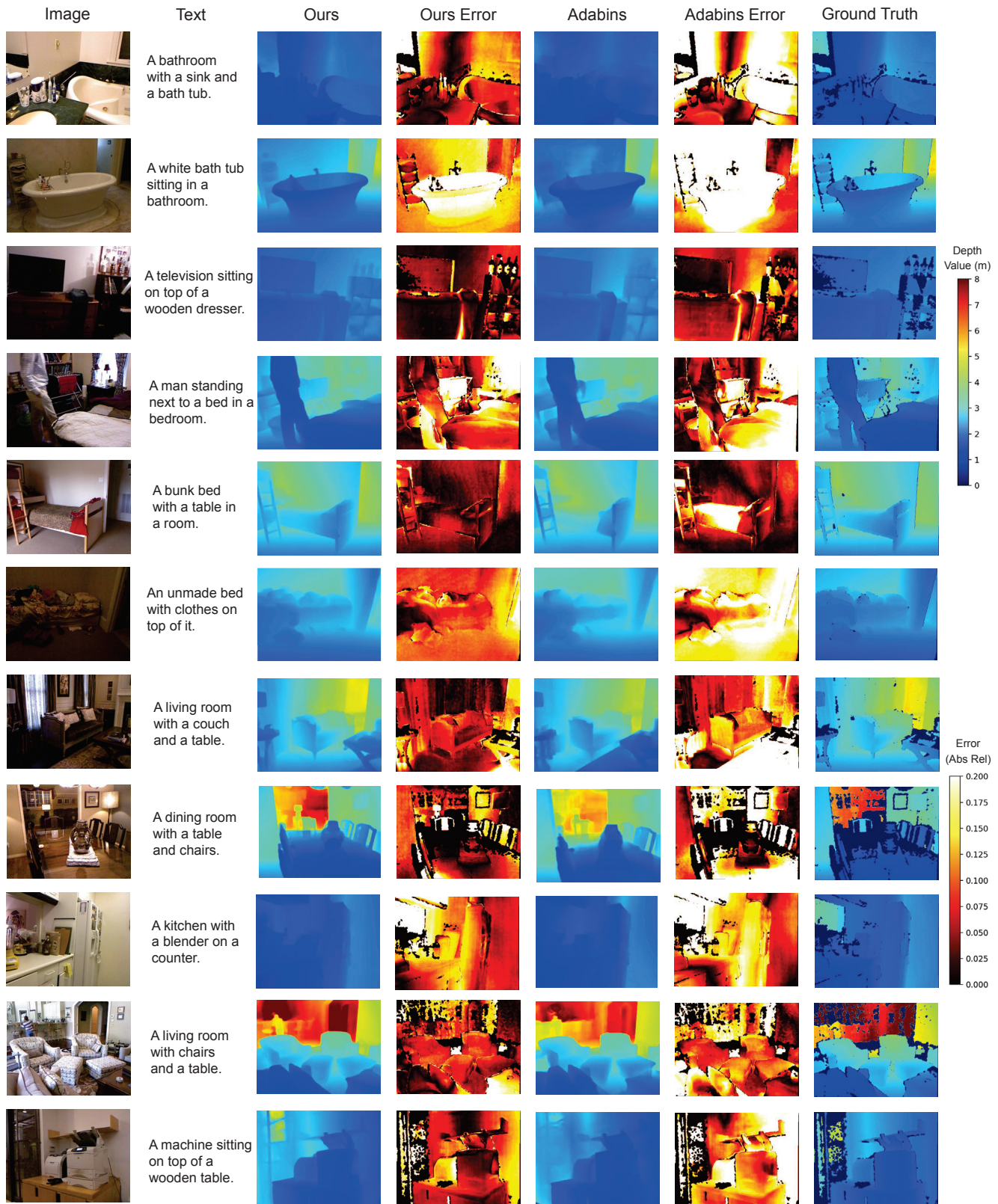
Depth Value (m)

Error (Abs Rel)

Figure 5. **Additional visualization of monocular depth estimation on NYU Depth V2.**
