BRIDGE: BUILDING REWARD OPTIMIZED DEPTH-TO-IMAGE DATA GENERATION ENGINE FOR MONOCULAR DEPTH ESTIMATION

Anonymous authors
Paper under double-blind review



Figure 1: BRIDGE uses a reward-optimized Depth-to-Image (D2I) engine to synthesize realistic, geometrically consistent RGB images from source depth maps. An MDE model trained on this generated RGB-D data achieves strong depth prediction across complex real-world scenes.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ABSTRACT

Monocular Depth Estimation (MDE) recovers scene geometry from a single RGB image, yet what a model can learn is capped less by its architecture than by the depth supervision available: sensor-captured RGB-D is metrically accurate but sparse, rendered synthetic RGB-D is dense but domain-shifted, and web-scale pseudo-labels cover broad appearance but inherit teacher boundary noise. We propose BRIDGE, a reward-optimized depth-to-image (D2I) engine that synthesizes over 20M RGB images from diverse source depth maps, each paired with the source depth map used to condition it. We then train a depth model on this data with similarity-guided hybrid supervision: source depth supervises regions where the generated image preserved geometry, and teacher pseudo-labels supervise where it drifted. BRIDGE achieves the strongest zero-shot relative-depth results among compared methods on NYUv2, ScanNet, ETH3D, and Sintel, using 20M generated images against Depth Anything V2's 62M.

1 INTRODUCTION

Monocular Depth Estimation (MDE) recovers scene geometry from a single RGB image, and it underpins 3D reconstruction (Kerbl et al., 2023), autonomous driving (Wang et al., 2019b), and robotics (Wofk et al., 2019). A single image, however, fixes scene shape only up to an unknown scale, which makes the depth supervision a model is trained on especially load-bearing. Recent progress (Hu et al., 2024; Yang et al., 2024b;c) has largely come from scaling that supervision, yet the supervision itself is where the field now hits a ceiling.

Depth supervision today comes from three regimes, each capped by a different failure. Sensor-captured real RGB-D (Bhat et al., 2023; Hu et al., 2024) is metrically accurate but sparse, and it often fails on challenging surfaces such as thin structures and transparent or reflective objects (Costanzino et al., 2023). Rendered synthetic RGB-D (Rajpal et al., 2023; Roberts et al., 2021) is dense and pixel-perfect but carries a domain gap that erodes generalization (Ganin & Lempitsky, 2015). Web-scale pseudo-labeling, pioneered by Depth Anything (Yang et al., 2024b;c), buys broad appearance coverage from unlabeled real images but inherits the teacher's boundary errors and fine-detail noise. No single regime delivers both appearance diversity and dense, accurate geometry.

This gap suggests turning the problem around. If dense, accurate geometry is easier to obtain than diverse real appearance, then rather than collecting more labeled RGB we can start from rendered depth maps with known scene geometry and synthesize diverse RGB appearance onto them. Depth-to-image (D2I) generation makes this possible, but off-the-shelf D2I models drift geometrically: the image looks plausible while quietly disagreeing with the depth that conditioned it, so the generated pair is no longer a valid RGB-D sample.

We propose BRIDGE, a reward-optimized D2I data engine that reduces this drift. BRIDGE fine-tunes a depth-conditioned generator by backpropagating a geometric-consistency reward, the agreement between a frozen depth estimator's prediction on the generated image and the conditioning depth, together with a visual-quality reward, directly through the sampling process, so its images stay aligned with their source depth while spanning diverse appearance. Running the engine over existing synthetic depth maps yields roughly 20M RGB-D pairs.

The generated pair is trustworthy only where the generator preserved the input geometry, so we supervise with a division of labor keyed to fidelity. We compute SSIM between each generated image and its original synthetic counterpart; where they stay similar, the geometry is preserved and the source depth supervises those pixels directly. Where the image has drifted, the source depth no longer describes what the pixels show, so we fall back to a teacher model's pseudo-label read from the generated image itself, which stays consistent with what the student sees. Source depth thus gives precision where geometry holds, and the teacher gives coverage everywhere else.

Our main contributions are as follows:

- We propose a reward-optimized depth-to-image engine that backpropagates a geometric-consistency reward and a visual-quality reward through the diffusion sampling process,

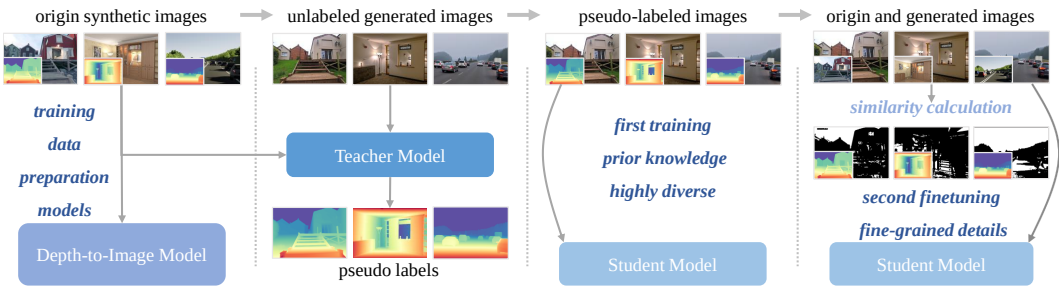


Figure 2: **BRIDGE Pipeline.** First training a depth-to-image model to synthesize millions of realistic RGB images with precise ground truth depths and a teacher model for pseudo labeling. Student model is then trained on this extensive synthetic dataset. Finally, it's fine-tuned using mask-based refinement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

generating ∼20M RGB-D pairs whose appearance is diverse and whose geometry is verified against the source depth.

- We propose similarity-guided hybrid supervision: an SSIM mask supervises with source depth where the generated image preserves geometry and falls back to teacher pseudo-labels where it drifts, so precision and coverage are combined rather than traded off.
- We show that a student trained on this data attains the strongest zero-shot δ_1/AbsRel among compared methods on NYUv2, ScanNet, ETH3D, and Sintel, using 20M generated images against Depth Anything V2's 62M.

