

Distill Any Depth: Distillation Creates a Stronger Monocular Depth Estimator

Xiankang He^{*1,2} Dongyan Guo^{*1} Hongji Li^{2,3} Ruibo Li⁴ Ying Cui¹ Chi Zhang^{†2}

¹Zhejiang University of Technology ²AGI Lab, Westlake University

³Lanzhou University ⁴Nanyang Technological University

{hexiankang577, 3420670269neon}@gmail.com {guodongyan, cuiying}@zjut.edu.cn

ruibo.li@ntu.edu.cn chizhang@westlake.edu.cn

<https://distill-any-depth-official.github.io/>

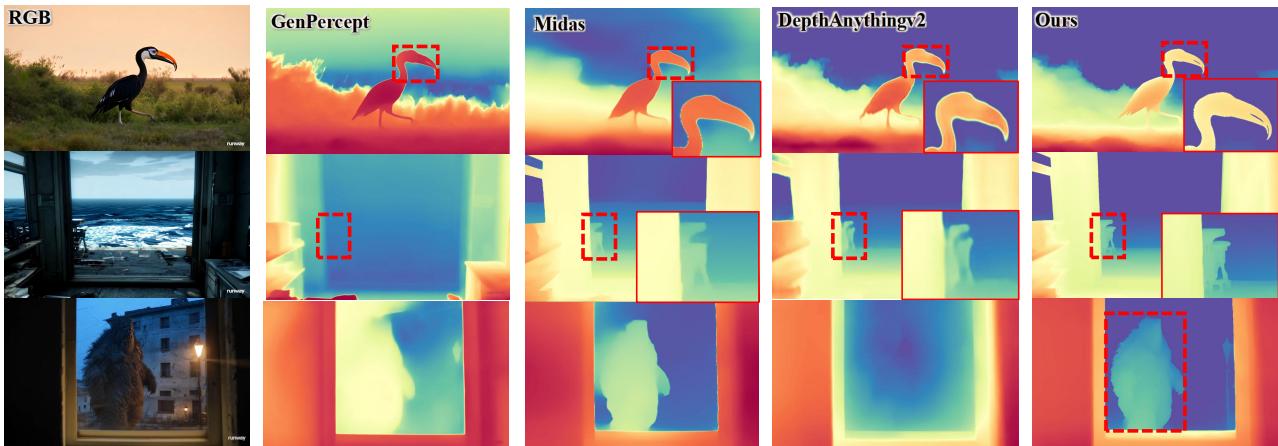


Figure 1: **Zero-shot prediction on in-the-wild images.** Our model, distilled from Genpercept [45] and DepthAnythingv2 [47], outperforms other methods by delivering more accurate depth details and exhibiting superior generalization for monocular depth estimation on in-the-wild images.

Abstract

Monocular depth estimation (MDE) aims to predict scene depth from a single RGB image and plays a crucial role in 3D scene understanding. Recent advances in zero-shot MDE leverage normalized depth representations and distillation-based learning to improve generalization across diverse scenes. However, current depth normalization methods for distillation, relying on global normalization, can amplify noisy pseudo-labels, reducing distillation effectiveness. In this paper, we systematically analyze the impact of different depth normalization strategies on pseudo-label distillation. Based on our findings, we propose Cross-Context Distillation, which integrates global and local depth cues to enhance pseudo-label quality. Additionally, we introduce a multi-teacher distillation framework that leverages complementary strengths of different depth estimation models,

leading to more robust and accurate depth predictions. Extensive experiments on benchmark datasets demonstrate that our approach significantly outperforms state-of-the-art methods, both quantitatively and qualitatively.

1. Introduction

Monocular depth estimation (MDE) predicts scene depth from a single RGB image, offering flexibility compared to stereo or multi-view methods. This makes MDE ideal for applications like autonomous driving and robotic navigation [10, 12, 16, 48, 28]. Recent research on zero-shot MDE models [34, 51, 43, 22] aims to handle diverse scenarios, but training such models requires large-scale, diverse depth data, which is often limited by the need for specialized equipment [29, 50]. A promising solution is using large-scale unlabeled data, which has shown success in tasks like classification and segmentation [25, 58, 44]. Studies like DepthAnything [46] highlight the effectiveness of using pseudo labels from teacher models for training student

^{*}denotes co-first authorship. This work was done while Xiankang He was a visiting student at the AGI Lab, Westlake University.

[†] denotes corresponding author.

models.

To enable training on such a diverse, mixed dataset, most state-of-the-art methods [47, 37, 51] employ scale-and-shift invariant (SSI) depth representations for loss computation. This approach normalizes raw depth values within an image, making them invariant to scaling and shifting, and ensures that the model learns to focus on relative depth relationships rather than absolute values. The SSI representation facilitates the joint use of diverse depth data, thereby improving the model’s ability to generalize across different scenes [35, 4]. Similarly, during evaluation, the metric depth of the prediction is recovered by solving for the unknown scale and shift coefficients of the predicted depth using least squares, ensuring the application of standard evaluation metrics.

Despite its advantages, using SSI depth representation for pseudo-label distillation in MDE models presents several issues. Specifically, the inherent normalization process in SSI loss makes the depth prediction at a given pixel not only dependent on the teacher model’s raw prediction at that location but also influenced by the depth values in other regions of the image. This becomes problematic because pseudo-labels inherently introduce noise. Even if certain local regions are predicted accurately, inaccuracies in other regions can negatively affect depth estimates after global normalization, leading to suboptimal distillation results. As shown in Fig. 2, we empirically demonstrate that normalizing depth maps globally tends to degrade the accuracy of local regions, as compared to only applying normalization within localized regions during evaluation.

Building on this insight, in this paper, we first investigate the issue of depth normalization in pseudo-label distillation. We begin by analyzing various depth normalization strategies, including global normalization, local normalization, hybrid global-local approaches, and the absence of normalization. Through empirical experiments, we explore how each technique affects the performance of various distillation designs, especially when using pseudo-labels for training. Our analysis provides valuable insights into how different normalization methods influence the MDE loss function and distillation outcomes, offering a set of best practices for optimizing performance in diverse scenarios.

Building on this empirical foundation, we introduce a Cross-Context Distillation method, designed to distill knowledge from the teacher model more effectively. We are motivated by our finding that local regions, when used for distillation, produce pseudo-labels that capture higher-quality depth details, improving the student model’s depth estimation accuracy. However, focusing solely on local regions might overlook the broader contextual relationships in the image. To address the issue, we combine local and global inputs within a unified distillation framework. By combining the context-specific advantages of local distillation with the broader understanding provided by global methods, our

method achieves more detailed and reliable depth predictions.

Furthermore, we propose a multi-teacher distillation framework that leverages the complementary strengths of multiple depth estimation models. Our design is motivated by the observation of recent advancements that diffusion-based models, benefiting from large-scale image priors, excel at capturing fine-grained details but are computationally expensive, while encoder-decoder models provide higher accuracy and efficiency but relatively lack fine-detail reconstruction. To harness these strengths, we randomly select different models to generate pseudo-labels, and then supervise the student model based on these labels. This operation enables the student model to learn from the detailed depth information of diffusion-based models while benefiting from the precision of encoder-decoder models.

To validate the effectiveness of our design, we conduct extensive experiments on various benchmark datasets. The empirical results show that our method significantly outperforms existing baselines qualitatively and quantitatively. The contributions can be summarized below:

- We systematically analyze the role of different depth normalization strategies in pseudo-label distillation, providing insights into their effects on MDE performance.
- We propose Cross-Context Distillation, a hybrid local-global distillation framework that enhances distillation by leveraging both fine-grained details and global depth relationships.
- We develop a multi-teacher distillation framework that integrates pseudo labels from multiple depth estimation models, combining the strengths of various depth models.
- We conduct extensive experiments on benchmark datasets, demonstrating that our method outperforms state-of-the-art approaches both quantitatively and qualitatively. Code and models are made publicly available.