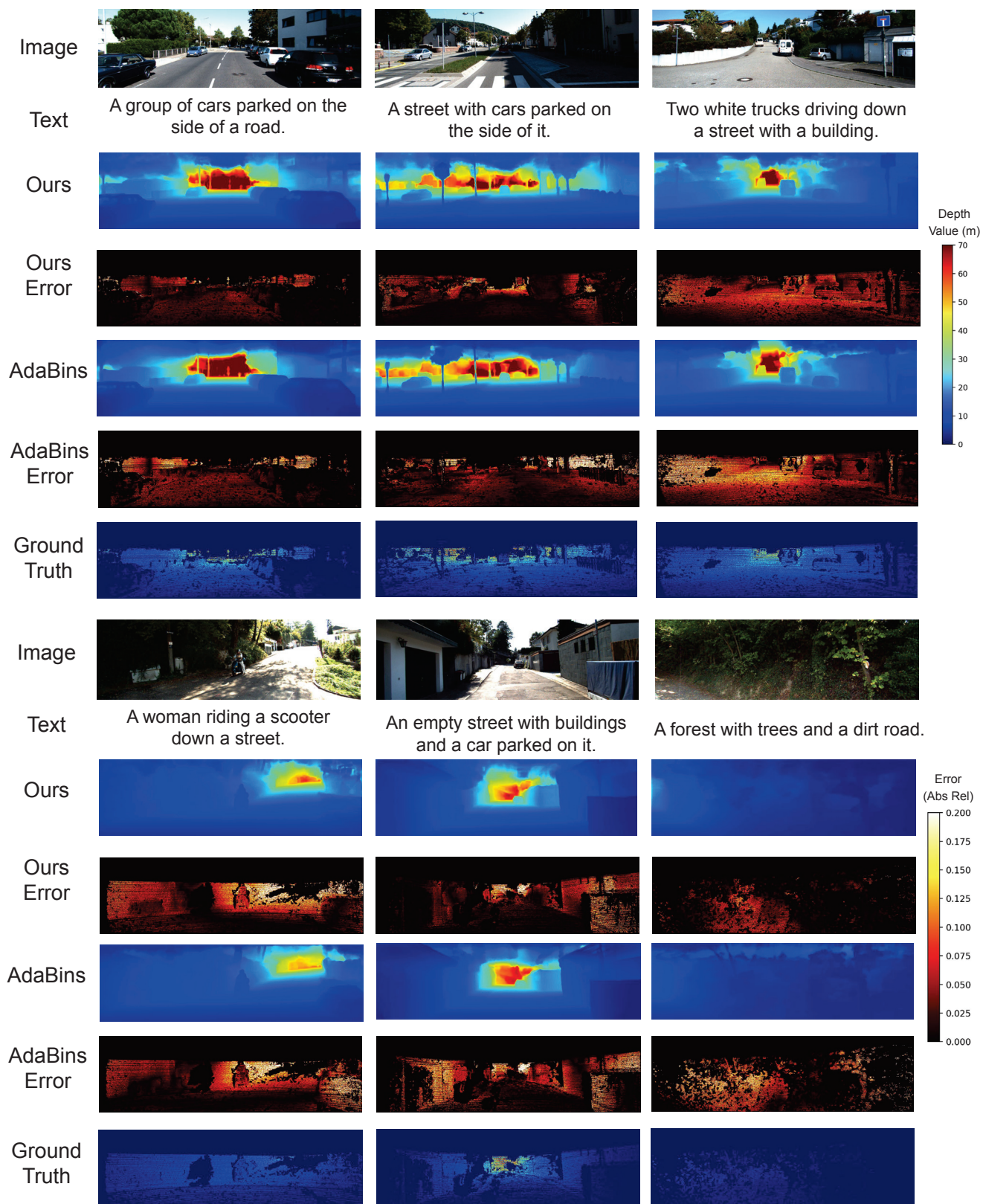
Figure 6. **Additional visualization of monocular depth estimation on KITTI Eigen split.**