Learning to Discriminate Scene Structures Makes Self-Supervised Depth Learning Scalable

Mengtan Zhang[1] **Yuchuan Cui**[2] **Yu Ma**[2] **Zizhan Guo**[2] **Wei Ye**[1,2] **Rui Fan**[1,2,3,✉]

[1]Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University
[2]College of Electronic and Information Engineering, Tongji University
[3]State Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University

zmt200110@tongji.edu.cn, rui.fan@ieee.org

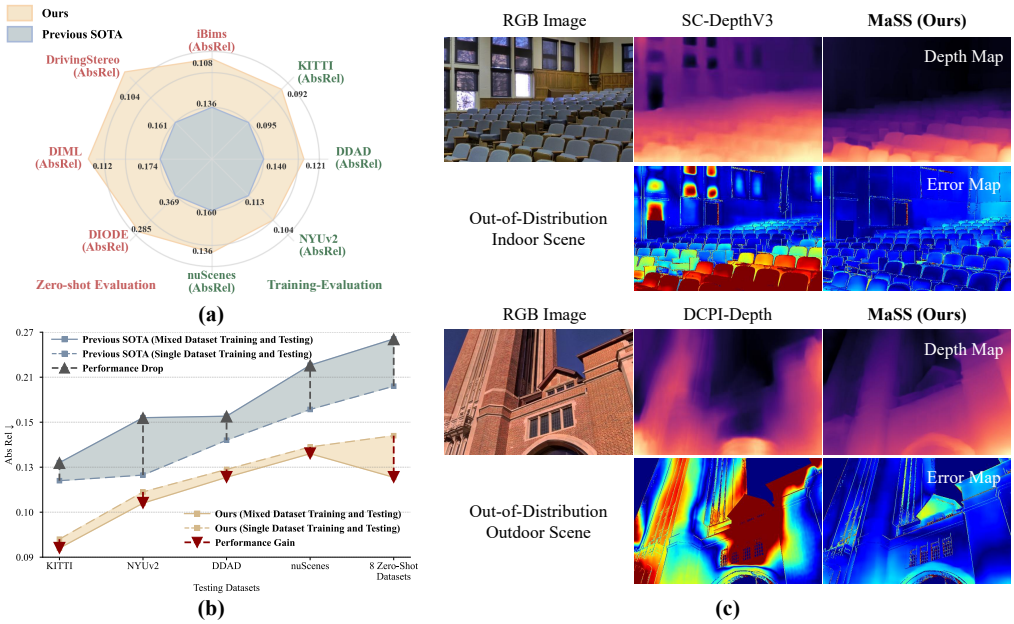


Figure 1: Beyond (a) achieving SoTA performance across diverse datasets with a unified model, MaSS is the first SSMDE method to (b) effectively benefit from increased data diversity and (c) exhibit remarkable generalizability to entirely out-of-distribution scenes.

Abstract

Self-supervised monocular depth estimation is appealing for its potential to scale with unlabeled data, yet existing methods often degrade when trained on diverse scenes. This study finds that, under self-supervision, models tend to encode dataset-specific structural biases rather than transferable scene geometry, leading to cross-domain optimization conflicts. To address this issue, this study proposes the MaSS framework, aiming to make self-supervised depth learning scalable to diverse data. MaSS introduces a dedicated structural representation pathway explicitly decoupled from contextual features, which is guided by a novel scene-structure discrimination objective to learn generalized, layout-aware structural representations for depth decoding. Trained and evaluated on four diverse datasets, both individually and jointly, MaSS substantially outperforms prior self-supervised methods and is the first to consistently benefit from mixed-domain training. Zero-shot evaluation on eight additional datasets further demonstrates its remarkable generalizability, particularly on those entirely out-of-distribution scenes. The source code is publicly available at https://mias.group/MaSS.

✉ Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

The fundamental advantage of self-supervised learning lies in its potential to leverage ubiquitous unlabeled data for scalable model training. However, most existing self-supervised monocular depth estimation (SSMDE) research [1, 2, 3, 4, 5] is predominantly confined to training models individually on single datasets or scenarios, leaving much of this potential unexplored. Recent efforts have attempted to directly train models with existing SSMDE methods on large-scale video sequences curated from online platforms [6, 7], as well as to incorporate vision foundation models (VFMs) as backbones [8, 7]. Despite these advances, overall performance improvements remain limited. As demonstrated by our preliminary experiments (see Fig. 1(b)), when existing methods are trained simultaneously on diverse monocular video data (*e.g.*, indoor, outdoor, dynamic, and adverse weather scenarios), they fail to benefit mutually from the increased diversity and do not exhibit improved generalization to out-of-distribution scenes. Instead, we observe a performance decline compared to models trained individually on a single dataset. This observation demonstrates that existing SSMDE methods have not yet achieved effective scalability to diverse data.