2. Related Work

2.1. Monocular Depth Estimation

Monocular depth estimation (MDE) has evolved from hand-crafted methods to deep learning, significantly improving accuracy [10, 27, 11, 15, 57, 35]. Architectural refinements, such as multi-scale designs and attention mechanisms, have further enhanced feature extraction [19, 5, 56]. However, most models remain reliant on labeled data and struggle to generalize across diverse environments. Zero-shot MDE improves generalization by leveraging large-scale datasets, geometric constraints, and multi-task learning [34, 51, 53, 55]. Metric depth estimation incorporates intrinsic data for absolute depth learning [2, 52, 20, 42], while

generative models such as Marigold refine depth details using diffusion priors [22, 45]. Despite these advances, effectively utilizing unlabeled data remains a challenge due to pseudo-label noise and inconsistencies across different contexts. DepthAnything [47] explores large-scale unlabeled data but struggles with pseudo-label reliability. PatchFusion [8, 30] improves depth estimation by refining high-resolution image representations but lacks adaptability in generative settings. To address these issues, we propose Cross-Context and Multi-Teacher Distillation, which enhances pseudo-label supervision by leveraging diverse contextual information and multiple expert models, improving both accuracy and generalization ability.

2.2. Semi-supervised Monocular Depth Estimation

Semi-supervised depth estimation has gained attention by utilizing temporal consistency to better use unlabeled data [26, 17]. Some methods [1, 40, 6, 49, 14] apply stereo geometric constraints, enforcing left-right consistency to enhance depth accuracy, while others use additional supervision like semantic priors [33, 18] or GANs, such as DepthGAN [21]. However, these approaches are limited by their reliance on temporal cues or stereo constraints, restricting their applicability. Recent work [32] explored pseudo-labeling for semi-supervised MDE but lacks generative modeling capabilities. DepthAnything [46] demonstrated the potential of large-scale unlabeled data, though pseudo-label reliability remains challenging. In contrast, our approach improves pseudo-label reliability and enhances MDE accuracy, relying solely on unlabeled data without additional constraints.

3. Method

In this section, we introduce a novel distillation framework designed to leverage unlabeled images for training zero-shot Monocular Depth Estimation (MDE) models. We begin by exploring various depth normalization techniques in Section 3.1, followed by detailing our proposed distillation method in Section 3.2, which combines predictions across multiple contexts. The overall framework is illustrated in Fig. 3. Finally, we describe a multi-teacher distillation mechanism in Section 3.2 that integrates diverse depth estimators as teacher models to train the student model.

3.1. Depth Normalization

Depth normalization is a crucial component of our framework as it adjusts the pseudo-depth labels \mathbf{d}^t from the teacher model and the depth predictions \mathbf{d}^s from the student model for effective loss computation. To understand the influence of normalization techniques on distillation performance, we systematically analyze several approaches commonly employed in prior works. These strategies are visually illustrated in Fig. 4.

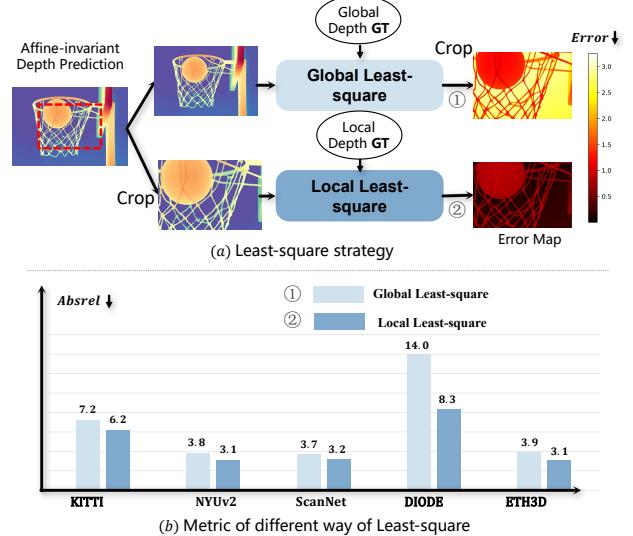


Figure 2: Issue with Global Normalization (SSI). In (a), we compare two alignment strategies for the central $w/2, h/2$ region: (1) *Global Least-Square*, where alignment is applied to the full image before cropping, and (2) *Local Least-Square*, where alignment is performed on the cropped region. Metrics are computed on the cropped region. As shown in (b), the outperformed local strategy demonstrates that **global normalization degrades local accuracy compared to local normalization**.

Global Normalization: The first strategy we examine is the global normalization [46, 47] used in recent distillation methods. Global normalization [34] adjusts depth predictions using global statistics of the entire depth map. This strategy aims to ensure scale-and-shift invariance by normalizing depth values based on the median and mean absolute deviation of the depth map. For each pixel i , the normalized depth for the student model and pseudo-labels are computed as:

$$\begin{aligned}\tilde{d}_i^s &= \mathcal{N}_{glo}(\mathbf{d}^s) = \frac{d_i^s - \text{med}(\mathbf{d}^s)}{\frac{1}{M} \sum_{j=1}^M |d_j^s - \text{med}(\mathbf{d}^s)|} \\ \tilde{d}_i^t &= \mathcal{N}_{glo}(\mathbf{d}^t) = \frac{d_i^t - \text{med}(\mathbf{d}^t)}{\frac{1}{M} \sum_{j=1}^M |d_j^t - \text{med}(\mathbf{d}^t)|},\end{aligned}\quad (1)$$

where $\text{med}(\mathbf{d}^s)$ and $\text{med}(\mathbf{d}^t)$ are the medians of the predicted depth and pseudo depth, respectively. The final regression loss for distillation is computed as the average absolute difference between the normalized predicted depth and the normalized pseudo depth across all valid pixels M :

$$\mathcal{L}_{\text{Dis}} = \frac{1}{M} \sum_{i=1}^M |\tilde{d}_i^s - \tilde{d}_i^t|. \quad (2)$$

Hybrid Normalization: In contrast to global normalization, Hierarchical Depth Normalization [54] employs a hybrid

