

# BRIDGE - BUILDING REINFORCEMENT-LEARNING DEPTH-TO-IMAGE DATA GENERATION ENGINE FOR MONOCULAR DEPTH ESTIMATION

Anonymous authors

Paper under double-blind review



Figure 1: We present BRIDGE, showcasing its RL-optimized Depth-to-Image (D2I) data generation engine which is used for generating realistic and geometrically accurate RGB images from source depth maps and Monocular Depth Estimation (MDE) model which after being trained on the massive high-quality data generated by the D2I engine, achieves superior depth prediction in complex scenes.

## ABSTRACT

Monocular Depth Estimation (MDE) is a foundational task for computer vision. Traditional methods are limited by data scarcity and quality, hindering their robustness. To overcome this, we propose BRIDGE, an RL-optimized depth-to-image (D2I) generation framework that synthesizes over 20M realistic and geometrically accurate RGB images, each intrinsically paired with its ground truth depth, from diverse source depth maps. Then we train our depth estimation model on this dataset, employing a hybrid supervision strategy that integrates teacher pseudo-labels with ground truth depth for comprehensive and robust training. This innovative data generation and training paradigm enables BRIDGE to achieve breakthroughs in scale and domain diversity, consistently outperforming existing state-of-the-art approaches quantitatively and in complex scene detail capture, thereby fostering general and robust depth features. Code and models are available at <https://dingning-liu.github.io/bridge.github.io/>.

## 1 INTRODUCTION

Monocular Depth Estimation (MDE) stands as a cornerstone in the field of 3D computer vision, providing essential geometric perception for various critical applications such as 3D reconstruction (Mildenhall et al., 2021; Kerbl et al., 2023; Charatan et al., 2024; Ye et al., 2024), autonomous driving (Wang et al., 2019b), robotics (Wofk et al., 2019), and AR/VR (Rasla & Beyeler, 2022). The field has witnessed notable progress in recent years (Atapour-Abarghouei & Breckon, 2018; Ranftl et al., 2020; Bhat et al., 2023; Hu et al., 2024; Rajpal et al., 2023; Yang et al., 2024b; Ke et al., 2024; Yang et al., 2024c;d; Fu et al., 2024; Wang et al., 2025), however, training robust MDE models with excellent generalization remains challenging due to the scarcity of high-quality, precise ground truth depth annotations, insufficient detail and diversity in existing labels, and the underutilization of available depth data, creating a critical bottleneck for MDE model training.

Existing data acquisition and utilization methods can be categorized into three groups. (1) Bhat et al. (2023); Hu et al. (2024); Piccinelli et al. (2024) primarily rely on real-world data. While valuable, this approach is constrained by the sparsity of depth maps from sensors and the difficulty of accurately capturing and annotating transparent or reflective objects (Costanzino et al., 2023).

(2) Atapour-Abarghouei & Breckon (2018); Rajpal et al. (2023); Yang et al. (2024b); Ke et al. (2024) are mainly based on synthetic data. Although synthetic data is theoretically precise, it may introduce significant domain gaps (Ganin & Lempitsky, 2015) and geometric artifacts, degrading models' generalization capabilities.

(3) Ranftl et al. (2020); Yang et al. (2024c;d); Fu et al. (2024); Wang et al. (2025) integrate diverse data sources. MiDaS (Ranftl et al., 2020) as the pioneer significantly enhances generalization by training on various mixed datasets; however, its data coverage and diversity remain limited. Building upon this, Depth Anything (Yang et al., 2024c;d) advances SOTA by leveraging massive real-world images and a teacher model for pseudo-label generation, significantly improving generalization. However, its dependence on large real datasets and pseudo-label inaccuracies (despite high generation quality) bottleneck further performance and efficiency. These data-level limitations—diversity, scale, and pseudo-label fidelity—hinder truly universal and efficient MDE models.

To address the above challenges, we propose BRIDGE, an RL-optimized, large-scale Depth-to-Image (D2I) generation data engine. This engine leverages precise ground truth from synthetic depth datasets to generate massive, high-quality, and diverse RGB-D training data. Data created by the data engine enable MDE models to achieve breakthroughs in both scale and domain diversity, thereby fostering general, robust depth features and excellent real-world performance.

Specifically, our core lies in an innovative data generation pipeline: we first train an RL-optimized Depth-to-Image (D2I) generation model on many synthetic datasets. This model can directly utilize diverse source depth maps (from existing synthetic depth datasets) to synthesize over 20 million diverse and information-rich RGB images. Based on this, we introduce a hybrid data supervision strategy. This strategy pairs generated RGB images with original high-precision ground truth depth, which is screened through similarity detection methods such as SSIM and gradient analysis. Si-

108 multaneously, it leverages a large number of pseudo-labels generated by a teacher model trained  
 109 on synthetic data, and both are jointly used for training. This ensures that the generated RGB im-  
 110 ages receive reliable supervision during training while improving the overall quality and utilization  
 111 efficiency of the labels.

112 Our experimental results consistently demonstrate BRIDGE’s superior performance across various  
 113 challenging benchmarks, including indoor, outdoor, and synthetic animation environments, surpassing  
 114 existing state-of-the-art methods. Qualitatively, BRIDGE excels at capturing fine-grained details  
 115 and maintaining robustness in complex scene structures. Furthermore, ablation studies confirm the  
 116 effectiveness of the RL-D2I generated data and the hybrid supervision training method.

117 Our main contributions are as follows:  
 118

- 119 • **An efficient data generation and utilization paradigm:** We adopt an RL-driven D2I  
 120 paradigm, efficiently generating massive high-quality RGB-D data, which effectively alle-  
 121 viates data scarcity and quality issues.
- 122 • **Key practice of hybrid depth supervision training strategy:** We practice a hybrid depth  
 123 supervision strategy, combining teacher model pseudo-labels with high-precision ground  
 124 truth depth, to achieve a more challenging optimization objective and learn geometric struc-  
 125 tural knowledge from RGB data.
- 126 • **Superior performance and high training efficiency:** Our data generation and training  
 127 strategy, using only approximately 20M data, surpasses SOTA models (e.g., Depth Any-  
 128 thing V2 which uses 62M data), significantly improving training efficiency and demon-  
 129 strating excellent detail capture capability and robustness.

## 131 2 RELATED WORKS

### 132 2.1 MONOCULAR DEPTH ESTIMATION

133 Monocular Depth Estimation (MDE), a foundational computer vision task, has seen significant  
 134 progress with deep learning, moving beyond early methods limited by data homogeneity (Saxena  
 135 et al., 2008). Deep learning approaches are broadly categorized into Discriminative Methods and  
 136 Generative Methods. The former, starting with CNN-based pioneers like (Krizhevsky et al., 2009;  
 137 Eigen et al., 2014; Eigen & Fergus, 2015; Chen et al., 2016) and later enhanced by multi-dataset  
 138 training in MiDaS (Ranftl et al., 2020), learns a direct mapping from image to depth (Fu et al.,  
 139 2018). The adoption of the Transformer architecture (Dosovitskiy et al., 2020; Ranftl et al., 2021;  
 140 Liu et al., 2021) and large-scale self-supervised learning, as in Depth Anything (Yang et al., 2024c),  
 141 further improves generalization, while models like NDDdepth (Shao et al., 2023a) and Depth Pro  
 142 (Bochkovskii et al., 2024) target high-precision boundary estimation. In contrast, generative meth-  
 143 ods utilize models like Diffusion Models (Ho et al., 2020; Song et al., 2020; Nichol & Dhariwal,  
 144 2021; Ji et al., 2023) to synthesize depth maps, with works such as LeReS (Yin et al., 2021) and  
 145 Metric3D v2 (Hu et al., 2024) focusing on accurate metric depth. The recent Depth Anything  
 146 V2 (Yang et al., 2024d) combines large-scale synthetic images and pseudo-labeling to boost per-  
 147 formance. Despite these advancements, a key challenge remains: achieving highly accurate and  
 148 generalizable depth estimation in complex, unseen scenarios. To tackle this, we propose BRIDGE,  
 149 a framework that generates a massive dataset of realistic and geometrically accurate RGB images to  
 150 expand the effective training data.  
 151