2 METHOD

To address data scarcity and label quality in monocular depth estimation, we propose a comprehensive three-stage pipeline, as illustrated in Figure 2. First, we train a powerful giant teacher model on large-scale synthetic data, and train a depth-to-image model using reinforcement learning. Second, we generate millions of visually realistic RGB images with accurate geometric structures, highly consistent with source depth maps, and their corresponding pseudo-labels, using the model we train in the previous stage. Then we use a simple similarity algorithm method to calculate the original images and their corresponding generated images to create a mask for similar regions, and use the high-precision mask area and the original gt depth map for fine-tuning. Finally, we train the MDE model on the large-scale generated data together with their pseudo-labels, yielding a student with strong zero-shot generalization and fine-grained detail.

2.1 REWARD-OPTIMIZED DEPTH-TO-IMAGE GENERATION

To acquire large-scale, high-quality RGB-depth image pairs, we develop and train a reward-optimized Depth-to-Image (D2I) generation model. This model generates visually realistic RGB images from source depth maps (from existing synthetic datasets like Hypersim (Roberts et al., 2021), TartanAir (Wang et al., 2020)) while precisely preserving their geometric structure. Generation is depth-conditioned at a fixed resolution and preserves the scene layout, so the source depth map stays pixel-aligned with the generated image and remains a valid supervision target for the regions where geometry is retained. This generation process results in a large-scale dataset comprising approximately 20M images.

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

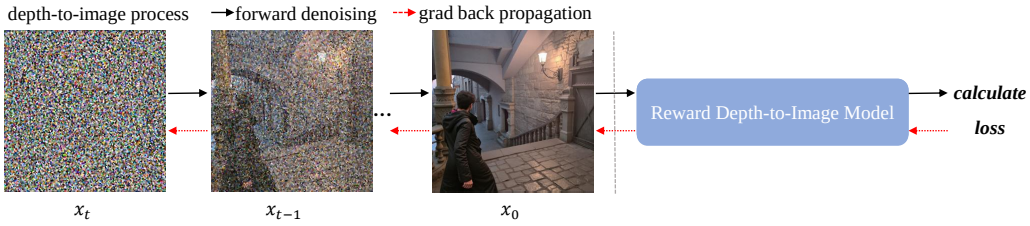


Figure 3: **Reward Model Process.** Our D2I model is trained via reward-gradient-driven direct optimization, avoiding complex proxy objective functions and improving training memory efficiency.

Training Strategy: As shown in Figure 3, following VADER (Prabhudesai et al., 2024), we optimize the generator by backpropagating reward gradients directly through the diffusion sampling process, rather than reweighting samples by a scalar reward. Differentiating the reward directly avoids the high-variance policy-gradient estimators used in RL-style fine-tuning and lets a few sampling steps carry a clean signal, at the cost of storing activations along the sampled trajectory. Let $I_{\text{gen}} = \mathcal{M}_\theta(D_{\text{source}})$ denote the generated RGB image and $D_{\text{gen}} = F_\phi(I_{\text{gen}})$ the depth map predicted by a frozen depth estimator F_ϕ. The overall optimization objective minimizes the following loss:

$$\begin{aligned} L_{\text{total}}(\theta) = \lambda_{\text{depth}} \cdot L_{\text{D}}(F_\phi(I_{\text{gen}}), D_{\text{source}}) \\ - \lambda_{\text{aesthetic}} \cdot R_{\text{aesthetic}}(I_{\text{gen}})\,. \end{aligned} \tag{1}$$

where D_{source} is the input source depth map, and θ and ϕ are the generator and fixed depth-estimator parameters, respectively. $L_{\text{D}}(D_{\text{gen}}, D_{\text{source}})$ is one minus the cosine similarity between the predicted and source depth maps. The depth estimator F_ϕ is frozen and kept independent of both the teacher model used for pseudo-labeling and the models we compare against, so the reward does not optimize toward an evaluated baseline. This ensures the D2I model precisely learns and maintains the scene's geometric structural information. $R_{\text{aesthetic}}(I_{\text{gen}})$ is the aesthetic reward function, quantifying the visual quality and aesthetic appeal of the generated image. It is formulated as:

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

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 and overall visual quality.

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

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

where η is the learning rate. The core challenge is computing the gradient of the loss function with respect to model parameters θ, $\nabla_\theta L_{\text{total}}(\theta)$. Since x_0 is the final output of the diffusion model's multi-step denoising process, and model parameters θ 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:

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

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. Propagating gradients through the full trajectory in this way lets the generator learn to preserve the geometric structure of the input depth map, at the cost of retaining the activations along that trajectory.

Backpropagating through Flux.1 Depth [dev]'s full 20-step sampling trajectory is memory-bound: retaining the DiT activations of every denoising step exceeds the 80 GB budget of a single A800. We keep training tractable with two choices. First, we adapt the generator with LoRA on the DiT attention projections (rank 16) rather than full fine-tuning, so only the adapter parameters are updated and their optimizer state and gradients stay small. Second, we apply gradient checkpointing across sampling steps, trading recomputation for activation memory. Under this setup the generator trains on 64×A800 GPUs at batch size 128 for 1,500 steps at learning rate 1e-5, and the mean reward rises from 5.07 to 6.59.

2.2 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 comprehensive 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 effectively 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}), evaluating their pixel-level similarity without explicit geometric registration. Regions where the SSIM value exceeds the same threshold form the second binary mask, capturing similar areas that feature-based registration alone may miss.

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 in the resulting mask.

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 reward-optimized D2I process, each paired with its teacher-model-derived depth pseudo-label. The second involves original synthetic depth labels: we directly integrate the high-precision ground truth depth from the source datasets, selected by our similarity-guided strategy.

2.3 Training Monocular Depth Estimation Model

After designing the data generation strategy, we train the Monocular Depth Estimation (MDE) model on the generated data. For model architecture, we adopt the established pretrained DINOv2-Giant encoder (Oquab et al., 2023) and DPT (Ranftl et al., 2021) head combination from Depth Anything (Yang et al., 2024b). To optimize our model, we utilize two loss terms, both of which are based on the approach proposed in MiDaS (Ranftl et al., 2020) because they are well-suited for affine-invariant depth. The first is the scale- and shift-invariant loss (L_{ssi}), which ensures the model's robustness to variations in the scale and offset of depth values. The second is the gradient matching loss (L_{gm}), which helps the model capture fine-grained scene structure and details by enforcing similarity between the gradients of the predicted and ground-truth depth maps. We set the L_{ssi} and L_{gm} as 1:4 and we ignore its top n largest loss regions during training, where n is set as 10% because the regions are usually considered as potentially noisy pseudo labels. To enable zero-shot metric depth estimation for arbitrary "in-the-wild" images, we further introduce a specialized scale head following the design principle of MoGe (Wang et al., 2025). The Scale Head is a lightweight two-layer MLP that takes the CLS token feature from the DINOv2 encoder as input, leveraging its global contextualized representation to predict a scalar metric scale factor s. Since scale prediction is a scalar regression task, we employ the standard L_1 loss:

$$L_{\text{scale}} = \frac{1}{N} \sum_{i=1}^{N} |s_i^{\text{pred}} - s_i^{\text{gt}}|, \tag{5}$$

where s_i^{pred} and s_i^{gt} represent the predicted and ground truth scale factors, respectively. We first fine-tune the metric depth model, then decouple and train the Scale Head separately. This improves training efficiency and provides high-quality, artifact-free, precise metric depth for downstream tasks like novel view synthesis, further enhancing the model's utility.