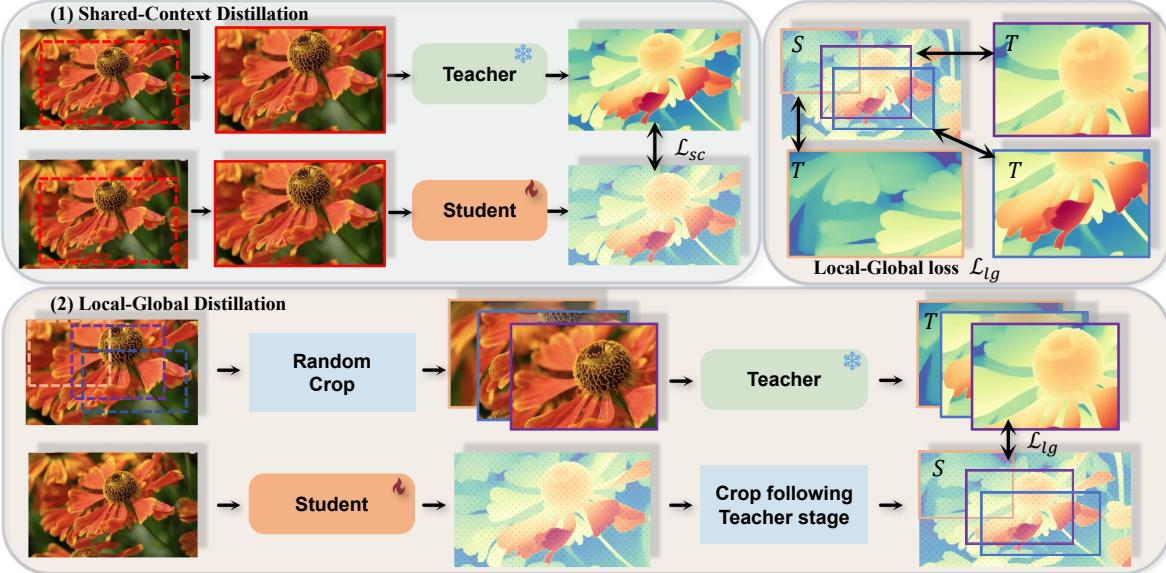


Figure 3: Overview of Cross-Context Distillation. Our method combines local and global depth information to enhance the student model’s predictions. It includes two scenarios: (1) *Shared-Context Distillation*, where both models use the same image for distillation; and (2) *Local-Global Distillation*, where the teacher predicts depth for overlapping patches while the student predicts the full image. The Local-Global loss \mathcal{L}_{lg} (Top Right) ensures consistency between local and global predictions, enabling the student to learn both fine details and broad structures, improving accuracy and robustness.

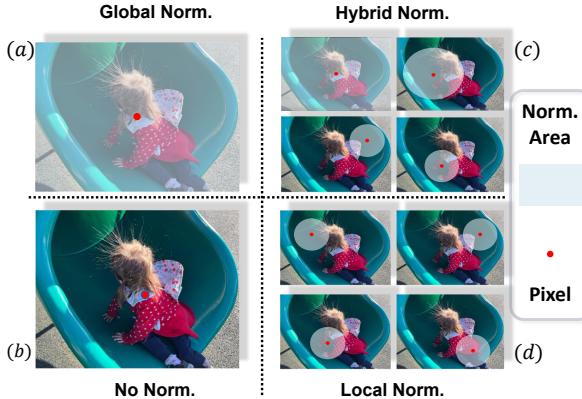


Figure 4: Normalization Strategies. We compare four normalization strategies: Global Norm [34], Hybrid Norm [54], Local Norm, and No Norm. The figure visualizes how each strategy processes pixels within the normalization region (Norm. Area). The red dot represents any pixel within the region.

normalization approach by integrating both global and local depth information. This strategy is designed to preserve both the global structure and local geometry in the depth map. The process begins by dividing the depth range into S segments, where S is selected from $\{1, 2, 4\}$. When $S = 1$, the entire depth range is normalized globally, treating all pixels as part of a single context, akin to global normalization. In the case of $S = 2$, the depth range is divided into two segments, with each pixel being normalized within

one of these two local contexts. Similarly, for $S = 4$, the depth range is split into four segments, allowing normalization to be performed within smaller, localized contexts. By adapting the normalization process to multiple levels of granularity, hybrid normalization achieves a balance between global coherence and local adaptability. For each context u , the normalized depth values for the student model $\mathcal{N}_u(d_i^s)$ and pseudo-labels $\mathcal{N}_u(d_i^t)$ are calculated within the corresponding depth range. The loss for each pixel i is then computed by averaging the losses across all contexts U_i to which the pixel belongs:

$$\mathcal{L}_{Dis}^i = \frac{1}{|U_i|} \sum_{u \in U_i} |\mathcal{N}_u(d_i^s) - \mathcal{N}_u(d_i^t)|, \quad (3)$$

where $|U_i|$ denotes the total number of groups (or contexts) that pixel i is associated with. To obtain the final loss \mathcal{L}_{Dis} , we average the pixel-wise losses across all valid pixels M :

$$\mathcal{L}_{Dis} = \frac{1}{M} \sum_{i=1}^M \mathcal{L}_{Dis}^i. \quad (4)$$

Local Normalization: In addition to global and hybrid normalization, we investigate Local Normalization, a strategy that focuses exclusively on the finest-scale groups used in hybrid normalization. This approach isolates the smallest local contexts for normalization, emphasizing the preservation of fine-grained depth details without considering hierarchical or global scales. Local normalization operates by dividing

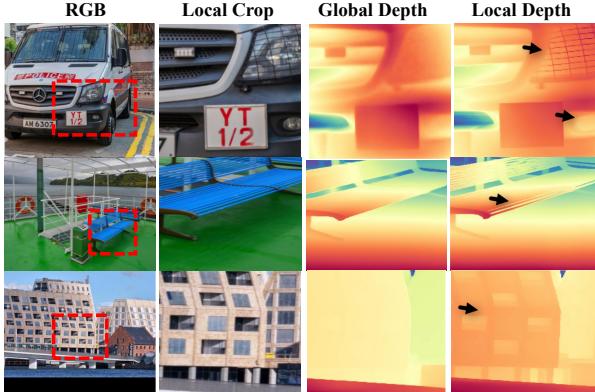


Figure 5: Different Inputs Lead to Different Pseudo Labels. Global Depth: The teacher model predicts depth using the entire image, and the local region’s prediction is cropped from the output. Local Depth: The teacher model directly takes the cropped local region as input, resulting in more refined and detailed depth estimates for that area, capturing finer details compared to using the entire image.

the depth range into the smallest groups, corresponding to $S = 4$ in the hybrid normalization framework, and each pixel is normalized within its local context. The loss for each pixel i is computed using a similar formulation as in hybrid normalization, but with u^i now representing the local context for pixel i , defined by the smallest four-part group:

$$\mathcal{L}_{\text{Dis}} = \frac{1}{M} \sum_{i=1}^M |\mathcal{N}_{u^i}(d_i^s) - \mathcal{N}_{u^i}(d_i^t)|. \quad (5)$$

No Normalization: As a baseline, we also consider a direct depth regression approach with no explicit normalization. The absolute difference between raw student predictions and teacher pseudo-labels is used for loss computation:

$$\mathcal{L}_{\text{Dis}} = \frac{1}{M} \sum_{i=1}^M |d_i^s - d_i^t|, \quad (6)$$

This approach eliminates the need for normalization, assuming pseudo-depth labels naturally reside in the same domain as predictions. It provides insight into whether normalization enhances distillation effectiveness or if raw depth supervision suffices.

3.2. Distillation Pipeline

In this section, we introduce an enhanced distillation pipeline that integrates two complementary strategies: Cross-Context Distillation and Multi-Teacher Distillation. Both strategies aim to improve the quality of pseudo-label distillation, enhance the model’s fine-grained perception, and boost generalization across diverse scenarios.

Cross-context Distillation. A key challenge in monocular depth distillation is the trade-off between local detail preservation and global depth consistency. As shown in Fig. 5, providing a local crop of an image as input to the teacher model enhances fine-grained details in the pseudo-depth labels, but it may fail to capture the overall scene structure. Conversely, using the entire image as input preserves the global depth structure but often lacks fine details. To address this limitation, we propose Cross-Context Distillation, a method that enables the student model to learn both local details and global structures simultaneously. Cross-context distillation consists of two key strategies:

1) Shared-Context Distillation: In this setup, both the teacher and student models receive the same cropped region of the image as input. Instead of using the full image, we randomly sample a local patch of varying sizes from the original image and provide it as input to both models. This encourages the student model to learn from the teacher model across different spatial contexts, improving its ability to generalize to varying scene structures. For the loss of shared-context distillation, the teacher and student models receive identical inputs and produce each depth prediction, denoted as $\mathbf{d}_{\text{local}}^t$ and $\mathbf{d}_{\text{local}}^s$:

$$\mathcal{L}_{\text{sc}} = \mathcal{L}_{\text{Dis}} (\mathbf{d}_{\text{local}}^s, \mathbf{d}_{\text{local}}^t), \quad (7)$$

This loss encourages the student model to refine its fine-grained predictions by directly aligning with the teacher’s outputs at local scales.

