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	$\frac{1}{N_e}\sum_{(i,j)\in\Omega}\frac{\lvert y^*(i,j)-y(i,j)\rvert}{y^*(i,j)}$
RMSE	$\sqrt{\frac{1}{N_e}\sum_{(i,j)\in\Omega}(y^*(i,j)-y(i,j))^2}$
$\log_{10}$	$\frac{1}{N_e}\sum_{(i,j)\in\Omega}\lvert\log_{10}(y^*(i,j))-\log_{10}(y(i,j))\rvert$
$\text{RMSE}_{\log}$	$\sqrt{\frac{1}{N_e}\sum_{(i,j)\in\Omega}(\ln(y^*(i,j))-\ln(y(i,j)))^2}$
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 $\downarrow$	$\log_{10}\downarrow$	RMSE $\downarrow$
$d=32$	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
$\boldsymbol{d=128}$	**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.

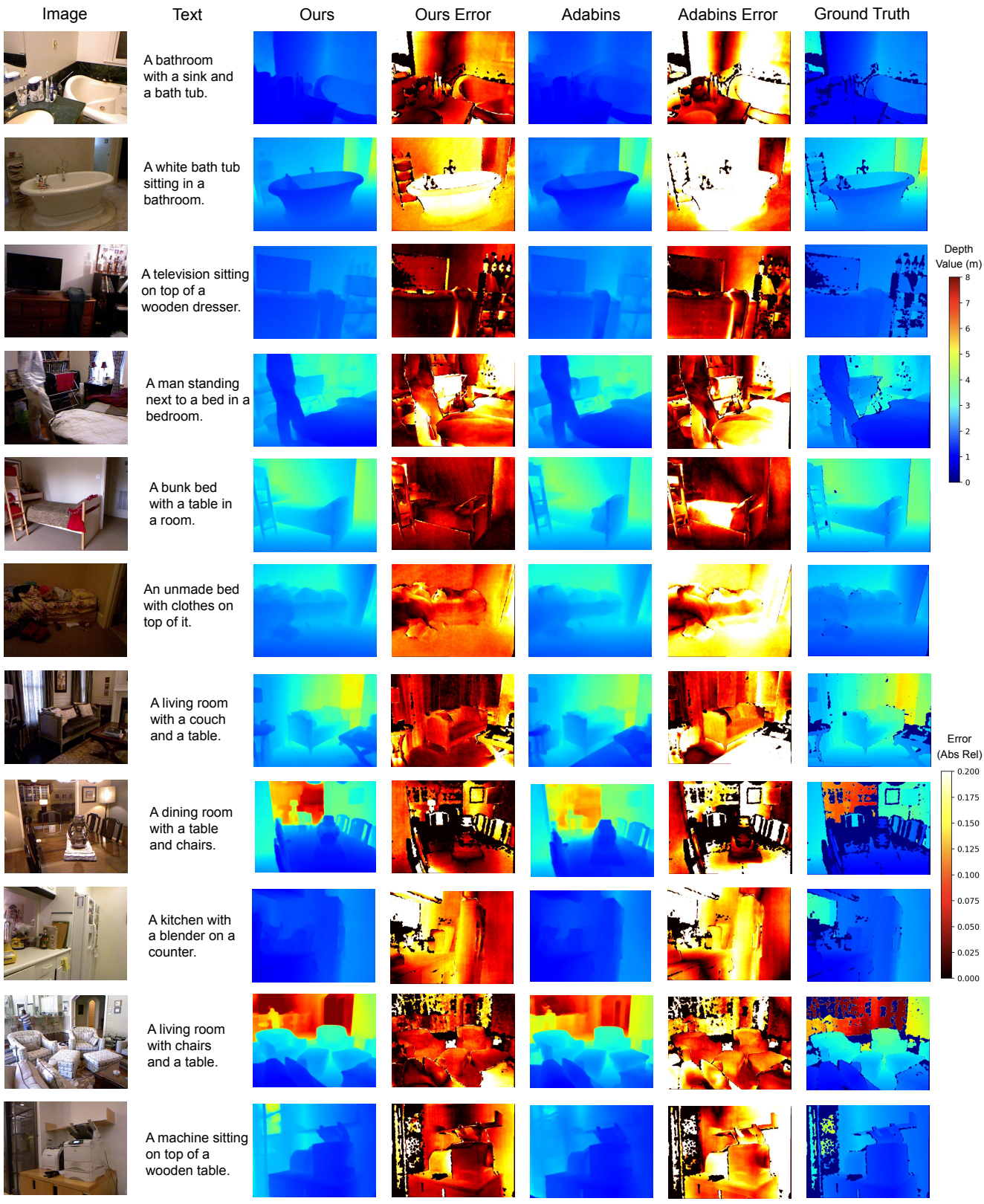

Figure 5. **Additional visualization of monocular depth estimation on NYU Depth V2.**