3 Experiments

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 best result for each metric is highlighted in **bold**, and the second best is underlined.

	Method	NYUv2		ScanNet		ETH3D		Sintel		DA2K
		$\delta_1 \uparrow$	AbsRel↓	$\delta_1 \uparrow$	AbsRel↓	$\delta_1 \uparrow$	AbsRel↓	$\delta_1 \uparrow$	AbsRel↓	Acc(%)
Generative	DepthFM (Gui et al., 2025)	0.956	0.065	0.947	0.068	0.958	0.069	0.700	0.577	85.8
	Marigold (Ke et al., 2024)	0.964	0.055	0.951	0.064	0.960	0.065	0.703	0.576	86.8
	GeoWizard (Fu et al., 2024)	0.966	0.052	0.953	0.061	0.961	0.064	0.705	0.574	88.1
Discriminative	LeReS (Yin et al., 2021)	0.916	0.090	0.917	0.091	0.777	0.171	0.590	0.785	–
	DPT (Ranftl et al., 2021)	0.919	0.091	0.932	0.084	0.929	0.115	0.605	0.700	–
	MiDaSv3.1 (Birkl et al., 2023)	0.980	0.048	0.957	0.069	0.947	0.078	0.670	0.611	–
	Metric3Dv2 (Hu et al., 2024)	0.963	0.058	0.941	0.074	0.960	0.066	0.662	0.619	–
	Depth Anything V2 (Yang et al., 2024c)	0.979	0.045	0.978	0.043	0.988	0.038	0.672	0.598	97.1
	Depth Pro (Bochkovskii et al., 2024)	0.580	0.245	0.667	0.207	0.687	0.201	0.476	0.883	–
	Ours	**0.982**	**0.041**	**0.981**	**0.033**	**0.991**	**0.029**	**0.719**	**0.513**	**97.3**

3.1 Implementation Details

Our reward-optimized D2I generator is based on the Flux.1 Depth[dev] (Labs, 2024) architecture and fine-tuned on depth data from synthetic datasets, using the training setup detailed in the Method section.

The generator synthesizes approximately 20 million visually realistic and geometrically accurate RGB images from source depth maps by optimizing a loss function that includes depth loss (weight 0.9) and aesthetic reward (weight 0.1). To achieve high geometric accuracy and depth similarity, we

ensure that the generated images meet stringent quality criteria: high-precision regions are identified by combining areas where: 1) ORB feature-based registered SSIM between the generated and original RGB images exceeds 0.85 (requiring at least 10 ORB matches), and 2) direct SSIM between them also exceeds 0.85. We only select mask samples where the valid region constitutes over 50% of the pixels, and a 3×3 erosion operation is performed to filter out excessively small regions. Subsequently, a 518×518 crop is randomly extracted, centered on the largest valid region's bounding box. For each original depth map, we generate four RGB images simultaneously, and use different random seeds. The prompt is set to None during the generation process.

We validate the effectiveness and advantages of the proposed method for monocular depth estimation through a series of experiments. We first introduce the datasets, evaluation metrics, and experimental settings. Then, we present the performance of our method on multiple benchmark datasets and conduct ablation experiments to analyze the contribution of each module to the overall performance of our model.

3.2 Evaluation Datasets

For monocular depth estimation, we conduct a series of experiments to evaluate the performance of our model on five widely used benchmarks. NYUv2 (Silberman et al., 2012) and ScanNet (Dai et al., 2017) provide indoor RGB-D data captured by depth sensors. ETH3D (Schops et al., 2017) includes both indoor and outdoor scenes, with depth data collected via laser scanners. For ScanNet, 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.

3.3 Evaluation Metrics

All evaluations are carried out under the zero-shot configuration. Following previous works (Yang et al., 2024c; 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 δ_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. δ_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.

3.4 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 preservation, and consistent depth estimation for diverse objects and scenes.

Quantitative Comparisons: As shown in Table 1, our method achieves the strongest results among compared methods across all evaluated benchmarks. On the indoor datasets NYUv2, ScanNet, and ETH3D it leads on both δ_1 and AbsRel, it is best on the synthetic Sintel, and it edges out Depth Anything V2 on DA2K. Depth Pro's lower relative-depth scores here stem from its optimization for metric rather than relative depth.

Qualitative Comparisons: Figure 4 highlights 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, it accurately estimates the depth of reflective surfaces such as mirrors on NYUv2, preserves complex details like distant table legs in ScanNet, and captures fine textures on Sintel. Crucially, it generalizes well to "in-the-wild" data, as exemplified by its accurate depth for transparent umbrellas.

3.5 Ablation Studies

We evaluate the contribution of each component in BRIDGE under consistent training configurations to ensure fair comparisons.