2) Local-Global Distillation: In this approach, the teacher and student models operate on different input contexts. The teacher model processes local cropped regions, generating fine-grained depth predictions, while the student model predicts a global depth map from the entire image. To ensure knowledge transfer, the teacher’s local depth predictions supervise the corresponding overlapping regions in the student’s global depth map. This strategy allows the student to integrate fine-grained local details into its holistic depth estimation. Formally, the teacher model produces multiple depth predictions for cropped regions, denoted as $\mathbf{d}_{\text{local}_n}^t$, while the student generates a global depth map, $\mathbf{d}_{\text{global}}^s$. The loss for Local-Global distillation is computed only over overlapping areas between the teacher’s local predictions and the corresponding regions in the student’s global depth map:

$$\mathcal{L}_{\text{lg}} = \frac{1}{N} \sum_{n=1}^N \mathcal{L}_{\text{Dis}} (\text{Crop}(\mathbf{d}_{\text{global}}^s), \mathbf{d}_{\text{local}_n}^t), \quad (8)$$

where $\text{Crop}(\cdot)$ extracts the overlapping region from the student’s depth prediction, and N is the total number of sampled patches. This loss ensures that the student benefits from the detailed local supervision of the teacher model while maintaining global depth consistency. The total loss function integrates both local and cross-context losses along with additional constraints, including feature alignment and gradient

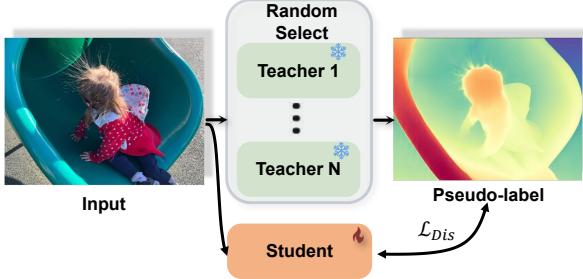


Figure 6: Multi-teacher Mechanism. We introduce a multi-teacher distillation approach, where pseudo-labels are generated from multiple teacher models. At each training iteration, one teacher is randomly selected to produce pseudo-labels for unlabeled images.

preservation, as proposed in prior works [47]:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sc}} + \lambda_1 \cdot \mathcal{L}_{\text{lg}} + \lambda_2 \cdot \mathcal{L}_{\text{feat}} + \lambda_3 \cdot \mathcal{L}_{\text{grad}}. \quad (9)$$

Here, λ_1 , λ_2 , and λ_3 are weighting factors that balance the different loss components. By incorporating cross-context supervision, this framework effectively allows the student model to integrate both fine-grained details from local crops and structural coherence from global depth maps.

Multi-teacher Distillation. In addition to cross-context distillation, we adopt a multi-teacher distillation strategy, illustrated in Fig. 6, to further enhance the quality and robustness of the distilled depth knowledge. This approach leverages multiple teacher models, each trained with distinct architectures, optimization strategies, or data distributions, to generate diverse pseudo-labels. By aggregating knowledge from multiple sources, the student model benefits from a richer and more generalized depth representation. Formally, given a set of pre-trained teacher models $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_N$, we employ a probabilistic teacher selection mechanism, where one teacher model is randomly selected at each training iteration to generate pseudo-labels for the input image. The inclusion of multiple teacher models allows the student to learn from a diverse set of predictions, effectively mitigating biases and limitations inherent to any single model.

4. Experiment

4.1. Experimental Settings

Datasets. To evaluate the effectiveness of our proposed distillation framework, we follow the methodology outlined in DepthAnythingv2 [47]. Specifically, we conduct our distillation experiments using a subset of 200,000 samples from the SA-1B dataset [24].

For evaluation, we assess the performance of the distilled student model on five widely used depth estimation benchmarks, ensuring that these datasets remain unseen during training to enable a robust zero-shot evaluation. The chosen

benchmarks include: NYUv2 [39], KITTI [13], ETH3D [38], ScanNet [7], and DIODE [41]. Additional dataset details are provided in the Appendix.

Metrics. We assess depth estimation performance using two key metrics: the mean absolute relative error (AbsRel) and δ_1 accuracy. Following previous studies [34, 52, 22] on zero-shot MDE, we align predictions with ground truth in both scale and shift before evaluation.

Implementation. Our experiments use state-of-the-art monocular depth estimation models as teachers to generate pseudo-labels, supervising various student models in a distillation framework with only RGB images as input. In shared-context distillation, both teacher and student receive the same global region, extracted via random cropping from the original image. The crop maintains a 1:1 aspect ratio and is sampled within a range from 644 pixels to the shortest side of the image, then resized to 560×560 for local predictions. In global-local distillation, the global region is cropped into overlapping local patches, each sized 560×560 , for the teacher model to predict pseudo-labels. We use GenPercept [45] and DepthAnythingv2 (DAv2) [47] as teacher models for the multi-teacher mechanism. The learning rate is in tune with that of the corresponding student model. For DAv2 [47], the decoder learning rate is set to 5×10^{-5} . For the total loss function, we set the parameters as follows: $\lambda_1 = 0.5$, $\lambda_2 = 1.0$ and $\lambda_3 = 2.0$.

4.2. Analysis

For the ablation study and analysis, we sample a subset of 50K images from **SA-1B** [24] as our training data, with an input image size of 560×560 for the network. We conduct experiments on two of the most challenging benchmarks, DIODE [41] and ETH3D [38], which include both indoor and outdoor scenes. The model was trained with a batch size of 4 and converged after approximately 20,000 iterations on a single NVIDIA V100 GPU.

Impact of Normalization across Cross-Context Distillation. We evaluate the effect of different normalization strategies on Cross-Context Distillation, as shown in Table 1. The results indicate that the optimal normalization method varies across different distillation strategies. For shared-context distillation, no normalization achieves the best performance, assuming that pseudo-depth labels naturally reside in the same domain as predictions. For Local-Global distillation, Hybrid Normalization proves most effective, maintaining consistent depth predictions across regions through hierarchical normalization within specific depth ranges.

Ablation Study of Cross-Context Distillation. To further validate the effectiveness of our distillation framework, we conduct ablation studies by removing Shared-Context Distillation and Local-Global Distillation in Table 2. The results show that both components contribute significantly to