### 152 2.2 SELF-SUPERVISED AND HYBRID SUPERVISION LEARNING

153 To train a robust MDE model despite the scarcity of precisely labeled data, researchers have explored  
 154 various weakly supervised and self-supervised approaches, such as utilizing unlabeled monocular or  
 155 stereo video data through photometric consistency loss for training (Godard et al., 2017; Kuznietsov  
 156 et al., 2017; Ren et al., 2020). Other self-supervised approaches also include consistency training  
 157 methods like FixMatch (Sohn et al., 2020) and unsupervised data augmentation (Xie et al., 2020),  
 158 as well as strategies like shrinking class space (Yang et al., 2023a) to enhance certainty. Build-  
 159 ing on this, Pseudo-labeling supervision strategies (Lee et al., 2013), exemplified by Depth Any-  
 160 thing V2 (Yang et al., 2024d), significantly improve model generalization by employing a powerful  
 161

162 teacher model to generate pseudo-labels for massive amounts of unlabeled images. However, de-  
 163 spite the broad coverage provided by pseudo-labels, their inherent noise and inaccuracy—especially  
 164 around boundaries and fine details—remain a bottleneck for further enhancing depth estimation per-  
 165 formance. To leverage both the scale of pseudo-labels and the precision of ground truth annotations,  
 166 we propose a hybrid supervision strategy. Unlike methods that solely rely on pseudo-labels, our  
 167 strategy integrates high-precision ground truth depth from the original data via a similarity detection  
 168 mechanism. By utilizing the teacher model’s pseudo-labels for broad training and the similarity-  
 169 guided ground truth for precision fine-tuning, our hybrid approach provides a more reliable and  
 170 comprehensive supervision signal.

### 171 2.3 DEPTH-TO-IMAGE GENERATION AND SYNTHETIC DATA

172 The performance of Monocular Depth Estimation (MDE) is directly linked to the scale and quality  
 173 of training data, yet acquiring high-quality ground truth depth from real-world data is expensive  
 174 and lacks sufficient diversity, leading researchers to increasingly utilize synthetic data. Traditional  
 175 synthetic methods, which rely on 3D rendering engines, provide geometrically accurate annotations  
 176 but often suffer from a pronounced domain gap (Atapour-Abarghouei & Breckon, 2018; Rajpal et al.,  
 177 2023) and low generation efficiency. More recently, methods based on Generative Models have been  
 178 explored, such as using Diffusion Models (Ke et al., 2025; Gui et al., 2025) to generate depth maps  
 179 from RGB images, but these primarily focus on depth prediction rather than the active generation  
 180 of high-fidelity RGB-D data pairs for model training. To address this data bottleneck, Depth-to-  
 181 Image (D2I) generation emerges as an active data augmentation strategy, synthesizing geometrically  
 182 consistent RGB images from existing depth data. However, conventional D2I models often struggle  
 183 with precise geometric alignment, leading to artifacts. For instance, Atlantis (Zhang et al., 2024)  
 184 pioneers the use of a D2I approach but relies on triplet supervision primarily for domain-specific  
 185 underwater scenes style transfer, rather than enforcing strict, geometric consistency across diverse  
 186 domains. Our BRIDGE framework alleviates this by employing a Reinforcement Learning (RL)  
 187 optimization mechanism in its D2I engine, ensuring both visual realism and geometric consistency  
 188 for the massive generated dataset.

## 189 3 METHOD

190 To address data scarcity and label quality in monocular depth estimation, we propose a comprehen-  
 191 sive three-stage pipeline, as illustrated in Figure 2. First, we train a powerful giant teacher model on  
 192 large-scale synthetic data, and train a depth-to-image model using reinforcement learning. Second,  
 193 we generate millions of visually realistic RGB images with accurate geometric structures, highly  
 194 consistent with source depth maps, and their corresponding pseudo-labels, using the model we train  
 195 in the previous stage (Section 3.1). Then we use a simple similarity algorithm method to calculate  
 196 the original images and their corresponding generated images to create a mask for similar regions,  
 197 and use the high-precision mask area and the original gt depth map for fine-tuning (Section 3.2).  
 198 Finally, we train the MDE model on the large-scale generated data and their pseudo-labels (Sec-  
 199 tion 3.3).

### 200 3.1 REINFORCEMENT LEARNING-OPTIMIZED DEPTH-TO-IMAGE GENERATION

201 To acquire large-scale, high-quality RGB-depth image pairs, we develop and train a reinforcement  
 202 learning (RL)-optimized Depth-to-Image (D2I) generation model. This model generates visually re-  
 203 alistic RGB images from source depth maps (from existing synthetic datasets like Hypersim (Roberts  
 204 et al., 2021), TartanAir (Wang et al., 2020)) while precisely preserving their geometric structure.  
 205 This generation process results in a large-scale dataset comprising approximately 20M images.

206 **Objectives and Advantages:** Leveraging the reinforcement learning paradigm, we ensure generated  
 207 images are not only visually realistic but also geometrically accurate and consistent. This enables  
 208 synthesizing diverse RGB images from a single depth map while preserving its true geometric struc-  
 209 ture. This approach mitigates geometric artifacts or structural distortions common in traditional D2I  
 210 models and expands the diversity of effective training data.

211 **Training Strategy:** As shown in Figure 3, our D2I model, inspired by VADER (Prabhudesai et al.,  
 212 2024), is trained via reward-gradient-driven direct optimization to efficiently and accurately generate  
 213

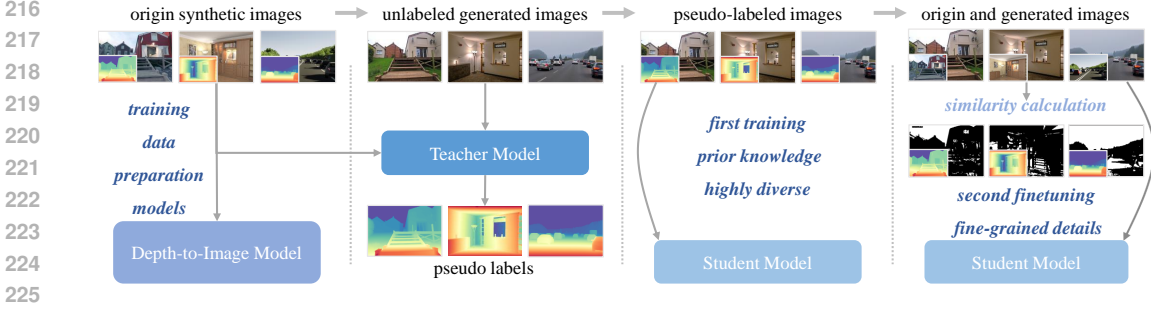


Figure 2: **BRIDGE Pipeline.** First training a depth-to-image model to synthesize millions of realistic RGB images with precise ground truth depths and a teacher model for pseudo labeling. Student model is then trained on this extensive synthetic dataset. Finally, it’s fine-tuned using mask-based refinement with original ground truth depth for robust generalization and detailed depth capture.

high-fidelity images. This direct usage of reward gradients guides the generation process towards desired visual quality, thereby avoiding complex proxy objective functions and improving training memory efficiency. The overall optimization objective minimizes the following loss:

$$L_{\text{total}}(\theta) = \lambda_{\text{depth}} \cdot L_{\text{D}}(\mathcal{M}(D_{\text{source}}, \theta), D_{\text{source}}) - \lambda_{\text{aesthetic}} \cdot R_{\text{aesthetic}}(\mathcal{M}(D_{\text{source}}, \theta)), \quad (1)$$

where  $D_{\text{source}}$  represents the input source depth map,  $\mathcal{M}(D_{\text{source}}, \theta)$  denotes the RGB image generated by the D2I model  $\mathcal{M}$  with parameters  $\theta$  based on  $D_{\text{source}}$ ,  $L_{\text{D}}(D_{\text{gen}}, D_{\text{source}})$  is a cosine similarity loss. Its corresponding loss (defined as  $1 - \text{cosine similarity}$ ) aims to minimize the difference between the depth map  $D_{\text{gen}}$  inverted from the generated RGB image and the source depth map  $D_{\text{source}}$ . This ensures the D2I model precisely learns and maintains the scene’s geometric structural information.  $R_{\text{aesthetic}}(\mathcal{M}(D_{\text{source}}, \theta))$  is the aesthetic reward function, quantifying the visual quality and aesthetic appeal of the generated image. It is formulated as:

$$R_{\text{aesthetic}}(I) = g_{\text{MLP}} \left( \frac{f_{\text{CLIP}}(I)}{\|f_{\text{CLIP}}(I)\|_2} \right), \quad (2)$$

where  $f_{\text{CLIP}}(I)$  represents the high-dimensional feature embedding extracted from the input image  $I$  by a pre-trained CLIP image encoder.  $\|\cdot\|_2$  denotes the L2 norm, normalizing the extracted features to unit length to prevent their magnitude from unduly influencing subsequent scoring and to align with the MLP’s training process. Finally,  $g_{\text{MLP}}(\cdot)$  is a specialized MLP head that receives the normalized CLIP features as input and maps them to a scalar aesthetic score  $S_{\text{aesthetic}} \in \mathbb{R}$ . During D2I model optimization, we maximize the aesthetic quality of generated images by minimizing the aesthetic loss term  $L_{\text{aesthetic}}(I) = -R_{\text{aesthetic}}(I)$ , thereby incentivizing the model to produce images with higher aesthetic scores.

