WorDepth: Variational Language Prior for Monocular Depth Estimation

CVPR JUNE 17-21, 2024 SEATTLE, WA

Ziyao Zeng[1] Daniel Wang[1] Fengyu Yang[1] Hyoungseob Park[1] Stefano Soatto[2] Dong Lao[2] Alex Wong[1]
{ziyao.zeng daniel.wang.dhw33 fengyu.yang hyougseob.park alex.wong}@yale.edu {soatto lao}@cs.ucla.edu
Yale Vision Lab[1] UCLA Vision Lab[2]

Paper Code

Preliminaries

3D reconstruction from a single image is an ill-posed problem – there exists infinitely many 3D scenes (e.g. scale) that can generate an image

3D reconstruction from a text caption is also an ill-posed problem – there exists infinitely many 3D scenes that fits a description

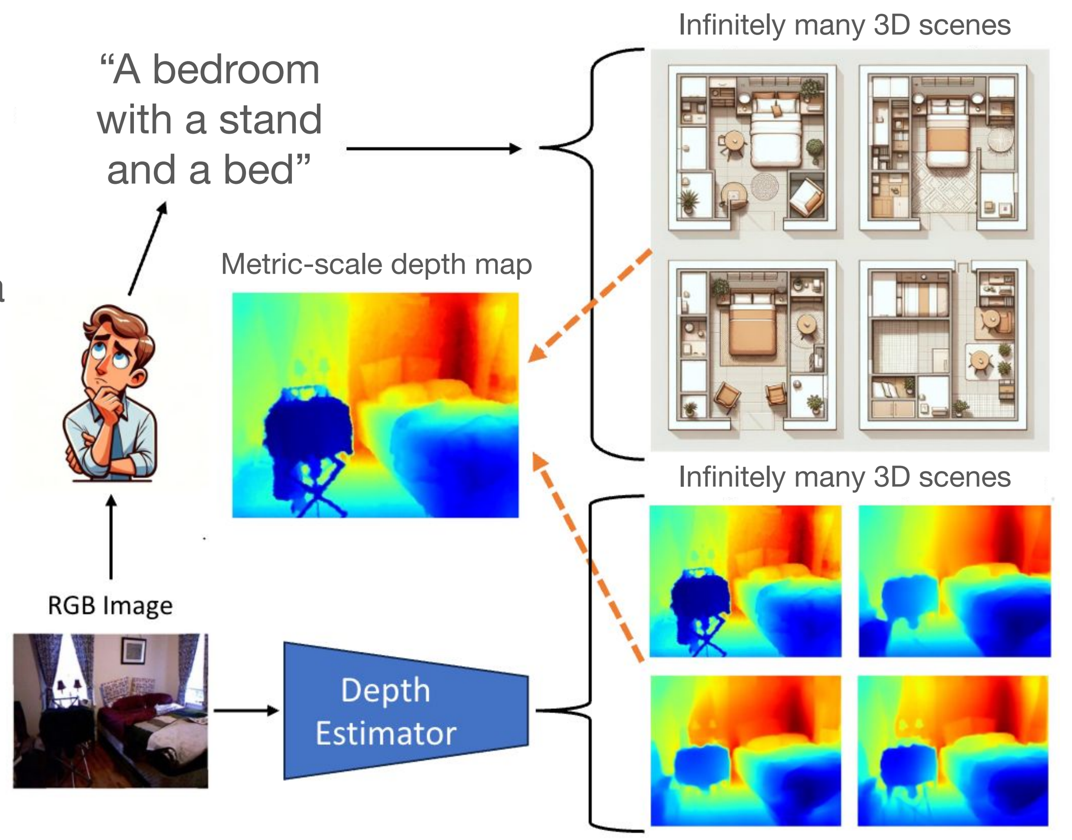
Motivation

To ground depth predictions to metric scale, one may use additional cameras with known position (stereo), or additional sensors (range, inertial, etc.)