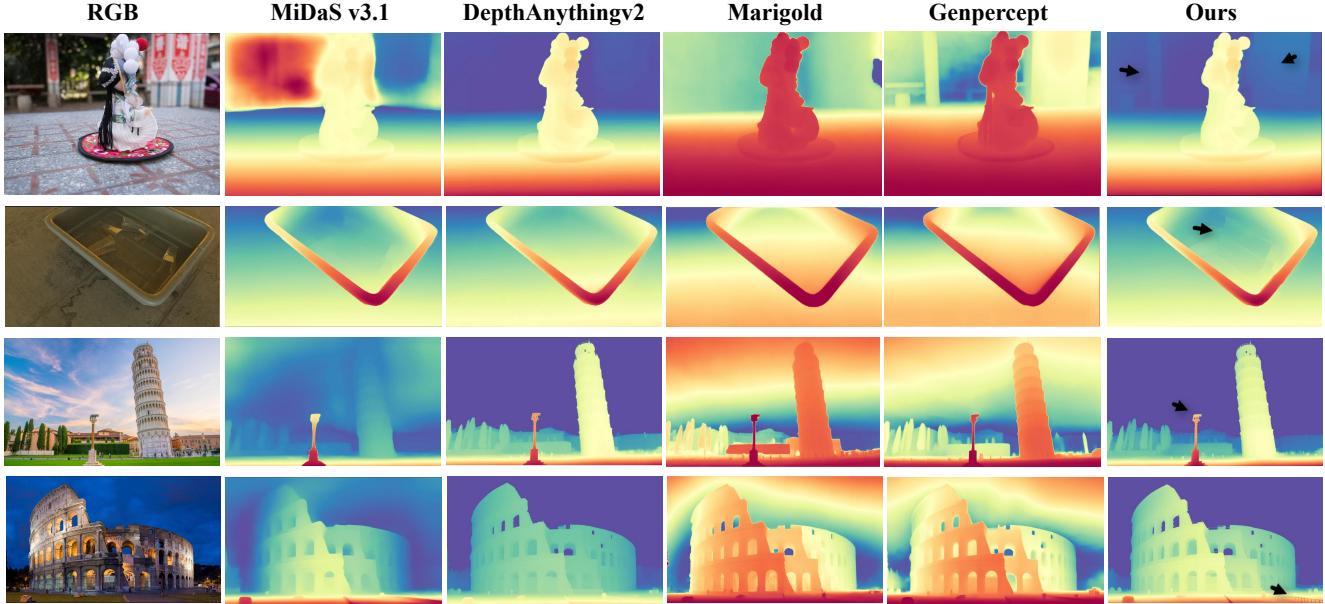


Figure 7: Qualitative Comparison of Relative Depth Estimations. We present visual comparisons of depth predictions from our method (“Ours”) alongside other classic depth estimators (“MiDaS v3.1” [3], and models using DINOv2 [31] or SD as priors (“DepthAnythingv2 [47]”, “Marigold” [22], “Genpercept” [45]). Compared to state-of-the-art methods, the depth map produced by our model, particularly at the position indicated by the **black arrow**, exhibits finer granularity and more detailed depth estimation.

Table 1: Analysis of Normalization Strategies. Performance comparison of different normalization strategies across Shared-Context Distillation and Local-Global Distillation.

Method	Normalization	ETH3D AbsRel↓	DIODE AbsRel↓
Shared-Context Distillation	Global Norm.	0.067	0.243
	No Norm.	0.058	0.236
	Local Norm.	0.060	0.238
	Hybrid Norm.	0.059	0.237
Local-Global Distillation	Global Norm.	0.065	0.253
	No Norm.	0.060	0.235
	Local Norm.	0.059	0.235
	Hybrid Norm.	0.056	0.232

Table 2: Effect of Cross-context Distillation. Performance comparison of various combinations of Shared-Context Distillation and Local-Global distillation on the ETH3D [38] and DIODE [41] datasets. The baseline corresponds to a simple shared-context approach with no random cropping. When neither method is applied, the model defaults to this baseline.

Shared-Context Distillation	Local-Global Distillation	ETH3D AbsRel↓	DIODE AbsRel↓
✗	✗	0.075	0.270
✗	✓	0.057 (-24.0%)	0.234 (-13.3%)
✓	✗	0.058 (-22.6%)	0.237 (-12.2%)
✓	✓	0.056 (-25.3%)	0.232 (-14.1%)

improving the student model’s ability to utilize pseudo-labels, demonstrating the robustness of our approach.

Table 3: Comparison in Cross-Architecture Distillation. Evaluation of our distillation pipeline in the context of Cross-Architecture Distillation. We adopt different architectures as teacher and student models, where the **Base** represents the previous distillation method [47]. Our method consistently improves the performance of the distilled student models.

Teacher	Student	Training Loss	DIODE	ETH3D
			AbsRel↓	AbsRel↓
DA-L	DA-S	Base	0.290	0.110
		Ours	0.262 (-9.6%)	0.098 (-10.9%)
DA-L	Midas-L	Base	0.313	0.147
		Ours	0.295 (-5.7%)	0.126 (-14.3%)
Midas-L	Midas-S	Base	0.303	0.150
		Ours	0.272 (-10.2%)	0.120 (-20.0%)

Cross-Architecture Distillation. To evaluate our normalization strategy, we conducted experiments using MiDaS [34] and DepthAnything [47], testing four configurations (DA-L, MiDaS-L, DA-S, MiDaS-S) as shown in Table 3. Our method consistently outperforms previous distillation approaches that use global normalization on the DIODE [41] and ETH3D [38] datasets, demonstrating superior performance both within and across architectures, and highlighting

Table 4: **Quantitative comparison of our multi-teacher distillation model on zero-shot benchmarks.** The **bold** values indicate the best performance. Our model, which integrates diverse depth estimation models, achieves higher accuracy than any individual teacher model.

Method	NYUv2		KITTI		DIODE		ScanNet		ETH3D		Avg. Rank
	AbsRel↓	$\delta 1 \uparrow$									
DepthAnything v2 NeurIPS'24	0.045	0.979	0.074	0.946	0.262	0.754	0.042	0.978	0.131	0.865	1.9
Genpercept(Disparity)ICLR'25	0.058	0.969	0.080	0.934	0.226	0.741	0.063	0.960	0.096	0.959	2.6
Ours(Multi-teacher)	0.043	0.981	0.077	0.945	0.298	0.756	0.042	0.979	0.065	0.983	1.4

Table 5: **Quantitative comparison with other affine-invariant depth estimators on several zero-shot benchmarks.** The **bold** values indicate the best performance, and underlined represent the second-best results.

Method	NYUv2		KITTI		DIODE		ScanNet		ETH3D	
	AbsRel↓	$\delta 1 \uparrow$								
DiverseDepth [51]	0.117	0.875	0.190	0.704	0.376	0.631	0.108	0.882	0.228	0.694
MiDaS [34]	0.111	0.885	0.236	0.630	0.332	0.715	0.111	0.886	0.184	0.752
LeReS [43]	0.090	0.916	0.149	0.784	0.271	0.766	0.095	0.912	0.171	0.777
Omnidata [9]	0.074	0.945	0.149	0.835	0.339	0.742	0.077	0.935	0.166	0.778
HDN [54]	0.069	0.948	0.115	0.867	0.246	<u>0.780</u>	0.080	0.939	0.121	0.833
DPT [36]	0.098	0.903	0.100	0.901	<u>0.182</u>	0.758	0.078	0.938	0.078	0.946
DepthAnything v2 [46]	<u>0.045</u>	0.979	0.074	0.946	0.262	0.754	0.042	<u>0.978</u>	0.131	0.865
Marigold [23]	0.055	0.961	0.099	0.916	0.308	0.773	0.064	0.951	0.065	0.960
Midas v3.1 [3]	-	0.980	-	<u>0.949</u>	-	-	-	-	0.061	0.968
Ours [†]	0.046	0.985	0.063	0.972	0.142	0.788	<u>0.049</u>	0.980	<u>0.057</u>	0.976
Ours*	0.043	<u>0.981</u>	<u>0.070</u>	<u>0.949</u>	0.233	0.753	0.042	0.980	0.054	0.981

[†] Refers to our method applied on the MiDaS v3.1. * Refers to our method applied on the DepthAnythingv2-Large.

the limitations of global normalization in pseudo-label distillation.