The model is trained using gradient descent, with the parameter update rule:

$$\theta \leftarrow \theta - \eta \cdot \nabla_{\theta} L_{\text{total}}(\theta), \quad (3)$$

where  $\eta$  is the learning rate. The core challenge is computing the gradient of the loss function with respect to model parameters  $\theta$ ,  $\nabla_{\theta} L_{\text{total}}(\theta)$ . Since  $x_0$  is the final output of the diffusion model’s multi-step denoising process, and model parameters  $\theta$  are involved in each denoising prediction  $\epsilon_{\theta}(x_t, t, D_{\text{source}})$ , the gradient must be backpropagated through the entire reverse diffusion process. Specifically, the gradient of the loss with respect to parameters can be expressed as:

$$\nabla_{\theta} L_{\text{total}}(\theta) = \sum_{t=0}^T \frac{\partial L_{\text{total}}(x_0)}{\partial x_t} \cdot \frac{\partial x_t}{\partial \theta}, \quad (4)$$

where  $\frac{\partial L_{\text{total}}(x_0)}{\partial x_t}$  represents the sensitivity of the loss to the intermediate diffusion step  $x_t$ , and  $\frac{\partial x_t}{\partial \theta}$  denotes the influence of model parameters on the intermediate state. This gradient propagation efficiently guides the diffusion model to generate high-quality RGB images while precisely capturing and maintaining the geometric structure of the input depth map, significantly enhancing training sample and computational efficiency.

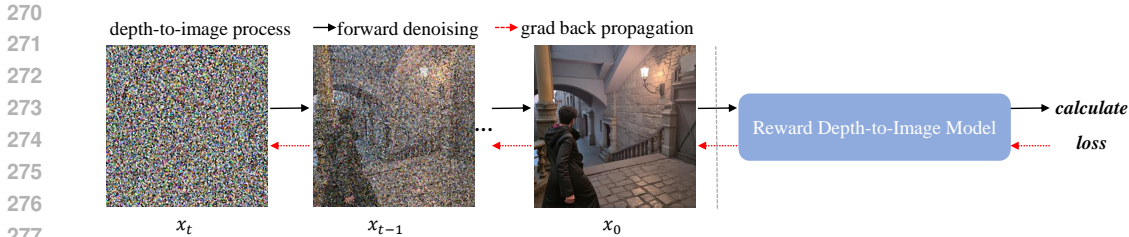


Figure 3: **Reward Model Process.** Our D2I model is trained via reward-gradient-driven direct optimization, avoiding complex proxy objective functions and improving training memory efficiency.

### 3.2 DEPTH PSEUDO-LABEL GENERATION AND MULTI-SOURCE DEPTH FUSION STRATEGY

After generating millions of high-fidelity RGB images, we construct their corresponding depth labels using a multi-strategy fusion approach. These are then combined with the original synthetic data to form the final hybrid training set.

**Teacher Model-Based Initial Pseudo-Label Generation:** We leverage a powerful teacher model trained on 1M synthetic data to infer initial relative depth pseudo-labels for all 20 million generated RGB images. These pseudo-labels provide valuable, geometrically consistent supervision, crucial for bridging the domain gap.

**Similarity-Guided Ground Truth Depth Utilization:** To achieve more precise and reliable depth labels and maximize the use of all available depth information (including original high-precision ground truth depth), we introduce a similarity-based method. This addresses pixel-level inaccuracies or detail loss from relying solely on teacher-generated pseudo-labels. Specifically, we construct a fusion mask for each generated RGB image  $I_{gen}$  by evaluating its similarity to the original synthetic RGB image  $I_{orig}$ . This mask generation involves two main steps:

- **Feature-based Registration for Structural Similarity:** First, we geometrically register the generated RGB image ( $I_{gen}$ ) with its original synthetic counterpart ( $I_{orig}$ ) using ORB feature detection and matching (Karami et al., 2017). This step compensates for minor shifts or deformations during image generation, aligning them at the pixel level. After registration, we compute the Structural Similarity Index Measure (SSIM) map (Nilsson & Akenine-Möller, 2020) between the aligned images. Regions with an SSIM value above a predefined threshold are identified as high-similarity areas, forming the first binary mask.
- **Direct SSIM-based Pixel-level Similarity:** Second, we directly compute the SSIM map between the size-adjusted generated RGB image ( $I_{gen}$ ) and the original synthetic RGB image ( $I_{orig}$ ). This evaluates their pixel-level similarity without explicit geometric registration. Regions where the SSIM value exceeds the same threshold form the second binary mask.

The final fusion mask is generated by applying a logical OR operation to these two binary masks. This ensures that any high-similarity region identified by either method is included. To enhance mask robustness and reduce noise, we further apply morphological opening and closing operations to smooth boundaries and fill small gaps.

Upon obtaining the final fusion mask, we integrate this refined depth supervision into a two-stage training process. Initially, the BRIDGE model undergoes pre-training on a large-scale dataset, primarily using the teacher-generated pseudo-labels. Following this, a fine-tuning stage is performed. In this fine-tuning stage, for regions covered by the fusion mask, we directly utilize the corresponding high-precision ground truth depth ( $D_{gt}$ ) from the original synthetic data as training labels. This strategy ensures the model first learns broad geometric consistency from vast pseudo-labeled data and then refines its accuracy and detail with the most accurate supervision in high-precision, geometrically consistent areas, thereby significantly improving depth estimation accuracy and detail.

**Final Hybrid Dataset Composition:** The composite dataset for training BRIDGE comprises two main components. The first consists of large-scale generated RGB images: approximately 20 million high-fidelity RGB images produced by the RL-optimized D2I process, each paired with its teacher-model-derived depth pseudo-label. The second involves original synthetic data depth labels: we

324 directly integrate depth labels from multiple original synthetic datasets, crucially including the high-  
325 precision ground truth depth filtered by our similarity-guided strategy.  
326

### 327 3.3 TRAINING MONOCULAR DEPTH ESTIMATION MODEL 328

329 After designing the data generation strategy, we train the Monocular Depth Estimation (MDE) model  
330 on the generated data. For model architecture, we adopt the established pretrained DINOv2-Giant  
331 encoder (Oquab et al., 2023) and DPT (Ranftl et al., 2021) head combination from Depth Any-  
332 thing (Yang et al., 2024c). To optimize our model, we utilize two loss terms, both of which are based  
333 on the approach proposed in MiDaS (Ranftl et al., 2020) because they are well-suited for affine-  
334 invariant depth. The first is the scale- and shift-invariant loss ( $L_{ssi}$ ), which ensures the model’s  
335 robustness to variations in the scale and offset of depth values. The second is the gradient matching  
336 loss ( $L_{gm}$ ), which helps the model capture fine-grained scene structure and details by enforcing sim-  
337 ilarity between the gradients of the predicted and ground-truth depth maps. We set the  $L_{ssi}$  and  $L_{gm}$   
338 as 1:4 and we ignore its top n largest loss regions during training, where n is set as 10% because the  
339 regions are usually considered as potentially noisy pseudo labels. To enable zero-shot metric depth  
340 estimation for arbitrary “in-the-wild” images, we further introduce a specialized scale head. This  
341 head autonomously predicts the image’s metric scale, significantly enhancing the model’s accuracy  
342 and generalization in real-world scenarios. We first fine-tune the metric depth model, then decouple  
343 and train the Scale Head separately. This improves training efficiency and provides high-quality,  
344 artifact-free, precise metric depth for downstream tasks like novel view synthesis, further enhancing  
345 the model’s utility.