But it is often costly in additional price, data collection, power, computational resources

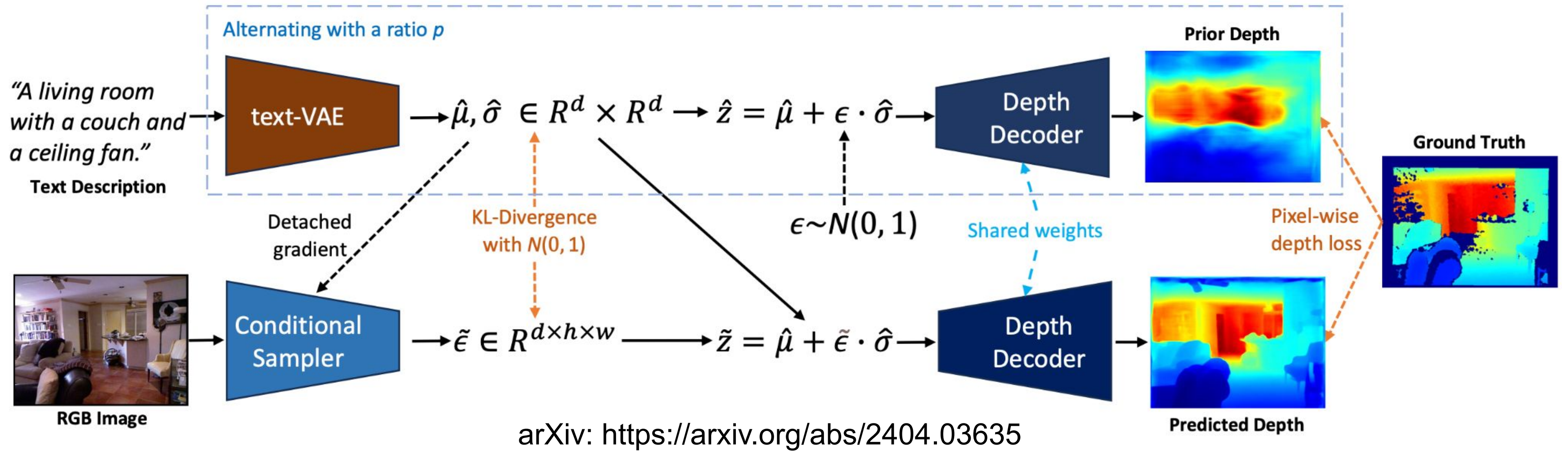
Question: Can two modalities that are inherently ambiguous resolve one another's ambiguity in 3D reconstruction?

Key idea: Use language to ground depth estimates to metric scale! Simply let the model know what objects are around and it can better estimate scale



Overview

Train a text-VAE to encode text into the mean and standard deviation parameterizing the distribution of 3D scenes for a description. Choose one of the infinitely many scenes matching the description that is compatible with the observed image.