Multi-teacher Mechanism. We evaluate the effectiveness of our multi-teacher distillation strategy across five benchmarks in Table 4. To handle the diverse output depth distributions of different teacher models, we use Hybrid Normalization for Shared-Context Distillation in this experiment. Using diffusion-based Genpercept [45] and Dinov2-based DepthAnythingv2 [47] as teacher models, we train a lightweight DPT-based depth estimation model. Our approach outperforms both teacher models overall, with only a minor gap on KITTI [13], demonstrating the effectiveness of multi-teacher distillation.

4.3. Comparison with State-of-the-Art

Quantitative Analysis. Our model achieves SOTA performance across both indoor and outdoor datasets, demonstrating strong generalization from structured indoor scenes (NYUv2 [39], ScanNet [7]) to complex outdoor environments (KITTI [13], DIODE [41], ETH3D [38]), as shown in Table 5. By optimizing pseudo-label distillation and depth normalization, our student model not only surpasses

its teacher but also achieves a new SOTA on multiple benchmarks, demonstrating the effectiveness of our approach.

Qualitative analysis. We show a qualitative comparison of different depth estimations between SOTA models and the proposed method in Fig. 7. Compared with DAv2 [47], our method preserves finer details, particularly in areas marked by arrows. Although Marigold [22] and Genpercept [45] generate detailed maps using generative priors, they struggle with correct relative depth relationships. In contrast, our model preserves fine details while maintaining accurate relative depth relationships, resulting in a visually consistent and reliable depth estimation.

5. Conclusion

In this work, we study pseudo-label distillation strategies for MDE. We find that the widely used SSI normalization amplifies noise in teacher-generated pseudo-labels, impairing local depth accuracy. To address the problem, we propose Cross-Context Distillation, which combines local refinement with global consistency, enabling the model to learn fine details and structural context. Our multi-teacher framework, integrating diffusion-based models and encoder-decoder net-

works, achieves state-of-the-art performance on multiple benchmarks. Future work could improve the efficiency of unlabeled data distillation.

References

- [1] Ali Jahani Amiri, Shing Yan Loo, and Hong Zhang. Semi-supervised monocular depth estimation with left-right consistency using deep neural network. In *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 602–607, 2019. [3](#)
- [2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv: 2302.12288*, 2023. [2](#)
- [3] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation, 2023. [7](#), [8](#)
- [4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. [2](#)
- [5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. [2](#)
- [6] Jaehoon Cho, Dongbo Min, Youngjung Kim, and Kwanghoon Sohn. A large rgb-d dataset for semi-supervised monocular depth estimation. *arXiv preprint arXiv: 1904.10230*, 2019. [3](#)
- [7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839. IEEE, 2017. [6](#), [8](#), [12](#)
- [8] John Doe and Jane Smith. Patchfusion: Multi-scale feature fusion for enhanced depth estimation. *International Journal of Computer Vision*, 131:1234–1250, 2023. [3](#)
- [9] Adnan Eftekhar, Mate Balog, et al. Omnidata: A pipeline for building synthetic data of complex 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. [8](#)
- [10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2366–2374, 2014. [1](#), [2](#)
- [11] Huazhu 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 (CVPR)*, pages 2002–2011, 2018. [2](#)
- [12] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised learning of depth and ego-motion from video. In *European Conference on Computer Vision*, pages 556–573. Springer, 2016. [1](#)
- [13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361. IEEE, 2012. [6](#), [8](#), [12](#)
- [14] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Monodepth2: Self-supervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 168–176, 2019. [3](#)
- [15] 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 (CVPR)*, pages 270–279, 2017. [2](#)
- [16] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2485–2494, 2020. [1](#)
- [17] Vitor Guizilini, Jie Li, Rares Ambrus, Sudeep Pillai, and Adrien Gaidon. Robust semi-supervised monocular depth estimation with reprojected distances. In *Conference on robot learning*, pages 503–512. PMLR, 2020. [3](#)
- [18] Lukas Hoyer, Dengxin Dai, Qin Wang, Yuhua Chen, and Luc Van Gool. Improving semi-supervised and domain-adaptive semantic segmentation with self-supervised depth estimation. *International Journal of Computer Vision*, 131(8):2070–2096, 2023. [3](#)
- [19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018. [2](#)
- [20] 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. *arXiv preprint arXiv:2404.15506*, 2024. [2](#)
- [21] Rongrong Ji, Ke Li, Yan Wang, Xiaoshuai Sun, Feng Guo, Xiaowei Guo, Yongjian Wu, Feiyue Huang, and Jiebo Luo. Semi-supervised adversarial monocular depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:2410–2422, 2020. [3](#)
- [22] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. 2024. [1](#), [3](#), [6](#), [7](#), [8](#), [12](#)
- [23] Qianli Ke, Hanxiao Lu, Yingcong Zhang, et al. Marigold: Multi-modal 3d perception with diffusion models. *arXiv preprint arXiv:2402.04567*, 2024. [8](#)
- [24] Alexander Kirillov, Eric Mintun, et al. Sa-1b: Segment anything 1-billion mask dataset. <https://segment-anything.com>, 2023. [6](#), [12](#)
- [25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint arXiv: 2304.02643*, 2023. [1](#)
- [26] Yevhen Kuznetsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction.

- In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6647–6655, 2017. 3
- [27] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *Proceedings of the Fourth International Conference on 3D Vision (3DV)*, pages 239–248, 2016. 2
- [28] Yanhua Li, Qixing Zhang, and Liqian Zhang. Ar shadow: Real-time 3d object tracking and shadow rendering for mobile augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 26(9):2871–2881, 2020. 1
- [29] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 1
- [30] S. M. H. Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yağız Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. *Computer Vision and Pattern Recognition*, 2021. 3
- [31] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DiNov2: Learning robust visual features without supervision. *TMLR*, 2024. 7
- [32] Andra Petrovai and Sergiu Nedevschi. Exploiting pseudo labels in a self-supervised learning framework for improved monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1578–1588, 2022. 3
- [33] Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, S. Mattoccia, and L. D. Stefano. Geometry meets semantics for semi-supervised monocular depth estimation. *Asian Conference on Computer Vision*, 2018. 3
- [34] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 1, 2, 3, 4, 6, 7, 8, 12
- [35] Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021. 2
- [36] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 8
- [37] Rene 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. 2
- [38] Thomas Schöps, Torsten Sattler, and Marc Pollefeys. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3260–3269. IEEE, 2017. 6, 7, 8, 12
- [39] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*, pages 746–760. Springer, 2012. 6, 8, 12
- [40] Nikolai Smolyanskiy, Alexey Kamenev, and Stan Birchfield. On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1007–1015, 2018. 3
- [41] Igor Vasiljevic, Ayan Chakrabarti, Vladlen Koltun, and Jack Tumblin. Diode: A dense indoor and outdoor depth dataset. In *IEEE International Conference on Computer Vision (ICCV)*, pages 896–905. IEEE, 2019. 6, 7, 8, 12
- [42] 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, 2024. 2
- [43] Zhenyu Wei, Andreas Geiger, et al. Leres: Learning-based monocular depth estimation for all scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021. 1, 8
- [44] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 1
- [45] Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What matters when repurposing diffusion models for general dense perception tasks? *arXiv preprint arXiv:2403.06090*, 2024. 1, 3, 6, 7, 8, 12
- [46] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data, 2024. 1, 3, 8
- [47] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 1, 2, 3, 6, 7, 8, 12
- [48] Nan Yang, Rui Wang, Jörg Stückler, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1281–1292, 2020. 1
- [49] Nan Yang, Rui Wang, J. Stückler, and D. Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. *European Conference on Computer Vision*, 2018. 3
- [50] Weicong Yin, Jianping Shi, and Yao Feng. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2042–2051, 2021. 1