## 346 4 IMPLEMENTATION DETAILS 347

348 Our RL-D2I generative model is based on the architecture of Flux.1 Depth[dev] (Labs, 2024) and  
349 fine-tuned using depth data from synthetic datasets, and the learning rate is  $1e-5$ . The generator syn-  
350 thesises approximately 20 million visually realistic and geometrically accurate RGB images from  
351 source depth maps by optimizing a loss function that includes depth loss (weight 0.9) and aesthetic  
352 reward (weight 0.1). To achieve high geometric accuracy and depth similarity, we ensure that the  
353 generated images meet stringent quality criteria: high-precision regions are identified by combin-  
354 ing areas where: 1) ORB feature-based registered SSIM between the generated and original RGB  
355 images exceeds 0.85 (requiring at least 10 ORB matches), and 2) direct SSIM between them also  
356 exceeds 0.85. We only select mask samples where the valid region constitutes over 50% of the pix-  
357 els, and a 3x3 erosion operation is performed to filter out excessively small regions. Subsequently, a  
358 518x518 crop is randomly extracted, centered on the largest valid region’s bounding box. For each  
359 original depth map, we generate four RGB images simultaneously, and use different random seeds.  
360 The prompt is set to None during the generation process.  
361

## 362 5 EXPERIMENTS 363

364 In this section, we will validate the effectiveness and advantages of the proposed method for monoc-  
365 ular depth estimation through a series of experiments. We first introduce the datasets, evaluation  
366 metrics, and experimental settings. Then, we present the performance of our method on multiple  
367 benchmark datasets and conduct ablation experiments to analyze the contribution of each module to  
368 the overall performance.  
369

### 370 5.1 EVALUATION DATASETS 371

372 For monocular depth estimation, we conduct a series of experiments to evaluate the performance of  
373 our model on five widely used benchmarks. NYUv2 (Silberman et al., 2012) and ScanNet (Dai et al.,  
374 2017) provide indoor RGB-D data captured by depth sensors. ETH3D (Schops et al., 2017) includes  
375 both indoor and outdoor scenes, with depth data collected via laser scanners. KITTI (Geiger et al.,  
376 2012) consists of outdoor driving scenes captured with cameras and LiDAR sensors. For ScanNet,  
377 we extract 1400 test frames. Sintel (Butler et al., 2012) is a synthetic dataset derived from animated  
short films, from which we extract images of 600 training samples for evaluation.

Table 1: **Quantitative Comparison with SOTA Methods on Zero-Shot Relative Depth Estimation.** These benchmarks aim to comprehensively measure model generalization ability and accuracy in diverse environments. While these standard benchmarks provide a direct quantitative comparison, some specific strengths of a model, such as details, complex layouts, may not be fully reflected by these benchmarks. The underline represents the best performance of SOTA models on this dataset.

		KITTI		NYUv2		ScanNet		ETH3D		Sintel		DA2K
Method		$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	Acc(%)
<i>Generative</i>	DepthFM (Gui et al., 2025)	0.932	0.085	0.956	0.065	0.947	0.068	0.958	0.069	0.700	0.577	85.8
	Marigold (Ke et al., 2024)	0.916	0.099	0.964	0.055	0.951	0.064	0.960	0.065	0.703	0.576	86.8
	GeoWizard (Fu et al., 2024)	0.921	0.097	0.966	0.052	0.953	0.061	0.961	0.064	0.705	0.574	88.1
<i>Discriminative</i>	LeReS (Yin et al., 2021)	0.784	0.149	0.916	0.090	0.917	0.091	0.777	0.171	0.590	0.785	-
	DPT (Ranftl et al., 2021)	0.881	0.111	0.919	0.091	0.932	0.084	0.929	0.115	0.605	0.700	-
	MiDaSv3.1 (Birkel et al., 2023)	0.851	0.126	0.980	0.048	0.957	0.069	0.947	0.078	0.670	0.611	-
	Metric3Dv2 (Hu et al., 2024)	<u>0.968</u>	0.062	0.963	0.058	0.941	0.074	0.960	0.066	0.662	0.619	-
	Depth Anything V2 (Yang et al., 2024d)	0.946	0.075	0.979	0.045	0.978	0.043	0.988	0.038	0.672	0.598	97.1
Depth Pro (Bochkovskii et al., 2024)	0.498	0.375	0.580	0.245	0.667	0.207	0.687	0.201	0.476	0.883	-	
<b>Ours</b>	<b>0.938</b>	<b>0.081</b>	<b>0.982</b>	<b>0.041</b>	<b>0.981</b>	<b>0.033</b>	<b>0.991</b>	<b>0.029</b>	<b>0.719</b>	<b>0.513</b>	<b>97.3</b>	

## 5.2 EVALUATION METRICS

All evaluations are carried out under the zero-shot configuration. Following previous works (Yang et al., 2024d; He et al., 2025), the evaluation of affine-invariant inverse depth prediction is performed by optimizing the scale and shift discrepancies between the estimated depth and the ground truth. For quantitative analysis, we utilize the absolute relative error (AbsRel) and  $\delta 1$  accuracy. AbsRel is calculated as  $\frac{1}{N} \sum_{k=0}^{N-1} \frac{|\hat{x}_d - x_d|}{x_d}$ , where  $N$  represents the total pixel count.  $\delta 1$  accuracy measures the percentage of pixels where the maximum ratio between the predicted affine-invariant inverse depth and the inverse true depth falls below 1.25.

## 5.3 ZERO-SHOT DEPTH ESTIMATION

Our model demonstrates exceptional zero-shot generalization in depth estimation across both indoor and outdoor datasets, highlighting its robustness, fine-grained detail, and consistent depth estimation for objects.

**Quantitative Comparisons:** Table 1 shows the results that our method achieves outstanding performance across multiple datasets. Our method achieves outstanding performance across multiple datasets and establishes new state-of-the-art (SOTA) results on several mainstream benchmarks. Specifically, our model demonstrates overwhelming superiority on indoor scene datasets like NYUv2, ScanNet, and ETH3D, with its ability to generate fine-structure predictions aligning perfectly with its objectives. Furthermore, our model slightly surpasses Depth Anything V2 on benchmarks like DA2K, further confirming its excellent zero-shot generalization capabilities. Notably, Depth Pro’s significantly lower performance in this evaluation is related to its primary optimization for metric depth rather than relative depth. Lastly, we do not achieve optimal performance on KITTI, primarily because of the dataset’s inherent sparsity. Our model is designed to capture fine-grained global and local depth information, which is not fully reflected in the KITTI evaluation.

**Qualitative Comparisons:** Figure 4 presents qualitative monocular depth estimation results, highlighting our model’s superior capability to capture fine-grained details and robustly handle challenging objects compared to Depth Anything V2, Depth Pro, and generative methods. Notably, our model successfully identifies and accurately estimates the depth of reflective surfaces, such as mirrors, on the NYUv2 dataset. It also preserves complex details, like distant table legs in ScanNet, and captures the fine textures of remote architectural and wooden structures on the Sintel dataset. Furthermore, it effectively processes similarly colored objects; for instance, on the KITTI dataset, it precisely delineates a person’s head from the background, a nuance Depth Anything V2 often struggles with. Crucially, our model exhibits excellent generalization to “in-the-wild” data, exem-

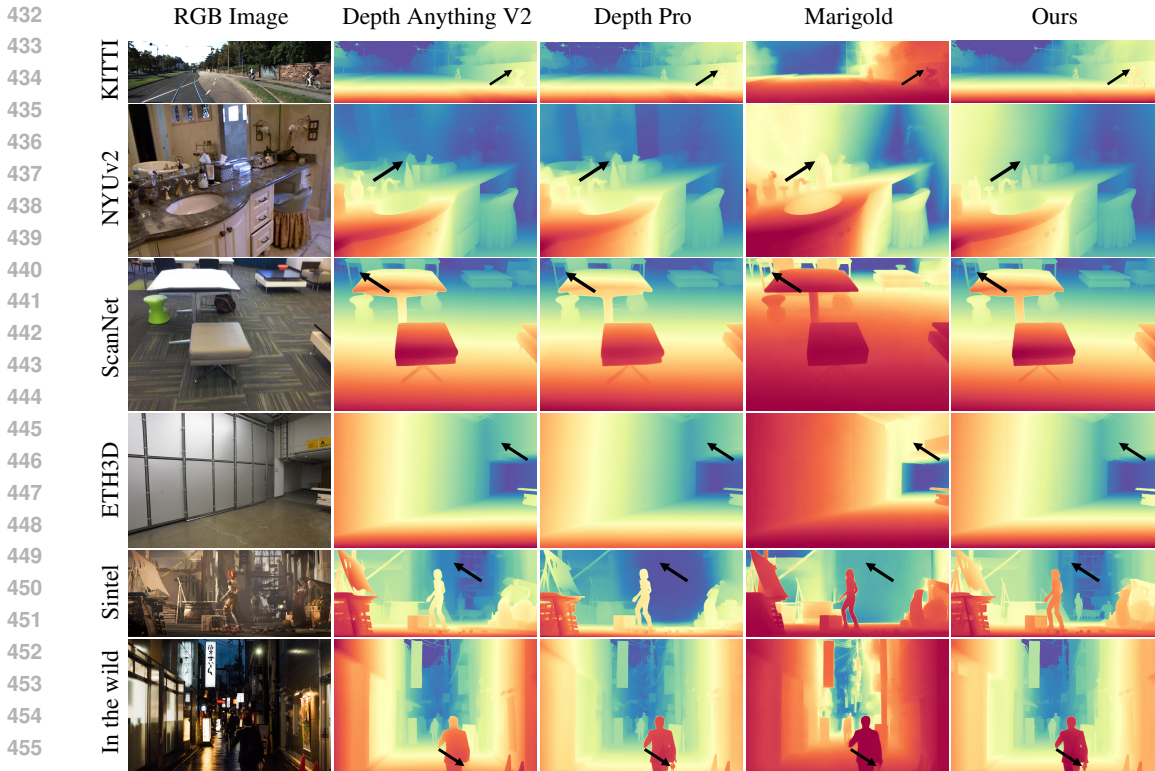


Figure 4: **Qualitative comparison of relative depth estimation across different datasets.** Our model captures fine-grained details and generalizes robustly on in-the-wild data.