Impact of BRIDGE Generated Data: We first investigate our reward-optimized generation engine, the core mechanism for alleviating data scarcity and enhancing training diversity. We compare a

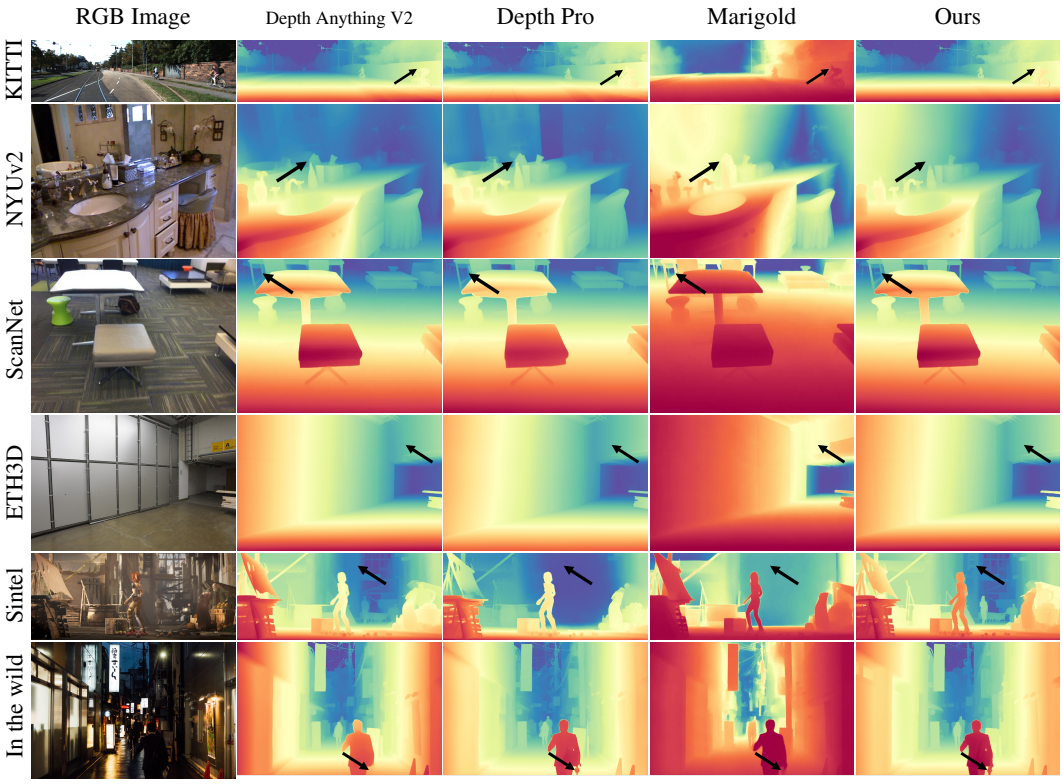


Figure 4: **Qualitative comparison of relative depth estimation across different datasets.** Our model captures fine-grained details and generalizes robustly on in-the-wild data.

Table 2: **Ablation on generated data and hybrid supervision.** Top: a ViT-L model trained on standard synthetic data vs. 1M BRIDGE-generated data. Bottom: training solely on pseudo-labels vs. our full hybrid supervision integrating similarity-guided high-precision ground truth depth. Both components improve performance across all benchmarks.

Method	NYUv2		ScanNet		ETH3D		Sintel	
	$\delta_1 \uparrow$	AbsRel$\downarrow$	$\delta_1 \uparrow$	AbsRel$\downarrow$	$\delta_1 \uparrow$	AbsRel$\downarrow$	$\delta_1 \uparrow$	AbsRel$\downarrow$
Impact of generated data								
Baseline	0.960	0.060	0.964	0.057	0.969	0.060	0.599	0.709
w/ Generated Data	**0.964**	**0.056**	**0.968**	**0.048**	**0.970**	**0.059**	**0.612**	**0.692**
Efficiency of hybrid supervision								
Pseudo-labels Only	0.958	0.069	0.957	0.069	0.949	0.078	0.595	0.751
Full Hybrid Strategy	**0.960**	**0.067**	**0.960**	**0.061**	**0.953**	**0.073**	**0.600**	**0.725**

baseline DINOv2-Base model trained on 1M synthetic data from various large-scale depth datasets against the same model trained on 1M BRIDGE-generated data (0.95M pseudo-labels and 0.05M masked source ground truth depth). As shown in Table 2, the generated data significantly boosts performance across all metrics, confirming that our engine produces realistic, geometrically accurate images that expand training scale and diversity.

Efficiency of Hybrid Training Strategy: As shown in Table 2, our hybrid depth supervision combines the broad coverage of teacher pseudo-labels with the precision of similarity-guided ground truth depth, providing more reliable supervision for learning geometric knowledge. Tested on the same 1M data, the BRIDGE model with our full hybrid strategy consistently outperforms a variant relying solely on teacher-generated pseudo-labels.

Reward Optimization vs. Standard SFT: Table 3 summarizes the comparison. We test whether reward optimization buys anything over plainly fine-tuning the same backbone, generating 1M images with each variant and training an identical DINOv2-Large student. Standard SFT, trained only on the noise-prediction objective, cannot enforce pixel-level agreement between image and conditioning depth, and the student reaches only 0.952 δ_1 / 0.063 AbsRel on ScanNet. The depth reward alone lifts this to 0.960 / 0.062, and the full reward reaches 0.960 / 0.061 on ScanNet and 0.952 / 0.074 on ETH3D. Since both variants see 1M images, the gain reflects reward-aligned geometry rather than data volume.

Table 3: **Reward optimization compared with standard SFT.** Downstream MDE results from 1M images generated by each D2I variant.

D2I Method	ScanNet		ETH3D	
	$\delta_1 \uparrow$	AbsRel$\downarrow$	$\delta_1 \uparrow$	AbsRel$\downarrow$
Standard SFT	0.952	0.063	0.946	0.079
RL (Depth Reward Only)	0.960	0.062	0.950	0.077
RL (Full: Depth + Aesthetic)	**0.960**	**0.061**	**0.952**	**0.074**

SSIM Threshold Sensitivity Analysis: Across thresholds from 0.80 to 0.90, δ_1 varies by at most 0.002 and AbsRel by at most 0.003. The 0.85 setting yields the lowest AbsRel on all three benchmarks and the best or joint-best δ_1 (Table 4); using 0.70 can yield worse performance than pseudo-label-only training.

Table 4: **SSIM Threshold Sensitivity Analysis.** Performance on key benchmarks under different SSIM thresholds for mask generation.