This study begins by investigating the challenges that arise when SSMDE is exposed to diverse data distributions during training. It is worth noting that the novel-view synthesis-based supervision inevitably introduces noisy and artifact-laden supervisory signals in scenarios involving occlusions, textureless regions, and illumination changes [9, 10, 11, 12]. Even setting aside such noise, the gradients induced by self-supervised losses often fail to reflect depth errors faithfully [4]. As a result, the optimization process is inherently prone to local minima, leading to poor convergence stability [4]. Although these challenges frequently occur during self-supervised training, existing methods can still achieve reasonable accuracy on a single, homogeneous dataset [3]. This is largely because the network can exploit the structural bias inherent to that dataset, thereby compensating for weak supervision. The key evidence we find is that under the existing SSMDE framework, the coarsest, high-level feature of the depth encoder does not capture depth-related cues, such as the foreground–background relation and perspective geometry. Instead, it primarily reflects the dataset-specific structural biases. When training on diverse datasets, different data domains often require the model to learn inconsistent structural biases for their respective convergence, leading to the observed performance drop.

These findings motivate us to explicitly guide the network to learn a robust, generalized scene-structure representation for depth decoding. Based on this core idea, we develop the **MaSS** framework that **Ma**ke **S**elf-supervised depth learning **S**calable to diverse data. MaSS bridges a conventional depth encoder with a frozen pretrained VFM through a bidirectional interaction architecture, drawing inspiration from [13]. In this design, the depth encoder focuses on extracting geometric cues that are difficult to recover from VFM features [8]. The interacted VFM features are then fed into two feature projectors to reconstruct complementary, collapse-free representations under explicit constraints and learning objectives. Specifically, one projector is adapted to learn a contextual representation informative for depth estimation, such as object-level spatial relationships, while retaining generalized semantics with the original VFM features as anchors. The other projector is then constrained to be decoupled from the contextual one, thereby reserving its representational capacity for scene structure. To explicitly define this representation, we introduce a scene-structure discrimination objective via contrastive learning. This objective organizes the feature space based on scene layout similarity, which is measured by the histogram distribution of predicted depth. Driven by this objective, the second projector learns transferable structural representations that capture global scene layout and perspective geometry, rather than dataset-specific biases. Together, these two complementary representations enable more reliable depth predictions while reducing their reliance on dataset-specific biases.

Extensive experiments demonstrate that MaSS outperforms previous state-of-the-art (SoTA) self-supervised methods by a large margin (as presented in Fig. 1(a)) and achieves accuracy comparable to certain supervised depth foundation models. Moreover, MaSS represents the first self-supervised approach to achieve clear performance gains from increasing the diversity of training data (Fig. 1(b)), resulting in notable generalizability improvements particularly on out-of-distribution scenes (Fig. 1(c)).

2 Related Works

To alleviate the reliance on extensive ground-truth data in monocular depth estimation (MDE), SSMDE from monocular video, first introduced in the study [14], has attracted growing attention. Within this paradigm, improvements have focused on enhancing photometric supervision [1], addressing

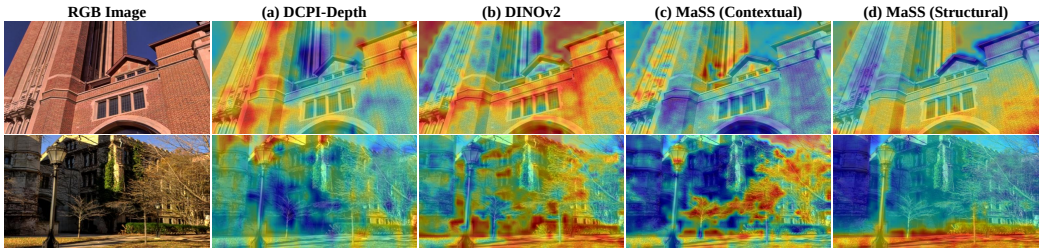


Figure 2: Feature activation map comparison among a conventional self-supervised depth model, DINOv2, and the two representations learned by MaSS. (a) The conventional depth model tends to encode dataset-specific biases, whereas (b) DINOv2 lacks sufficient geometric cues. In contrast, MaSS decouples features into contextual and structural representations: (c) contextual ones capture fine-grained object details and foreground–background relations; (d) structural ones encode global scene layouts that are transferable across diverse datasets.

scale inconsistency [15], and incorporating additional visual cues to handle occlusions and dynamic scenes [16, 3, 17, 18, 4]. Though significant progress has been made, these SSMDE methods are typically trained on single domains with limited-scale datasets and rely on task-specific network architectures [19, 20, 21, 22]. The study [23] first demonstrates the feasibility of self-supervised depth learning from heterogeneous in-the-wild videos by jointly estimating depth, camera motion, object motion, and camera intrinsics. With the rise of VFMs [24, 25], increasing efforts have been devoted to developing depth foundation models (DFMs) [26, 27, 28, 29] trained on large-scale annotated data, achieving strong performance and generalization. Building upon this trend, recent studies [6, 7] have taken initial steps toward scaling SSMDE training, while the others [9, 30, 5] demonstrate that developing DFMs in a self-supervised fashion requires not only harnessing the large pretrained models' capacity, but also addressing challenges unique to self-supervised learning. Among them, cross-domain learning conflicts under mixed-domain training pose a fundamental yet underexplored obstacle to scalability, which this work primarily aims to address.