plified by its accurate depth estimation for transparent umbrellas, fully demonstrating its robustness in complex real-world scenarios.

#### 5.4 FINE-TUNED TO METRIC DEPTH ESTIMATION

To validate our model’s generalization ability in metric depth estimation, we also transfer our pre-trained encoder and design an additional scale head to adapt to both indoor and outdoor domains. The scale head is designed to validate the model’s generalization ability in metric depth estimation, adapting to both indoor and outdoor domains. Structurally, it uses the DINOv2 encoder and an MLP head to predict an accurate metric scale. Following the pipeline of Depth Anything V2, our model maintains a consistently strong performance level on both NYUv2 and KITTI datasets shown in Table 4 and Table 5.

#### 5.5 ABLATION STUDIES

In this section, we systematically evaluate the contribution of each factor in BRIDGE to the overall performance, aiming to quantify the impact of our proposed innovative data generation scheme. Our ablation experiments are conducted under consistent training configurations to ensure fair comparisons.

**Impact of BRIDGE Generated Data:** We first investigate the effectiveness of our RL-D2I generation engine, which serves as the core mechanism for alleviating data scarcity and enhancing training data diversity. To demonstrate its significant impact, we compare a baseline model using DINOv2 Base trained on 1M synthetic data from various large-scale synthetic depth datasets with the BRIDGE model generated 1M data (comprising 0.95M pseudo-labels and 0.05M masked original ground truth depth). As shown in Table 2, the integration of RL-D2I-generated data significantly boosts the performance across all evaluated metrics. This improvement underscores the effectiveness of our RL-D2I engine in generating visually realistic and geometrically accurate images, which effectively expands the scale and diversity of the training data.

Table 2: **Impact of Generated Data.** We demonstrate the effectiveness of our RL-D2I engine by comparing a model trained on standard synthetic data with one trained on 1M generated data with ViT-L. The results show that our generated data boosts performance across all metrics.

	KITTI		NYUv2		ScanNet		ETH3D		Sintel	
Method	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$
Baseline	0.921	0.097	0.960	0.060	0.964	0.057	0.969	0.060	0.599	0.709
<b>w/ Generated Data</b>	<b>0.929</b>	<b>0.092</b>	<b>0.964</b>	<b>0.056</b>	<b>0.968</b>	<b>0.048</b>	<b>0.970</b>	<b>0.059</b>	<b>0.612</b>	<b>0.692</b>

Table 3: **Ablation Study of the Hybrid Training Strategy.** We compares the performance of BRIDGE’s ViT-L model, trained solely on pseudo-labels with hybrid supervision integrating similarity-guided high-precision ground truth depth on benchmarks. The results highlight the importance of incorporating original ground truth depth for improving depth estimation accuracy and detail capture.

	KITTI		NYUv2		ScanNet		ETH3D		Sintel	
Method	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$
Pseudo-labels Only	0.924	0.099	0.958	0.069	0.957	0.069	0.949	0.078	0.595	0.751
<b>Full Hybrid Strategy</b>	<b>0.926</b>	<b>0.098</b>	<b>0.960</b>	<b>0.067</b>	<b>0.960</b>	<b>0.061</b>	<b>0.953</b>	<b>0.073</b>	<b>0.600</b>	<b>0.725</b>

Table 4: The metric results on NYUv2.

Method	Higher $\uparrow$		Lower $\downarrow$	
	$\delta_1$	AbsRel	RMSE	
AdaBins (Bhat et al., 2021)	0.903	0.103	0.364	
DPT (Ranftl et al., 2021)	0.904	0.110	0.357	
P3Depth (Patil et al., 2022)	0.898	0.104	0.356	
SwinV2 (Liu et al., 2022)	0.949	0.083	0.287	
IEBins (Shao et al., 2023b)	0.936	0.087	0.314	
ZoeDepth (Bhat et al., 2023)	0.951	0.077	0.282	
Metric3Dv2 (Hu et al., 2024)	0.987	0.046	0.181	
<b>Ours</b>	<b>0.986</b>	<b>0.052</b>	<b>0.197</b>	

Table 5: The metric results on KITTI.

Method	Higher $\uparrow$		Lower $\downarrow$	
	$\delta_1$	AbsRel	RMSE	
AdaBins (Bhat et al., 2021)	0.964	0.058	2.360	
P3Depth (Patil et al., 2022)	0.953	0.071	2.842	
SwinV2 (Liu et al., 2022)	0.977	0.050	1.966	
GEDepth (Yang et al., 2023b)	0.976	0.048	2.044	
IEBins (Shao et al., 2023b)	0.978	0.050	2.011	
ZoeDepth (Bhat et al., 2023)	0.971	0.054	2.281	
Metric3Dv2 (Hu et al., 2024)	0.985	0.044	1.993	
<b>Ours</b>	<b>0.983</b>	<b>0.045</b>	<b>1.862</b>	

**Efficiency of Hybrid Training Strategy:** As shown in Table 3, our hybrid depth supervision strategy is crucial for combining the broad coverage of teacher model pseudo-labels with the precision of similarity-guided high-precision ground truth depth. This approach aims to provide more robust and reliable supervision signals for learning geometric structural knowledge. We conduct an ablation study to quantify the benefits of this dual supervision. We also test on 1M data, the BRIDGE model trained with our full hybrid strategy yields superior results compared to a variant that relies solely on teacher-generated pseudo-labels for the generated RGB images.

## 6 CONCLUSION

In this work, we introduce BRIDGE, an innovative Monocular Depth Estimation framework that addresses data scarcity and quality issues. This is achieved by leveraging an RL-optimized Depth-to-Image (RL-D2I) engine to generate over 20M RGB-D data, which, along with a hybrid strategy of teacher pseudo-labels and high-precision ground truth, achieves state-of-the-art performance, consistently outperforming methods like Depth Anything V2 on benchmarks with significantly less training data (20M vs. 62M). Our findings highlight BRIDGE’s superior capability in capturing fine-grained details, ensuring geometric consistency, and demonstrating robust zero-shot generalization to complex in-the-wild scenes, paving the way for more efficient and generalizable MDE solutions.

## REFERENCES

- 540  
541  
542 Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using syn-  
543 thetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE conference*  
544 *on computer vision and pattern recognition*, pp. 2800–2810, 2018.
- 545 Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adap-  
546 tive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,  
547 pp. 4009–4018, 2021.
- 548 Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-  
549 shot transfer by combining relative and metric depth, 2023. URL [https://arxiv.org/](https://arxiv.org/abs/2302.12288)  
550 [abs/2302.12288](https://arxiv.org/abs/2302.12288).
- 551 Reiner Birkel, Diana Wofk, and Matthias Müller. Midas v3. 1—a model zoo for robust monocular  
552 relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.
- 553 Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R  
554 Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second.  
555 *arXiv preprint arXiv:2410.02073*, 2024.
- 556 Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source  
557 movie for optical flow evaluation. In *European conference on computer vision*, pp. 611–625.  
558 Springer, 2012.
- 559 Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint*  
560 *arXiv:2001.10773*, 2020.
- 561 David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaus-  
562 sian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the*  
563 *IEEE/CVF conference on computer vision and pattern recognition*, pp. 19457–19467, 2024.
- 564 Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild.  
565 *Advances in neural information processing systems*, 29, 2016.
- 566 Alex Costanzino, Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi  
567 Di Stefano. Learning depth estimation for transparent and mirror surfaces. In *Proceedings of the*  
568 *IEEE/CVF International Conference on Computer Vision*, pp. 9244–9255, 2023.
- 569 Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias  
570 Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer*  
571 *Vision and Pattern Recognition (CVPR), IEEE*, 2017.
- 572 Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas  
573 Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An  
574 image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint*  
575 *arXiv:2010.11929*, 2020.
- 576 David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common  
577 multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on*  
578 *computer vision*, pp. 2650–2658, 2015.
- 579 David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using  
580 a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.
- 581 Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal  
582 regression network for monocular depth estimation. In *Proceedings of the IEEE conference on*  
583 *computer vision and pattern recognition*, pp. 2002–2011, 2018.
- 584 Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and  
585 Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a  
586 single image. In *European Conference on Computer Vision*, pp. 241–258. Springer, 2024.