Method	NYUv2		ScanNet		ETH3D	
	$\delta_1 \uparrow$	AbsRel$\downarrow$	$\delta_1 \uparrow$	AbsRel$\downarrow$	$\delta_1 \uparrow$	AbsRel$\downarrow$
Pseudo-labels Only	0.958	0.069	0.957	0.069	0.949	0.078
SSIM = 0.70	0.958	0.070	0.955	0.072	0.948	0.079
SSIM = 0.80	0.959	0.068	0.958	0.062	0.951	0.075
SSIM = 0.85 (Ours)	**0.960**	**0.067**	**0.960**	**0.061**	**0.953**	**0.073**
SSIM = 0.90	0.960	0.069	0.960	0.064	0.952	0.074

Scaling Appearance per Geometry: We also vary K, the number of RGB images generated per source depth map, on a DINOv2-Large student. Starting from the original synthetic images with no generation gives 0.955 δ_1 on ScanNet and 0.590 on Sintel; a single generation per map barely moves this, but $K{=}4$ (2M images) reaches 0.969 / 0.615 and the full 20M set reaches 0.972 / 0.688. Performance improves monotonically with K and shows no sign of overfitting, confirming that varying appearance over fixed geometry drives the student to learn texture-invariant structure.

Fine-tuned to Metric Depth Estimation: We further transfer our pre-trained encoder with a lightweight scale head (a DINOv2 encoder plus an MLP predicting the metric scale). Our features remain competitive with dedicated metric models such as Metric3Dv2 (0.987 vs. 0.986 δ_1 on NYUv2), confirming that features learned from generated data transfer beyond relative depth. Full metric-depth results, model-size selection, and additional qualitative comparisons are provided in the supplementary material.

4 Conclusion

We introduce BRIDGE, a simple yet effective framework that mitigates the scarcity of depth annotations via a reward-optimized Depth-to-Image (D2I) engine that generates over 20M diverse RGB-D pairs. Combining it with a hybrid strategy that pairs noisy teacher pseudo-labels and similarity-guided source depth consistently yields the strongest zero-shot results on NYUv2, ScanNet, ETH3D, and Sintel, using only 20M images versus Depth Anything V2's 62M.

AI USE STATEMENT

Generative AI was used only for language polishing and grammar checking.

REPRODUCIBILITY STATEMENT

The Method section describes reward optimization and similarity-guided supervision. The appendix provides experimental configurations, generator and student training settings, the two-stage supervision algorithm, data-source composition, and additional results.

REFERENCES

Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2800–2810, 2018.

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*, pp. 4009–4018, 2021.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. URL https://arxiv.org/abs/2302.12288.

Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.

Aleksei Bochkovskii, AmaÃĢl Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024.

Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pp. 611–625. Springer, 2012.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

Alex Costanzino, Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Learning depth estimation for transparent and mirror surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9244–9255, 2023.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

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.

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.

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*, pp. 2002–2011, 2018.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pp. 241–258. Springer, 2024.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

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*, pp. 270–279, 2017.

Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast generative monocular depth estimation with flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3203–3211, 2025.

Xiankang He, Dongyan Guo, Hongji Li, Ruibo Li, Ying Cui, and Chi Zhang. Distill any depth: Distillation creates a stronger monocular depth estimator. *arXiv preprint arXiv:2502.19204*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. Ddp: Diffusion model for dense visual prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21741–21752, 2023.

Ebrahim Karami, Siva Prasad, and Mohamed Shehata. Image matching using sift, surf, brief and orb: performance comparison for distorted images. *arXiv preprint arXiv:1710.02726*, 2017.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9492–9502, 2024.

Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

LAION-AI. LAION Aesthetics Predictor V1. `https://github.com/LAION-AI/aesthetic-predictor`, 2022. GitHub repository.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896. Atlanta, 2013.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12009–12019, 2022.

Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4981–4991, 2023.

Jim Nilsson and Tomas Akenine-Möller. Understanding ssim. *arXiv preprint arXiv:2006.13846*, 2020.

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.

Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3depth: Monocular depth estimation with a piecewise planarity prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1610–1621, 2022.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.

Aakash Rajpal, Noshaba Cheema, Klaus Illgner-Fehns, Philipp Slusallek, and Sunil Jaiswal. High-resolution synthetic rgb-d datasets for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1188–1198, 2023.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

Haoyu Ren, Aman Raj, Mostafa El-Khamy, and Jungwon Lee. Suw-learn: Joint supervised, unsupervised, weakly supervised deep learning for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 750–751, 2020.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3260–3269, 2017.

Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, and Zhengguo Li. Nddepth: Normal-distance assisted monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7931–7940, 2023a.

Shuwei Shao, Zhongcai Pei, Xingming Wu, Zhong Liu, Weihai Chen, and Zhengguo Li. Iebins: Iterative elastic bins for monocular depth estimation. *Advances in Neural Information Processing Systems*, 36:53025–53037, 2023b.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. *arXiv preprint arXiv:1912.09678*, 2019a.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5261–5271, 2025.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916. IEEE, 2020.

Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8445–8453, 2019b.

Diana Wofk, Fangchang Ma, Tien-Ju Yang, Sertac Karaman, and Vivienne Sze. Fastdepth: Fast monocular depth estimation on embedded systems. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 6101–6108. IEEE, 2019.

Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. *arXiv preprint arXiv:2410.10815*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10371–10381, 2024b.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024c.

Xiaodong Yang, Zhuang Ma, Zhiyu Ji, and Zhe Ren. Gedepth: Ground embedding for monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12719–12727, 2023.

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1790–1799, 2020.

Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 204–213, 2021.

Fan Zhang, Shaodi You, Yu Li, and Ying Fu. Atlantis: Enabling underwater depth estimation with stable diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11852–11861, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

A Related Work

A.1 Monocular Depth Estimation

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

A.2 Self-supervised and Hybrid Supervision Learning

To train a robust MDE model despite the scarcity of precisely labeled data, researchers have explored various weakly supervised and self-supervised approaches, such as utilizing unlabeled monocular or stereo video data through photometric consistency loss for training (Godard et al., 2017; Ren et al., 2020). Other self-supervised approaches also include consistency training methods like FixMatch (Sohn et al., 2020) and unsupervised data augmentation to enhance certainty. Building on this, Pseudo-labeling supervision strategies (Lee et al., 2013), exemplified by Depth Anything V2 (Yang et al., 2024c), significantly improve model generalization by employing a powerful teacher model to generate pseudo-labels for massive amounts of unlabeled images. However, despite the broad coverage provided by pseudo-labels, their inherent noise and inaccuracy—especially around boundaries and fine details—remain a bottleneck for further enhancing depth estimation performance. To leverage both the scale of pseudo-labels and the precision of ground truth annotations, we propose a hybrid supervision strategy. Unlike methods that solely rely on pseudo-labels, our strategy integrates high-precision ground truth depth from the original data via a similarity detection mechanism. By utilizing the teacher model's pseudo-labels for broad training and the similarity-guided ground truth for precision fine-tuning, our hybrid approach provides a more reliable and comprehensive supervision signal.