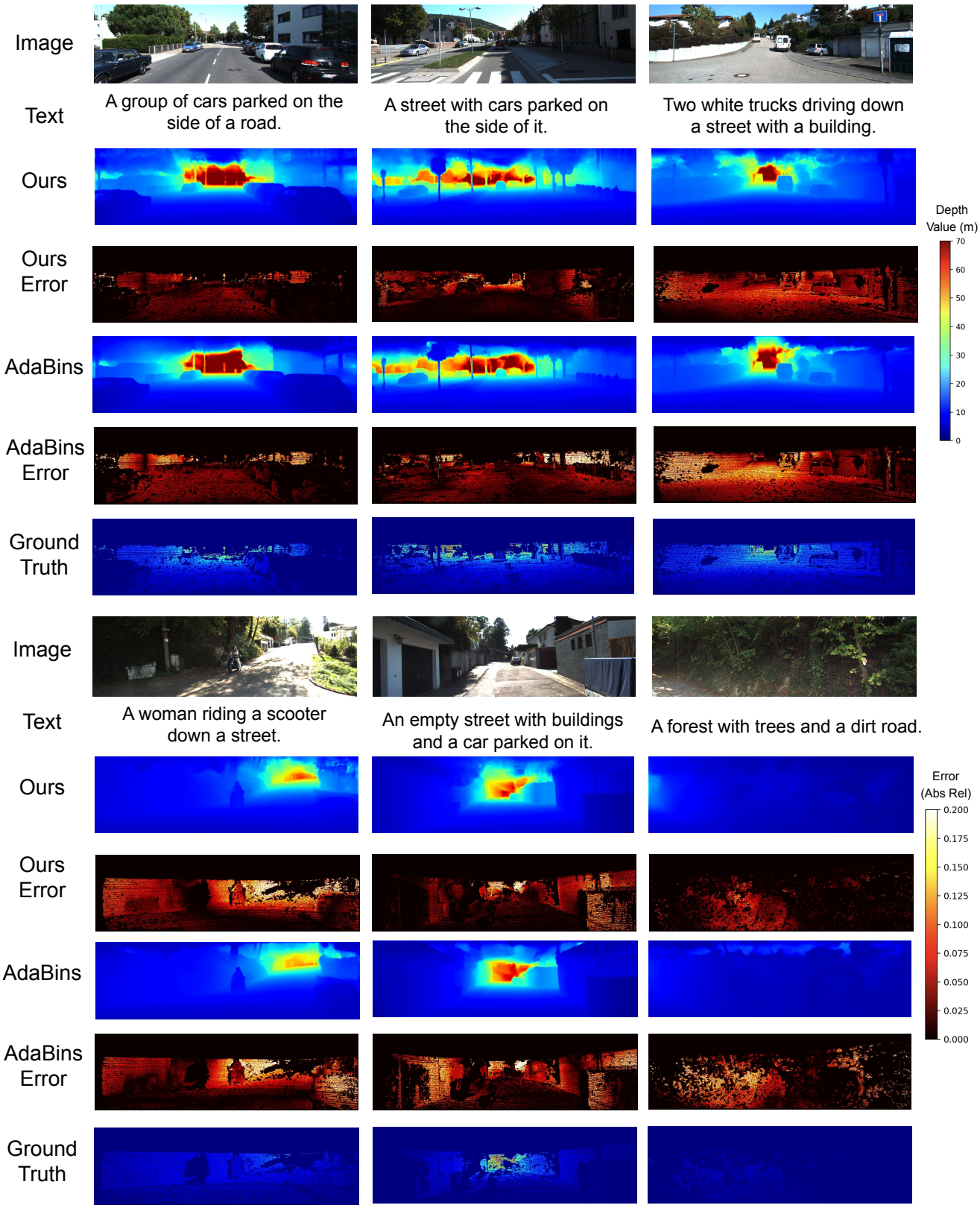

Figure 6. **Additional visualization of monocular depth estimation on KITTI Eigen split.**