- 594 Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In  
595 *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.
- 596
- 597 Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti  
598 vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*,  
599 pp. 3354–3361. IEEE, 2012.
- 600
- 601 Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estima-  
602 tion with left-right consistency. In *Proceedings of the IEEE conference on computer vision and*  
603 *pattern recognition*, pp. 270–279, 2017.
- 604
- 605 Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga  
606 Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast gen-  
607 erative monocular depth estimation with flow matching. In *Proceedings of the AAAI Conference*  
*on Artificial Intelligence*, volume 39, pp. 3203–3211, 2025.
- 608
- 609 Xiankang He, Dongyan Guo, Hongji Li, RuiBo Li, Ying Cui, and Chi Zhang. Distill any depth:  
610 Distillation creates a stronger monocular depth estimator. *arXiv preprint arXiv:2502.19204*, 2025.
- 611
- 612 Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in*  
*neural information processing systems*, 33:6840–6851, 2020.
- 613
- 614 Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu,  
615 Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation  
616 model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern*  
*Analysis and Machine Intelligence*, 2024.
- 617
- 618 Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li,  
619 and Ping Luo. Ddp: Diffusion model for dense visual prediction. In *Proceedings of the IEEE/CVF*  
620 *International Conference on Computer Vision*, pp. 21741–21752, 2023.
- 621
- 622 Ebrahim Karami, Siva Prasad, and Mohamed Shehata. Image matching using sift, surf, brief and  
623 orb: performance comparison for distorted images. *arXiv preprint arXiv:1710.02726*, 2017.
- 624
- 625 Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad  
626 Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In  
627 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9492–  
9502, 2024.
- 628
- 629 Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov,  
630 and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for  
631 image analysis, 2025.
- 632
- 633 Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splat-  
634 ting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.
- 635
- 636 Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny im-  
637 ages.(2009), 2009.
- 638
- 639 Yevhen Kuznetsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular  
640 depth map prediction. In *Proceedings of the IEEE conference on computer vision and pattern*  
*recognition*, pp. 6647–6655, 2017.
- 641
- 642 Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- 643
- 644 Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for  
645 deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3,  
pp. 896. Atlanta, 2013.
- 646
- 647 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*, pp. 10012–10022, 2021.

- 648 Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng  
649 Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of*  
650 *the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12009–12019, 2022.
- 651
- 652 Lukas Mehl, Jenny Schmalfluss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A  
653 high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Pro-*  
654 *ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4981–  
655 4991, 2023.
- 656 Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and  
657 Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications*  
658 *of the ACM*, 65(1):99–106, 2021.
- 659 Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models.  
660 In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.
- 661
- 662 Jim Nilsson and Tomas Akenine-Möller. Understanding ssim. *arXiv preprint arXiv:2006.13846*,  
663 2020.
- 664
- 665 Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov,  
666 Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning  
667 robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- 668 Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3depth: Monocular  
669 depth estimation with a piecewise planarity prior. In *Proceedings of the IEEE/CVF Conference*  
670 *on Computer Vision and Pattern Recognition*, pp. 1610–1621, 2022.
- 671
- 672 Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and  
673 Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the*  
674 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10106–10116, 2024.
- 675 Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak.  
676 Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.
- 677
- 678 Aakash Rajpal, Noshaba Cheema, Klaus Illgner-Fehns, Philipp Slusallek, and Sunil Jaiswal. High-  
679 resolution synthetic rgb-d datasets for monocular depth estimation. In *Proceedings of the*  
680 *IEEE/CVF conference on computer vision and pattern recognition*, pp. 1188–1198, 2023.
- 681 René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust  
682 monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transac-*  
683 *tions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.
- 684 René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction.  
685 In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188,  
686 2021.
- 687
- 688 Alex Rasla and Michael Beyeler. The relative importance of depth cues and semantic edges for  
689 indoor mobility using simulated prosthetic vision in immersive virtual reality. In *Proceedings of*  
690 *the 28th ACM symposium on virtual reality software and technology*, pp. 1–11, 2022.
- 691 Haoyu Ren, Aman Raj, Mostafa El-Khamy, and Jungwon Lee. Suw-learn: Joint supervised, unsu-  
692 pervised, weakly supervised deep learning for monocular depth estimation. In *Proceedings of the*  
693 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 750–751,  
694 2020.
- 695
- 696 Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan  
697 Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for  
698 holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference*  
699 *on computer vision*, pp. 10912–10922, 2021.
- 700 Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single  
701 still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840,  
2008.

- 702 Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc  
703 Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and  
704 multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern  
705 recognition*, pp. 3260–3269, 2017.
- 706 Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, and Zhengguo Li. Ndepth: Normal-  
707 distance assisted monocular depth estimation. In *Proceedings of the IEEE/CVF International  
708 Conference on Computer Vision*, pp. 7931–7940, 2023a.
- 709 Shuwei Shao, Zhongcai Pei, Xingming Wu, Zhong Liu, Weihai Chen, and Zhengguo Li. Iebins:  
710 Iterative elastic bins for monocular depth estimation. *Advances in Neural Information Processing  
711 Systems*, 36:53025–53037, 2023b.
- 712 Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and sup-  
713 port inference from rgb-d images. In *European conference on computer vision*, pp. 746–760.  
714 Springer, 2012.
- 715 Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel,  
716 Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised  
717 learning with consistency and confidence. *Advances in neural information processing systems*,  
718 33:596–608, 2020.
- 719 Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv  
720 preprint arXiv:2010.02502*, 2020.
- 721 Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs:  
722 A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface  
723 normal estimation. *arXiv preprint arXiv:1912.09678*, 2019a.
- 724 Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang.  
725 Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal  
726 training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*,  
727 pp. 5261–5271, 2025.
- 728 Wenshan Wang, DeLong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu,  
729 Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In  
730 *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–  
731 4916. IEEE, 2020.
- 732 Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Wein-  
733 berger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for  
734 autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern  
735 recognition*, pp. 8445–8453, 2019b.
- 736 Diana Wofk, Fangchang Ma, Tien-Ju Yang, Sertac Karaman, and Vivienne Sze. Fastdepth: Fast  
737 monocular depth estimation on embedded systems. In *2019 International Conference on Robotics  
738 and Automation (ICRA)*, pp. 6101–6108. IEEE, 2019.
- 739 Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation  
740 for consistency training. *Advances in neural information processing systems*, 33:6256–6268,  
741 2020.
- 742 Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin,  
743 Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. *arXiv preprint  
744 arXiv:2410.10815*, 2024a.
- 745 Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin,  
746 Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. *arXiv preprint  
747 arXiv:2410.10815*, 2024b.
- 748 Lihe Yang, Zhen Zhao, Lei Qi, Yu Qiao, Yinghuan Shi, and Hengshuang Zhao. Shrinking class space  
749 for enhanced certainty in semi-supervised learning. In *Proceedings of the IEEE/CVF international  
750 conference on computer vision*, pp. 16187–16196, 2023a.