- [51] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Renyin Dou. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*, 2020. [1](#), [2](#), [8](#)
- [52] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. *IEEE International Conference on Computer Vision*, 2023. [2](#), [6](#), [12](#)
- [53] 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. *Computer Vision and Pattern Recognition*, 2020. [2](#)
- [54] Chi Zhang, Wei Yin, Billzb Wang, Gang Yu, Bin Fu, and Chunhua Shen. Hierarchical normalization for robust monocular depth estimation. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. [3](#), [4](#), [8](#)
- [55] Chi Zhang, Wei Yin, Gang Yu, Zhibin Wang, Tao Chen, Bin Fu, Joey Tianyi Zhou, and Chunhua Shen. Robust geometry-preserving depth estimation using differentiable rendering. *IEEE International Conference on Computer Vision*, 2023. [2](#)
- [56] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017. [2](#)
- [57] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1858, 2017. [2](#)
- [58] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, and Quoc V Le. Rethinking pre-training and self-training. In *Advances in Neural Information Processing Systems*, volume 33, pages 3833–3845, 2020. [1](#)

6. Appendix

6.1. Dataset Details

Datasets. We train our model on **SA-1B** [24], a large-scale dataset covering diverse indoor and outdoor environments, enabling robust depth learning for real-world scenes. For evaluation, we use established monocular depth benchmarks:

- **NYUv2** [39]: Indoor depth estimation and semantic segmentation.
- **KITTI** [13]: Autonomous driving dataset with outdoor scenes and high-quality LiDAR depth.
- **ETH3D** [38]: High-resolution stereo images for indoor/outdoor depth estimation and 3D reconstruction.
- **ScanNet** [7]: Large-scale RGB-D dataset for 3D scene reconstruction and semantic segmentation.
- **DIODE** [41]: Dense, high-quality depth maps for both indoor and outdoor environments.

Metrics. We evaluate depth estimation using mean absolute relative error (AbsRel) and δ_1 accuracy. AbsRel is defined as:

$$AbsRel = \frac{1}{M} \sum_{i=1}^M \frac{|d_i - d_i^*|}{d_i^*} \quad (10)$$

where d_i is the predicted depth, d_i^* is the ground truth, and M is the total number of depth values. δ_1 accuracy measures the percentage of pixels where:

$$\delta_1 = \max \left(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i} \right) < 1.25 \quad (11)$$

indicating prediction accuracy within a specific tolerance. Following Metric3D [34, 52, 22], we align predictions with ground truth in scale and shift before evaluation.

6.2. More Experiments

Effect of Data Scaling. To investigate how dataset size affects model performance, we conducted experiments using progressively larger training datasets and compared our method against the SSI Loss baseline. Fig. 8 shows the Absolute Relative Error (AbsRel) as the dataset size increases from 10K to 200K images.

Distilling Generative Models vs. DepthAnything2. Beyond distilling encoder-decoder depth models, we extend our approach to generative models, specifically Genpercept [45], aiming to transfer their superior detail preservation to a more efficient student model. While diffusion-based depth estimators achieve fine-grained depth reconstruction, their high computational cost limits practical applications. We investigate whether their depth estimation capability can be effectively distilled into a lightweight DPT-based model. Experimental results in Fig. 9 show that compared to using

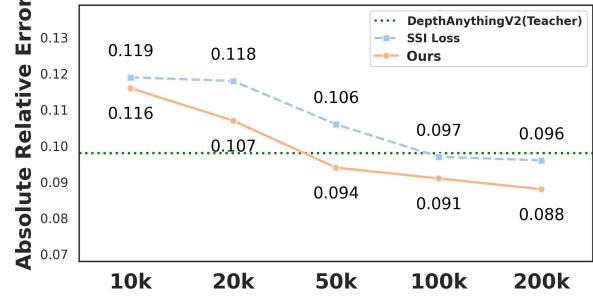


Figure 8: **Comparison of Data Scaling**. Performance comparison of our model with SSI Loss as the dataset size increases, measured by the average AbsRel. The results indicate that our method consistently outperforms the baseline method.

DepthAnythingv2 as the teacher, distilling from a diffusion-based model yields a student model with significantly enhanced fine-detail prediction.

Qualitative Comparison with Baseline Distillation. We present a qualitative comparison between our method and the previous distillation method [47], where the **Base** model relies solely on global normalization. We analyze the depth map details and the distribution differences between predicted and ground truth depths. The red diagonal lines represent the ground truth, with results closer to these lines indicating better performance. As shown in Fig. 10, our method produces smoother surfaces, sharper edges, and more detailed depth maps.

Qualitative Comparison: Additional Results on Depth Estimation in the Wild. We present additional depth maps generated by our model on in-the-wild scenes, emphasizing its robustness and precision. As shown in Fig. 11, our method produces sharper edges and more detailed depth maps, even in challenging regions such as hair, cartoon scenes, and other diverse environments.

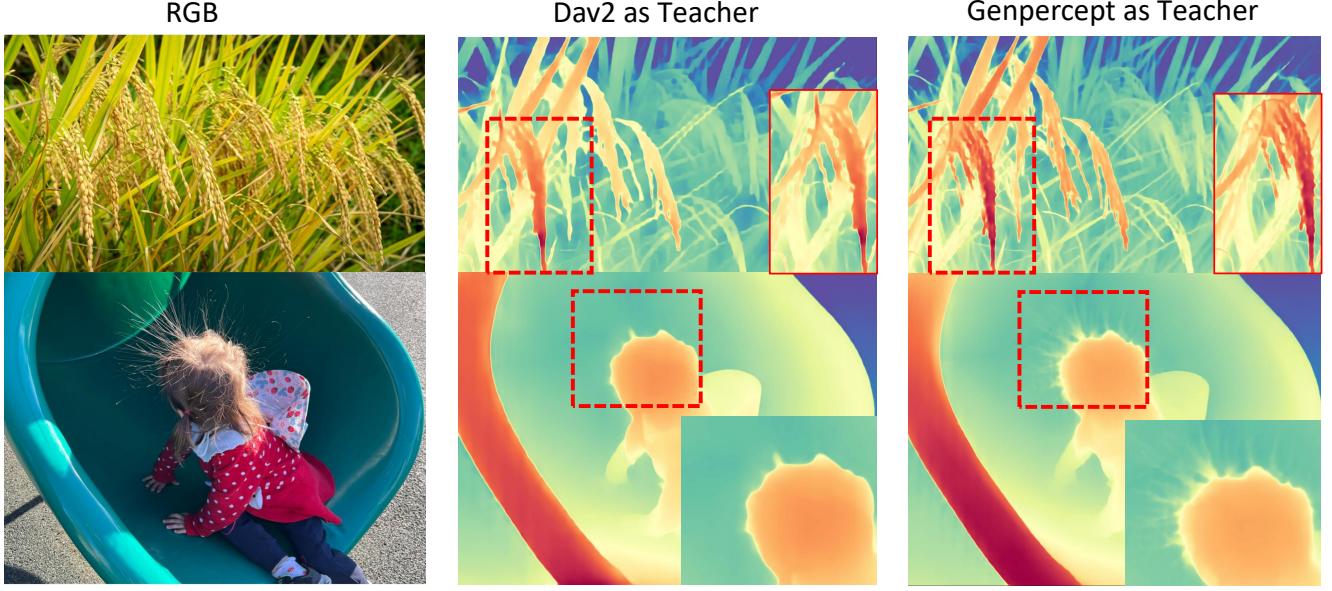


Figure 9: **Distilled Generative Models:** Instead of just distilling classical depth models, we also apply distillation to generative models, aiming for the student model to capture their rich details.