A.3 Depth-to-Image Generation and Synthetic Data

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

B Additional Experiments

B.1 Experimental Configurations

Table 5 summarizes the student model, training-set size, and comparison performed in each experimental setting.

Table 5: Configurations used across the experiments and ablations.

Study	MDE student	Training data	Within-study comparison
Generated-data ablation	DINOv2-Base	1M	Original synthetic vs. BRIDGE-generated images
Hybrid-supervision ablation	DINOv2-Large	1M	Pseudo-label pretraining vs. two-stage hybrid training
SFT vs. reward optimization	DINOv2-Large	1M per generator	D2I optimization objective
Generation ratio	DINOv2-Large	0.5M–20M	Number of appearances per source depth map
Model selection	DINOv2-Small/Base/Large/Giant	20M	Encoder capacity
Final relative-depth model	DINOv2-Giant	20M	Final configuration
Metric-depth transfer	DINOv2-Giant	NYUv2 or KITTI train split	Dataset-specific scale head

B.2 Reward and MDE Training Details

The RL-D2I model is initialized using the Flux.1 Depth [dev] checkpoint (Labs, 2024). Given a generated image I_{gen} and the corresponding source depth map D_{source}, a frozen depth estimator produces D_{gen} from I_{gen}. We define the depth loss as

$$L_{\text{D}} = 1 - \frac{\langle D_{\text{gen}}, D_{\text{source}} \rangle}{\|D_{\text{gen}}\|_2 \, \|D_{\text{source}}\|_2}. \tag{6}$$

We minimize this cosine loss directly, without scale-and-shift fitting. The aesthetic term is computed with a CLIP ViT-L/14 image encoder and the publicly released aesthetic scoring head (LAION-AI, 2022). The total loss is

$$L_{\text{total}} = 0.9 L_{\text{D}} - 0.1 R_{\text{aesthetic}}(I_{\text{gen}}). \tag{7}$$

Following the reward-gradient formulation of VADER (Prabhudesai et al., 2024), we backpropagate through the 20-step sampling trajectory and use gradient checkpointing to reduce activation memory. We optimize rank-16 LoRA adapters on the Flux attention projections with AdamW at a learning rate of 1×10^{-5}. Training uses 64 A800 GPUs, a total batch size of 128, and 1,500 optimization steps.

All relative-depth models in these experiments use a DPT head. Following Depth Anything V2, we optimize the encoder and decoder with Adam at learning rates of 5×10^{-6} and 5×10^{-5}, respectively, using a total batch size of 64 and 518×518 image crops. For zero-shot relative-depth evaluation, we align predictions with ground truth using scale and shift before computing δ_1 and AbsRel.

B.3 Two-Stage Hybrid Supervision

A DINOv2-Giant encoder with a DPT head serves as the pseudo-label teacher. We train this teacher on 1M synthetic images with dense ground-truth depth drawn from the sources in Table 7. The student is pretrained on dense pseudo-labels for the generated images and then fine-tuned with source-depth supervision in regions where the generated image retains the source geometry; pixels outside the fusion mask are ignored. We compute one SSIM mask after ORB registration and a second mask directly, without registration. In the 1M hybrid-supervision ablation, the two stages use 0.95M pseudo-labeled samples and 0.05M masked source-depth samples, respectively.

Algorithm 1 Two-stage hybrid supervision

Require: Generated RGB images I_{gen}, original synthetic RGB images I_{orig}, source depth maps D_{source}, frozen teacher T, threshold $\tau = 0.85$

1: Predict $\widetilde{D} = T(I_{\text{gen}})$ for each generated image.
2: Pretrain the student on $(I_{\text{gen}}, \widetilde{D})$.
3: Register I_{gen} to I_{orig} using ORB feature matches; threshold the registered SSIM map at τ to obtain M_{reg}.
4: Threshold the direct SSIM map at τ to obtain M_{dir} without registration.
5: Set $M = \text{Erode}_{3\times3}(M_{\text{reg}} \vee M_{\text{dir}})$ and retain samples for which the valid mask covers more than 50% of the image.
6: Fine-tune on I_{gen} using D_{source} within the valid region defined by M; pixels outside M are ignored.

B.4 Model Selection

We compare DINOv2 (Oquab et al., 2023) encoders using the same 20M-image training set and optimization protocol while varying only encoder capacity. For DINOv2-Large and DINOv2-Giant, we use checkpoints without register tokens. Replacing DINOv2-Large with DINOv2-Giant improves the ScanNet δ_1/AbsRel scores from 0.972/0.049 to 0.981/0.033 and the corresponding Sintel scores from 0.688/0.588 to 0.719/0.513 (Table 6), so the final model uses DINOv2-Giant.

Table 6: Comparison of pretrained DINOv2 encoders of different sizes after training on the same 20M images.

Method	NYUv2		ScanNet		ETH3D		Sintel	
	$\delta_1 \uparrow$	AbsRel$\downarrow$	$\delta_1 \uparrow$	AbsRel$\downarrow$	$\delta_1 \uparrow$	AbsRel$\downarrow$	$\delta_1 \uparrow$	AbsRel$\downarrow$
DINOv2-Small	0.964	0.062	0.921	0.085	0.978	0.045	0.637	0.684
DINOv2-Base	0.968	0.054	0.941	0.065	0.980	0.039	0.647	0.644
DINOv2-Large	0.978	0.045	0.972	0.049	0.989	0.032	0.688	0.588
DINOv2-Giant	**0.982**	**0.041**	**0.981**	**0.033**	**0.991**	**0.029**	**0.719**	**0.513**

B.5 Training Data Composition

The pseudo-label teacher is trained using ground-truth depth from 1M synthetic images drawn from the seven sources in Table 7. Depth maps from the same collections condition the D2I generator, yielding 20M RGB images. For example, the generator uses indoor source geometry from Hypersim to produce images with both indoor and outdoor appearances.