- 756 Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth  
757 anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF*  
758 *conference on computer vision and pattern recognition*, pp. 10371–10381, 2024c.
- 759  
760 Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang  
761 Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911,  
762 2024d.
- 763 Xiaodong Yang, Zhuang Ma, Zhiyu Ji, and Zhe Ren. Gedepth: Ground embedding for monocular  
764 depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*,  
765 pp. 12719–12727, 2023b.
- 766 Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan.  
767 Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of*  
768 *the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1790–1799, 2020.
- 769  
770 Chongjie Ye, Yinyu Nie, Jiahao Chang, Yuantao Chen, Yihao Zhi, and Xiaoguang Han. Gaustudio:  
771 A modular framework for 3d gaussian splatting and beyond. *arXiv preprint arXiv:2403.19632*,  
772 2024.
- 773 Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua  
774 Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF*  
775 *conference on computer vision and pattern recognition*, pp. 204–213, 2021.
- 776  
777 Fan Zhang, Shaodi You, Yu Li, and Ying Fu. Atlantis: Enabling underwater depth estimation with  
778 stable diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern*  
779 *recognition*, pp. 11852–11861, 2024.
- 780 Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image  
781 diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*,  
782 pp. 3836–3847, 2023.
- 783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809

## A EXPERIMENTS

### A.1 MODEL SELECTION

We compare the learning capabilities of different DINOv2 (Oquab et al., 2023) versions on our hybrid data. For DINOv2-Giant and Large, we select the non-register version as their representative. As shown in Table 6, we find that the Giant and Large versions of DINOv2 exhibit excellent learning ability on synthetic data and generalize remarkably well to real-world test datasets. In contrast, the Small and Base versions do not achieve satisfactory results. This might be attributed to their limited model capacity, which is insufficient to capture the rich fine-grained information and diversity present in generation data, thus leading to difficulties in generalizing to more complex real-world scenarios. Therefore, we ultimately choose to adopt the DINOv2-Giant for our model, to fully leverage its powerful generalization capabilities and accuracy.

Method	KITTI		NYUv2		ScanNet		ETH3D		Sintel	
	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$
DINOv2-S	0.928	0.082	0.964	0.062	0.921	0.085	0.978	0.045	0.637	0.684
DINOv2-B	0.932	0.081	0.968	0.054	0.941	0.065	0.980	0.039	0.647	0.644
DINOv2-L	0.937	0.081	0.978	0.045	0.972	0.049	0.989	0.032	0.688	0.588
DINOv2-G	<b>0.938</b>	<b>0.081</b>	<b>0.982</b>	<b>0.041</b>	<b>0.981</b>	<b>0.033</b>	<b>0.991</b>	<b>0.029</b>	<b>0.719</b>	<b>0.513</b>

Table 6: Comparison among various size pre-trained DINOv2 encoders trained on our dataset.

Table 7: **Our training sources.** \* represents the quantity sampled from the dataset.

Dataset	Indoor	Outdoor	# Images
<b>Precise Synthetic Images (1M)</b>			
BlendedMVS (Yao et al., 2020)	✓	✓	115K
TartanAir (Wang et al., 2020)	✓	✓	306K
Hypersim (Roberts et al., 2021)	✓		60K
IRS (Wang et al., 2019a)	✓		103K
VKITTI 2 (Cabon et al., 2020)		✓	20K
Spring (Mehl et al., 2023)		✓	5K
DA-V (Yang et al., 2024a)	✓	✓	400K*
<b>Generated Images (20M)</b>			
BlendedMVS	✓	✓	2.2M
TartanAir		✓	2M
Hypersim	✓	✓	2.4M
IRS	✓	✓	4.2M
VKITTI 2		✓	1M
Spring		✓	20K
DA-V	✓	✓	8.2M

### A.2 ADDITIONAL QUALITATIVE COMPARISON

**Training Datasets:** As shown in Table 7, we train both a generative model and a teacher model using seven synthetic datasets to enhance label precision. To effectively mitigate the limitations of synthetic images regarding distribution shift and limited diversity, we generate 20M synthetic data based on the aforementioned datasets, thereby significantly broadening the diversity of the overall samples. Notably, for datasets such as Hypersim, where the original data primarily covers indoor

864 scenes, our generation strategy successfully produces outdoor data largely consistent with depth  
865 maps, effectively expanding the scene coverage of the training samples.  
866

867 **Qualitative Results:**Figure 5 and Figure 6 show additional qualitative comparisons with other state-  
868 of-the-art monocular depth estimation methods. For indoor, outdoor, and non-real scenes alike, our  
869 model consistently produces depth maps with higher fidelity than the other methods. Our method  
870 particularly excels in outdoor scenes, showing clear distinctions between distant objects and the sky,  
871 while Depth Pro (Bochkovskii et al., 2024) and Marigold (Ke et al., 2025) fail to distinguish objects  
872 that are farther away.

873 **Conditional Synthesis:** Figure 7 shows a comparison of depth-conditioned synthesis results. We  
874 first infer depth based on an image using our BRIDGE and the Depth Anything V2 model, and then  
875 use the most basic pre-trained depth-to-image ControlNet (Zhang, Rao, and Agrawala 2023) with  
876 Stable Diffusion 1.5 to synthesize new samples based on the depth maps and a text prompt. We can  
877 clearly observe that the depth maps obtained with Depth Anything V2 do not reflect the actual depth  
878 well and are additionally inaccurate for some parts of the image. In contrast, our method yields  
879 sharp and realistic depth maps. This result is also reflected in the synthesized results, where images  
880 created based on our depth map more closely resemble the actual image.  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

918  
 919  
 920  
 921  
 922  
 923  
 924  
 925  
 926  
 927  
 928  
 929  
 930  
 931  
 932  
 933  
 934  
 935  
 936  
 937  
 938  
 939  
 940  
 941  
 942  
 943  
 944  
 945  
 946  
 947  
 948  
 949  
 950  
 951  
 952  
 953  
 954  
 955  
 956  
 957  
 958  
 959  
 960  
 961  
 962  
 963  
 964  
 965  
 966  
 967  
 968  
 969  
 970  
 971

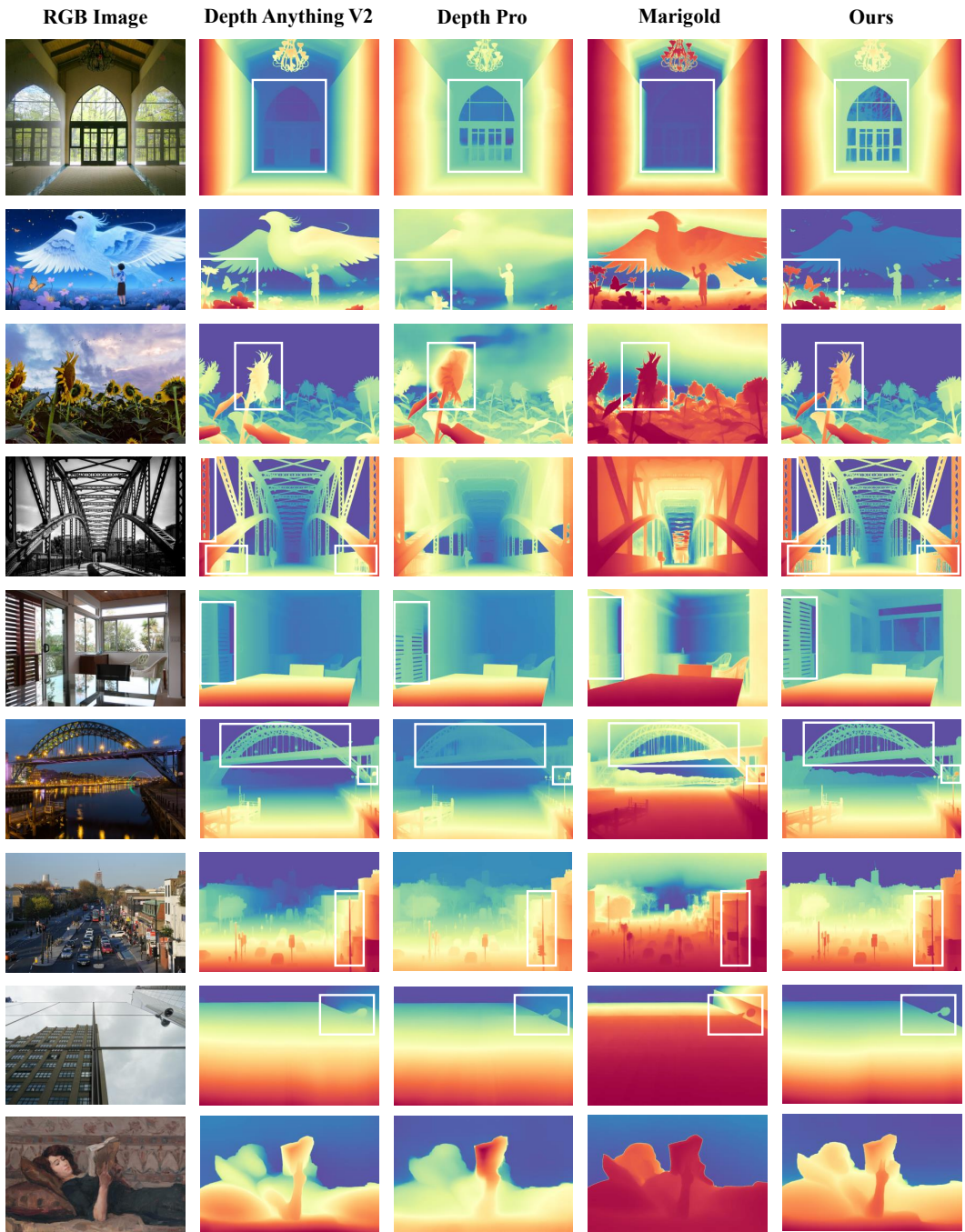


Figure 5: Comparison between our model and leading methods, including Depth Anything V2 (Yang et al., 2024d), Depth Pro(Bochkovskii et al., 2024) and Marigold(Ke et al., 2025) on open-world images.