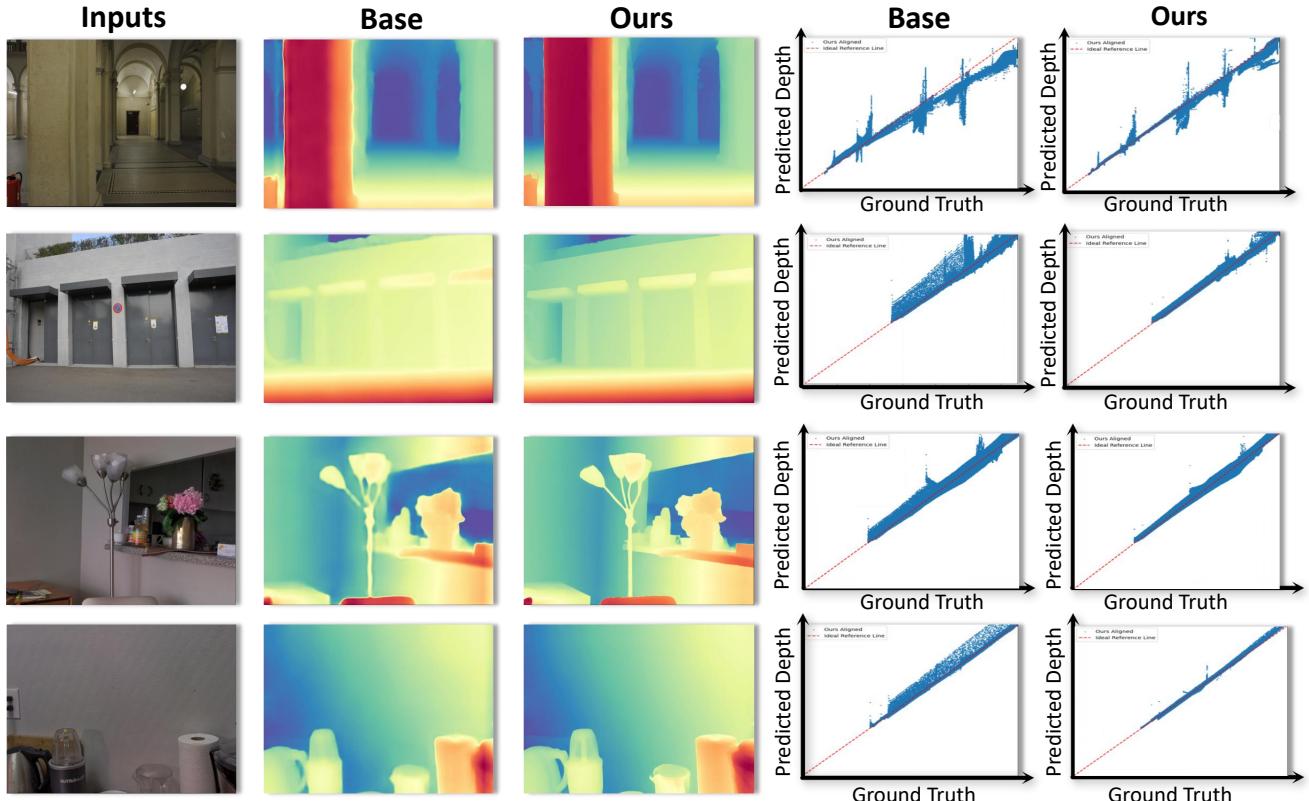


Figure 10: **Qualitative Comparison with Baseline Distillation.** We compare our method with the baseline as the previous distillation method, which uses only global normalization. The red diagonal lines represent the ground truth, with results closer to the lines indicating better performance. Our method produces smoother surfaces, sharper edges, and more detailed depth maps.

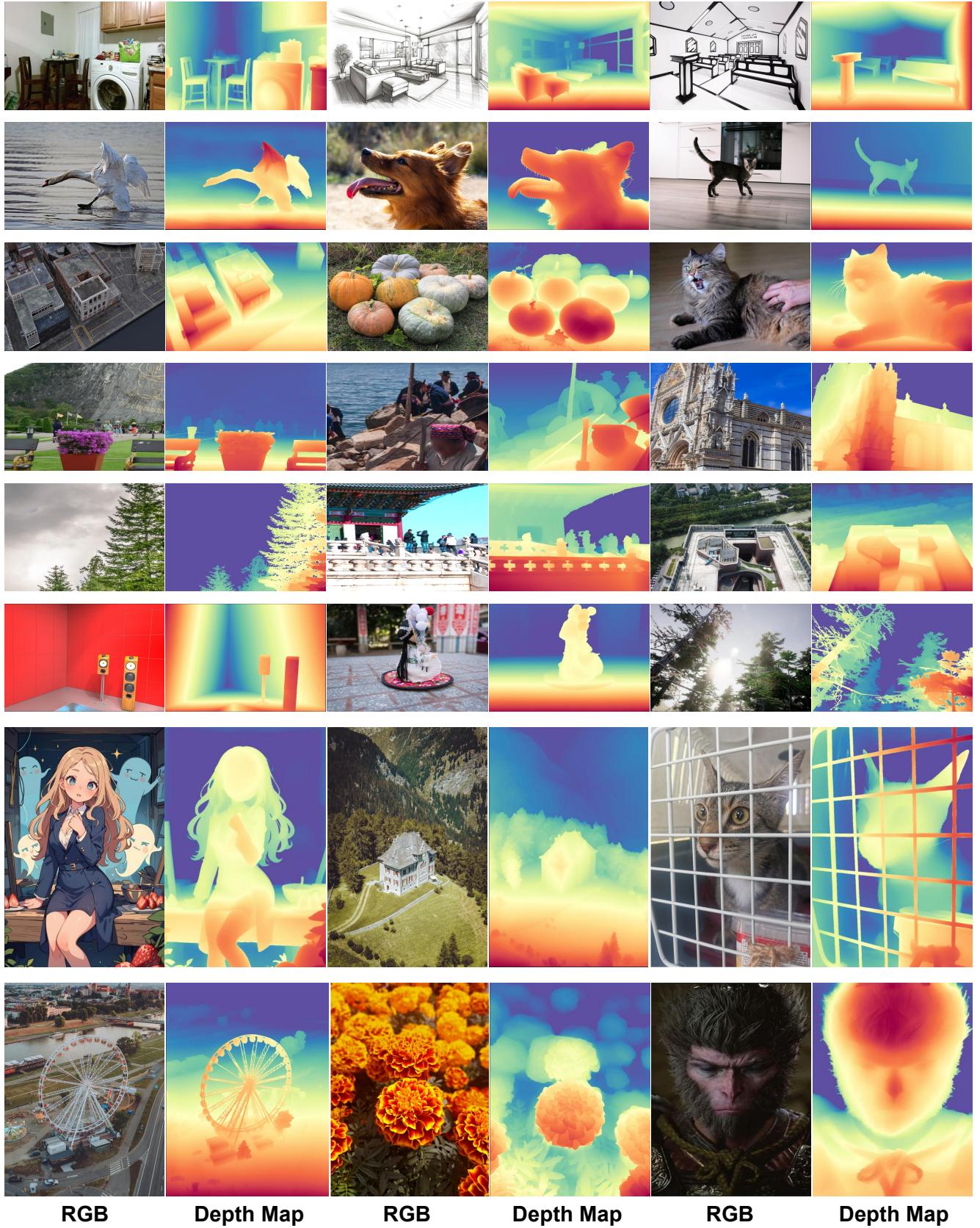


Figure 11: **Additional Results on Depth Estimation in the Wild.** We showcase more depth maps generated by our model on in-the-wild scenes, highlighting its robustness and precision.