B.6 Metric-Depth Transfer

We adapt the final DINOv2-Giant encoder and DPT head to metric-depth prediction by adding a dataset-specific scale head implemented as a two-layer MLP. For NYUv2 and KITTI, we freeze both the encoder and DPT head and train only the scale head on the corresponding official training split. Let $\hat{d}_i(p)$ denote the frozen model's relative-depth prediction at a valid pixel $p \in \Omega_i$, and let $d_i^{\text{gt}}(p)$ denote the corresponding metric ground truth. The image-level target scale is estimated without a shift,

$$s_i^{\text{gt}} = \arg\min_s \sum_{p\in\Omega_i} \left(s\hat{d}_i(p) - d_i^{\text{gt}}(p)\right)^2, \tag{8}$$

and the scale head takes the encoder's CLS token as input and is trained to predict this scalar with an L_1 loss. At inference time, multiplying $\hat{d}_i$ by s_i^{pred} yields the metric-depth estimate. On NYUv2, our method obtains δ_1/AbsRel/RMSE scores of 0.986/0.052/0.197, compared with 0.987/0.046/0.181 for Metric3Dv2 (Table 8a). On KITTI, our method obtains δ_1/AbsRel/RMSE scores of 0.983/0.045/1.862 (Table 8b). For each benchmark, the scale head is trained on the corresponding dataset's training split; the metric-transfer results are therefore supervised.

Table 7: Training sources for the 1M dense-depth synthetic set and the 20M generated set. Indoor and Outdoor denote source domains in the upper block and rendered appearances in the lower block. Counts are rounded; * denotes the number of images sampled from the corresponding dataset.

Dataset	Indoor	Outdoor	# Images
Synthetic Images with Dense Ground-Truth Depth (1M; Source Domain)			
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*
Generated Images (20M; Rendered Appearance)			
BlendedMVS	✓	✓	2.2M
TartanAir		✓	2M
Hypersim	✓	✓	2.4M
IRS	✓	✓	4.2M
VKITTI 2		✓	1M
Spring		✓	20K
DA-V	✓	✓	8.2M

B.7 Scaling Generated Appearances and Data Volume

Because total image count grows for $K>1$, Table 9 evaluates appearance multiplicity together with data volume. The baseline and $K=1$ settings contain 0.5M images each, $K=4$ contains 2M, and variable $K>8$ contains 20M. The $K=1$ setting performs slightly worse than the original-image baseline on every metric, whereas $K=4$ improves all four benchmarks. Among the settings in Table 9, the 20M-image setting achieves the highest δ_1 on all four benchmarks and the lowest AbsRel on NYUv2, ETH3D, and Sintel; its ScanNet AbsRel is 0.049, compared with 0.046 for $K=4$.

B.8 Reward Optimization versus Standard SFT

The reward-gradient variant and standard SFT use the same Flux.1 Depth [dev] initialization, source depth maps, LoRA rank, and number of training steps; standard SFT uses the Flux flow-matching objective. Each generator produces 1M images, and the same DINOv2-Large student is trained on each set. Relative to standard SFT, training with the depth reward improves δ_1/AbsRel from 0.952/0.063 to 0.960/0.062 on ScanNet and from 0.946/0.079 to 0.950/0.077 on ETH3D (Table 3). Adding the aesthetic reward leaves δ_1 on ScanNet unchanged while reducing AbsRel to 0.061. The aesthetic reward produces its largest additional improvement on ETH3D, where the scores reach 0.952/0.074. Data volume and downstream student training remain fixed.

B.9 RL-D2I Training Reward Progression

The mean total reward rises from 5.0683 before fine-tuning to 6.5884 at step 1,500, with an increase at every recorded checkpoint (Table 10).

Table 8: Metric depth estimation after training a dataset-specific scale head. Published comparison results use each method's reported protocol; the table provides context rather than matched retraining. The best result in each column is highlighted in **bold**.

Method	Higher ↑	Lower ↓	
	δ_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**
Ours	0.986	0.052	0.197

(a) Metric-depth results on NYUv2.

Method	Higher ↑	Lower ↓	
	δ_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., 2023)	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
Ours	0.983	0.045	**1.862**

(b) Metric-depth results on KITTI.

Table 9: **Scaling generated appearances and data volume.** DINOv2-Large results for settings that increase K and, beyond $K{=}1$, the total number of training images. The 20M setting uses a variable $K{>}8$ with no fixed upper bound.

Setting	NYUv2		ScanNet		ETH3D		Sintel	
	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓
Baseline (K=0, 500K)	0.957	0.063	0.955	0.060	0.962	0.063	0.590	0.785
K=1, 500K	0.955	0.064	0.950	0.064	0.960	0.066	0.584	0.789
K=4, 2M	0.964	0.054	0.969	**0.046**	0.971	0.058	0.615	0.687
Variable $K{>}8$, 20M	**0.978**	**0.045**	**0.972**	0.049	**0.989**	**0.032**	**0.688**	**0.588**

Table 10: **RL-D2I training reward progression.** Mean total reward at recorded training steps on 64×A800 GPUs with batch size 128.

Training Step	Mean Total Reward
0	5.0683
200	5.2497
700	5.9673
1200	6.2833
1500	6.5884