972  
 973  
 974  
 975  
 976  
 977  
 978  
 979  
 980  
 981  
 982  
 983  
 984  
 985  
 986  
 987  
 988  
 989  
 990  
 991  
 992  
 993  
 994  
 995  
 996  
 997  
 998  
 999  
 1000  
 1001  
 1002  
 1003  
 1004  
 1005  
 1006  
 1007  
 1008  
 1009  
 1010  
 1011  
 1012  
 1013  
 1014  
 1015  
 1016  
 1017  
 1018  
 1019  
 1020  
 1021  
 1022  
 1023  
 1024  
 1025

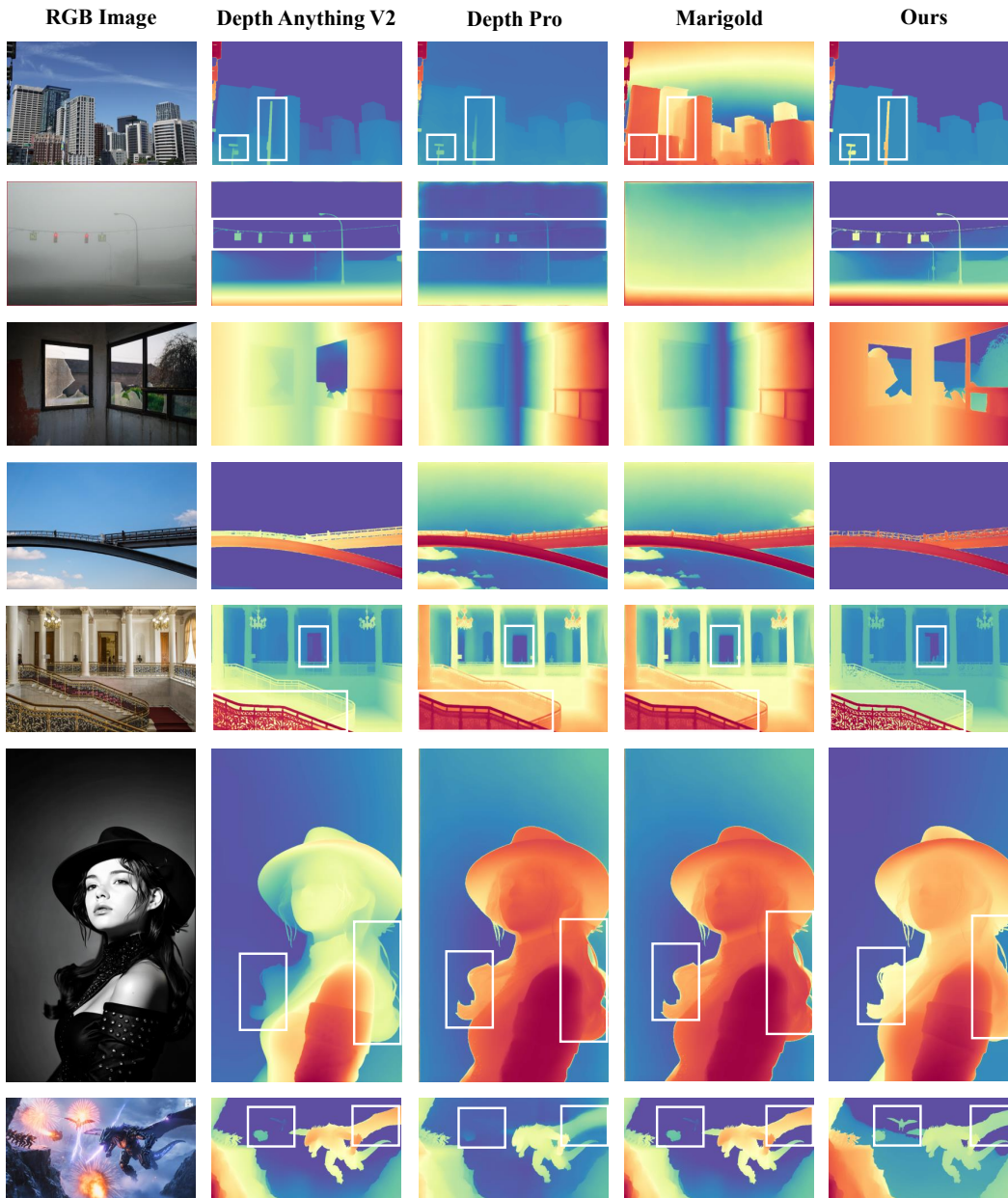


Figure 6: Additional comparison between our model and leading methods, including Depth Anything V2 (Yang et al., 2024d), Depth Pro(Bochkovskii et al., 2024) and Marigold(Ke et al., 2025) on “in-the-wild” images.

1026  
 1027  
 1028  
 1029  
 1030  
 1031  
 1032  
 1033  
 1034  
 1035  
 1036  
 1037  
 1038  
 1039  
 1040  
 1041  
 1042  
 1043  
 1044  
 1045  
 1046  
 1047  
 1048  
 1049  
 1050  
 1051  
 1052  
 1053  
 1054  
 1055  
 1056  
 1057  
 1058  
 1059  
 1060  
 1061  
 1062  
 1063  
 1064  
 1065  
 1066  
 1067  
 1068  
 1069  
 1070  
 1071  
 1072  
 1073  
 1074  
 1075  
 1076  
 1077  
 1078  
 1079

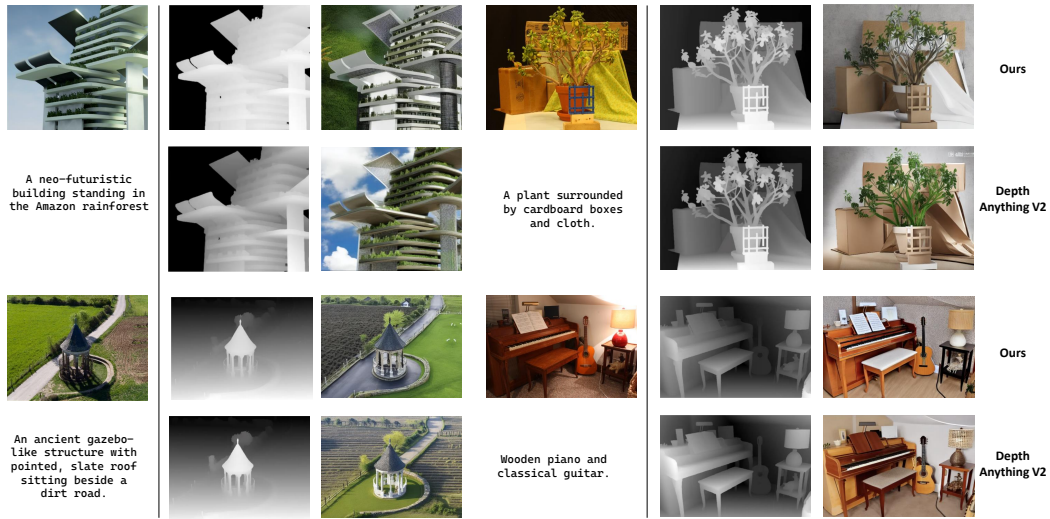


Figure 7: Our model, BRIDGE, generates superior, high-fidelity depth maps that enable ControlNet (Zhang et al., 2023) to synthesize new images with a zero-shot capability, precisely replicating the depth field of the source image. In contrast, Depth Anything V2 (Yang et al., 2024d) struggles to produce an accurate depth field, as demonstrated by the clear discrepancies between its corresponding ControlNet output and the source images. The prompts used for ControlNet are displayed in the lower left corners, and all images were generated with the same random seed.



Figure 8: **Comparison between Standard SFT and RL-optimized fine-tuning.** We show sample generated images. Our RL-optimized variants enable the model to capture finer details than the Standard SFT baseline.