arXiv: https://arxiv.org/abs/2404.03635

Training WorDepth

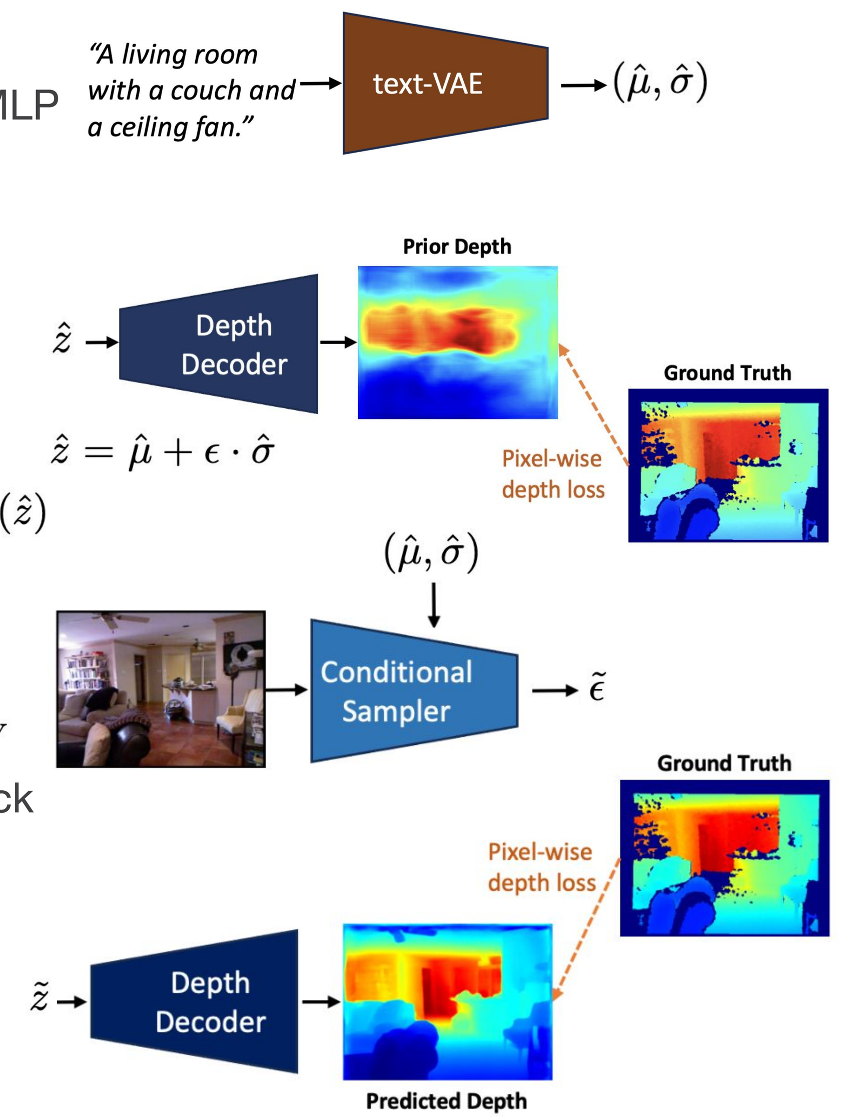
(1) Encode the text caption using the CLIP text encoder and estimate its mean and standard deviation $(\hat{\mu}, \hat{\sigma})$ using an MLP

(2) Alternatingly optimize (a), (b)

(a) Update text-VAE (freeze Conditional Sampler)
- (i) Draw from a standard Gaussian $\epsilon \sim \mathcal{N}(0, 1)$
- (ii) Sample latent vector using the reparameterization trick $\hat{z} = \hat{\mu} + \epsilon \cdot \hat{\sigma}$
- (iii) Predict depth map from using depth decoder $\hat{y} = h_\phi(\hat{z})$
- (iv) Minimize $\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{SI}}(y^*, \hat{y}) + \alpha \cdot \mathcal{L}_{\text{KL}}(\hat{\mu}, \hat{\sigma})$

(b) Update the Conditional Sampler (freeze text-VAE)
- (i) Sampler using an image $\tilde{\epsilon} = f_\varphi(x, \hat{\mu}, \hat{\sigma}),\ x \in \mathbb{R}^{3 \times H \times W}$
- (ii) Sample a latent vector using the reparameterization trick
- (iii) Predict depth map using depth decoder $\tilde{y} = h_\phi(\tilde{z})$
- (iv) Minimize $\mathcal{L}_{\text{CS}} = \mathcal{L}_{\text{SI}}(y^*, \tilde{y}) + \beta \cdot \mathcal{L}_{\text{KL}}(\tilde{\mu}, \tilde{\sigma})$

(3) Repeat steps 1, 2(a) and 1, 2(b) until convergence



Scale invariant loss:

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

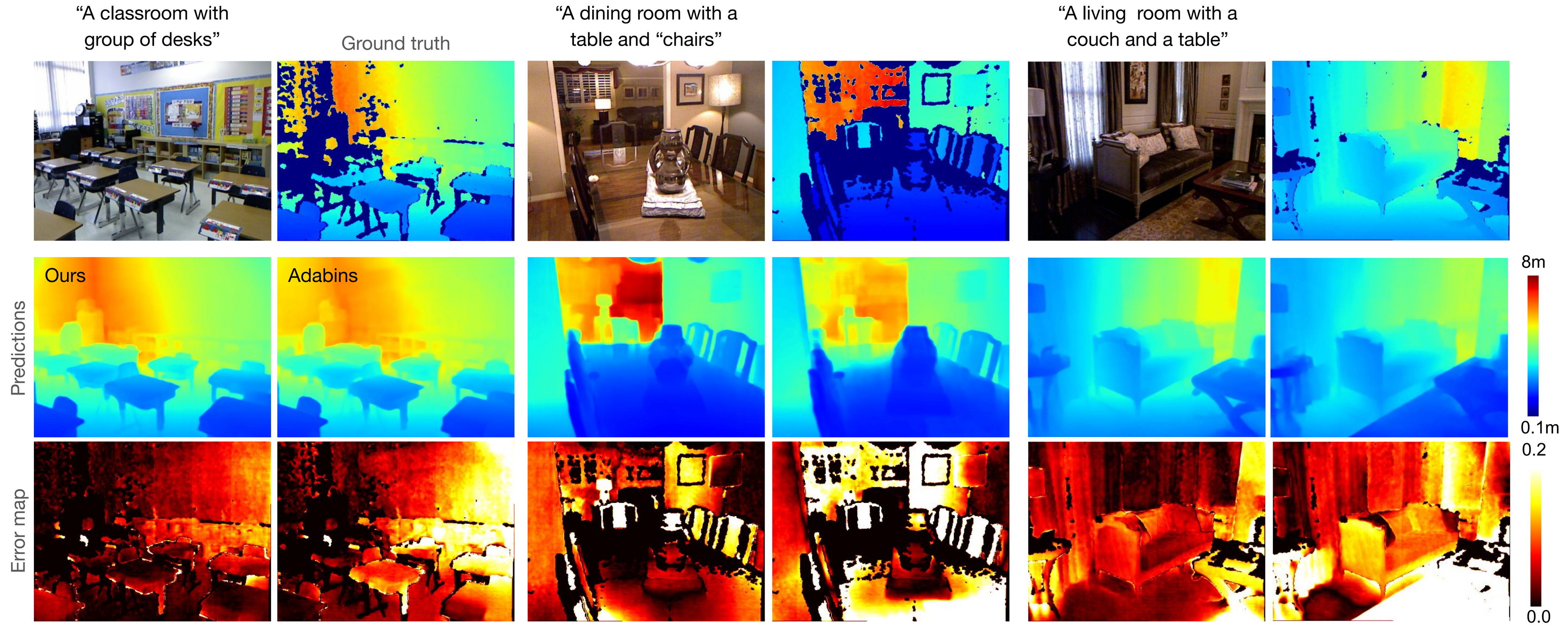
Kullback-Leibler (KL) divergence loss:

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

where y^* denotes ground truth, N_e the number of elements in the image, γ a scaling factor

Qualitative Results

Knowing that certain objects (and that they are typically of certain sizes) exist in the scene, we can better estimate the scale as evident by the uniform improvement over the error maps.



github: https://github.com/Adonis-galaxy/WorDepth

Quantitative Results

NYUv2 Benchmark

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%

KITTI Eigen Split Benchmark

Method	Backbone	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	Abs Rel $\downarrow$	$\text{RMSE}_{\log} \downarrow$	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%

Zero-shot Generalization: NYUv2 → SUN RGBD

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**

WorDepth consistently improves over existing methods across both indoor (NYUv2) and outdoor (KITTI) benchmarks.

Due to the flexibility of using text captions to ground predictions to scale, WorDepth shows consistent better zero-shot generalization

Different alternating ratios: NYUv2

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**
50%	0.763	0.942	0.987	0.163	0.068	0.527
100%	0.590	0.889	0.973	0.225	0.097	0.746
t - C	0.926	0.990	**0.998**	0.091	0.039	0.330

"t - C" indicates training text-VAE to convergence then freeze it and train Conditional Sampler.

Limitation. Caption specificity controls regularization. Vague captions provide little information on object shape or size, yielding minimal gains, and specific but incorrect captions can mislead the model.

This work was supported by NSF 2112562 (Athena AI Institute).