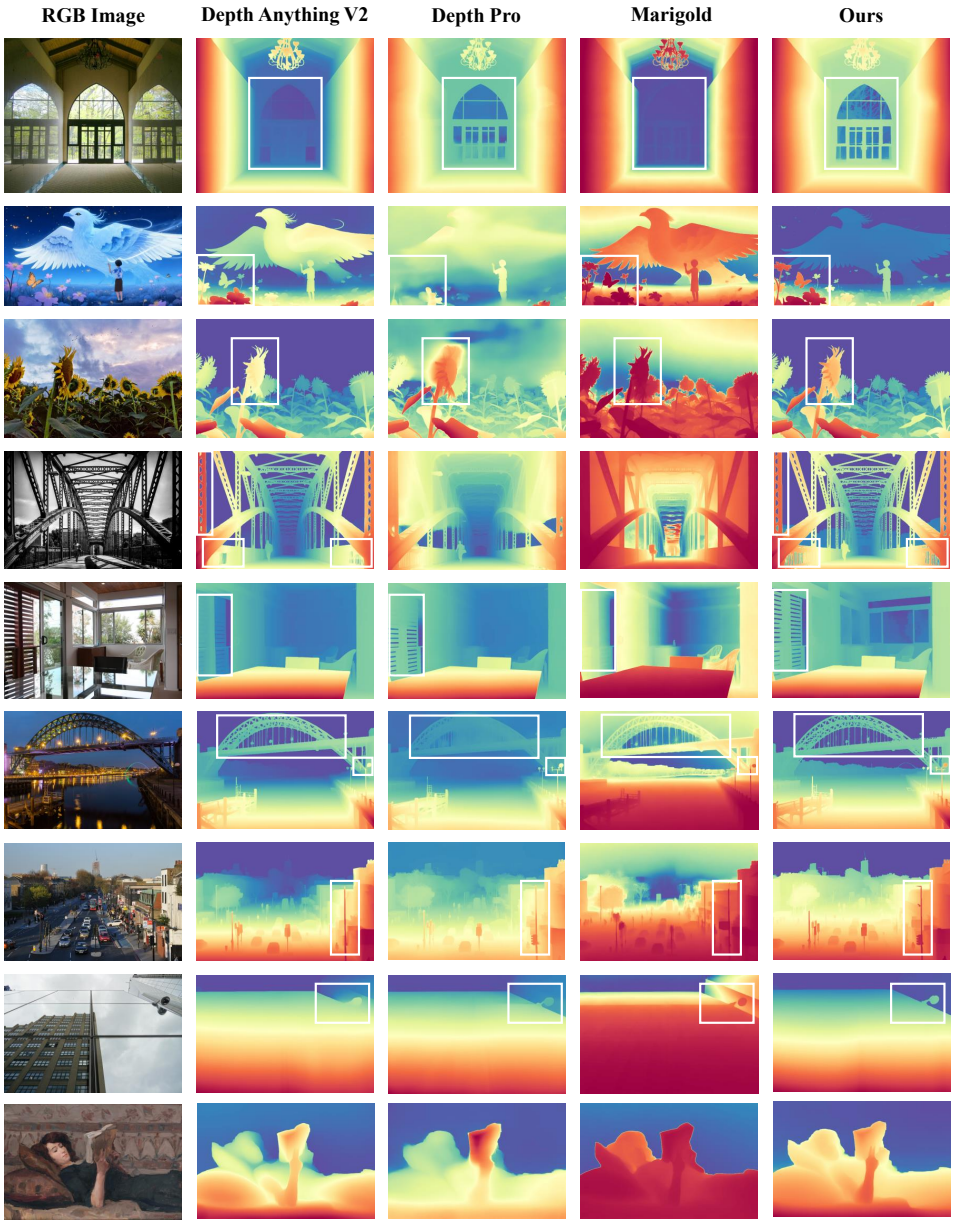


Figure 5: In-the-wild relative-depth comparison with Depth Anything V2 (Yang et al., 2024c), Depth Pro (Bochkovskii et al., 2024), and Marigold (Ke et al., 2025).

B.10 Additional Qualitative Comparisons

Relative depth: Figures 5 and 6 compare in-the-wild predictions from BRIDGE, Depth Anything V2 (Yang et al., 2024c), Depth Pro (Bochkovskii et al., 2024), and Marigold (Ke et al., 2025). The boxes highlight thin structures, object boundaries, and distant objects. In these examples, BRIDGE delineates these regions with sharper transitions and clearer separation from their surroundings.

Conditional synthesis: Figure 7 visualizes how each predicted depth map conditions the same Stable Diffusion 1.5 ControlNet (Zhang et al., 2023). Both conditioning runs use the same prompt and random seed. The BRIDGE-conditioned samples better preserve the source layout and object extents in these examples.



Figure 6: Additional in-the-wild relative-depth comparisons across indoor, outdoor, and stylized scenes.

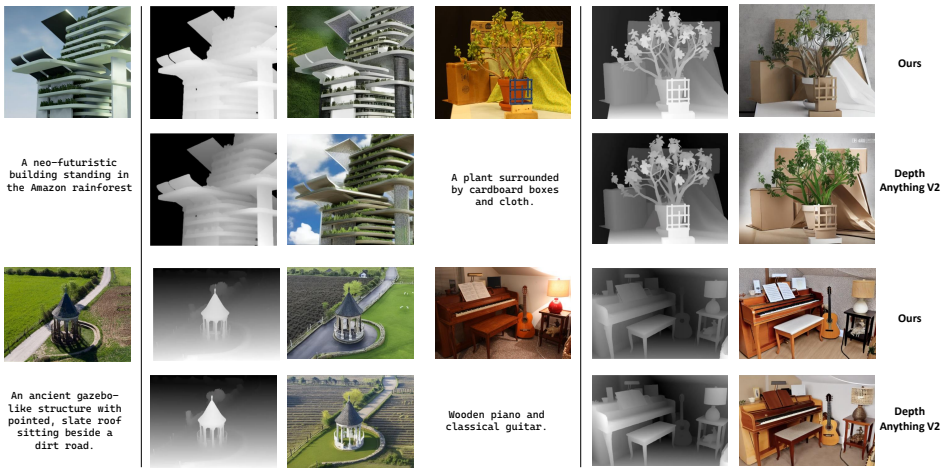


Figure 7: Depth-conditioned synthesis using predictions from BRIDGE and Depth Anything V2 (Yang et al., 2024c). Both branches use the same ControlNet (Zhang et al., 2023), prompt, and random seed; the prompts appear in the lower-left corners.

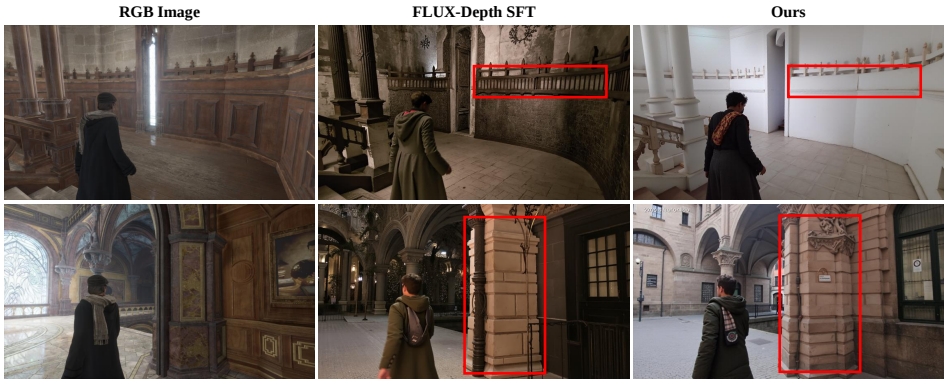


Figure 8: Each row shows the reference RGB image, the standard SFT baseline initialized from Flux.1 Depth [dev], and the output of our reward-optimized model.