Three lines of research inform our approach. The first draws on pretrained knowledge to improve depth representations. For example, the study [31] uses a pretrained semantic network to guide the learning of geometric features. The studies [32, 33] further emphasize preserving transferable pretrained representations during adaptation. The second focuses on organizing depth representations. The prior arts [34, 35, 36, 37, 30] have introduced contrastive objectives that structure the feature space according to weather variations, color or geometric consistency, historical predictions, and relationships guided by language. The third conditions depth prediction distributions on scene characteristics to account for domain shifts. For example, the study [38] conditions these distributions on scene categories. Drawing on these insights, MaSS explicitly decouples contextual and structural representations, preserves contextual relationships derived from VFMs, and adaptively organizes structural features according to similarities in scene layout. Depth prediction is then conditioned on the resulting structural representations to reduce learning conflicts across domains.

3 Methodology

SSMDE models are prone to learning dataset-specific inductive biases and struggle to generalize across varying depth distributions during both training and inference. For example, Fig. 1(c) presents depth predictions on out-of-distribution indoor and outdoor scenes from two representative self-supervised depth models [3, 4], trained on the NYUv2 [39] and KITTI [40] datasets, respectively. It can be observed that the model trained on driving scenes tends to predict depth with a vanishing point, while the model trained on small rooms fails to capture the more uniform depth distributions in large halls. To find the source of such inductive biases, we investigate the internal feature representations. As shown in Fig. 2(a), the DepthNet encoder [4] trained on driving data produces feature activation maps with a consistent low-activation region. This pattern closely resembles the biased depth predictions in Fig. 1(c) and fails to reflect variations aligned with the actual scene geometry. These results indicate that the dataset-specific inductive biases are already embedded in the intermediate representations. The above findings inspire us with a possible solution to make self-supervised learning scalable: explicitly guide the network to learn robust, generalized structural representations that are discriminative of scene layout before decoding depth.

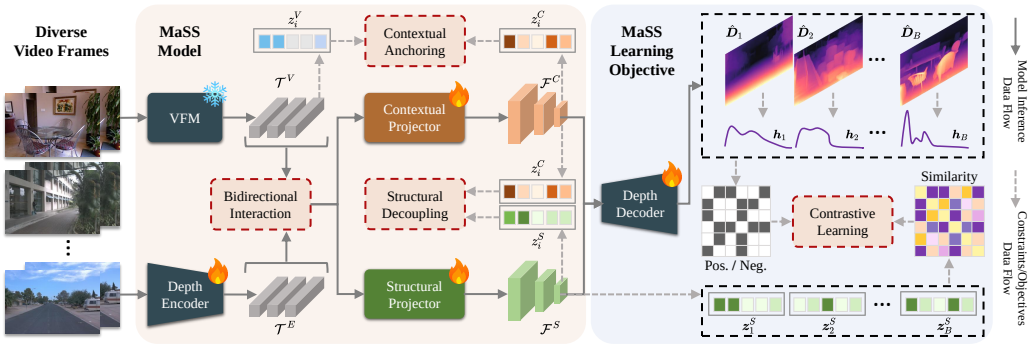


Figure 3: The overall architecture of the proposed MaSS framework. MaSS bridges a conventional depth encoder with a frozen pretrained VFM via a bidirectional interaction module. The interacted VFM features are fed into a pair of anchored–decoupled feature projectors, producing collapse-free contextual and structural representations. The structural projectors learn generalized representations from discriminating scene structures, making the self-supervised depth learning scalable.

3.1 MaSS Model

The first restriction of existing approaches towards this goal is from the model architecture design. Although the pretrained VFM (*e.g.*, DINOv2 [24]) exhibits strong generalization and is widely adopted as initialization for supervised depth foundation models [26, 41, 42], its adaptation is unreliable under self-supervision due to noisy training signals [9]. Direct fine-tuning may deteriorate the pretrained representations [43], yet keeping the VFM frozen leaves its semantic features containing less sufficient geometric cues for structural modeling [8] (as seen in Fig. 2(b)). This dilemma motivates a bidirectional interaction architecture between a trainable traditional depth encoder and a frozen VFM encoder, drawing on the ViT-Adapter design [13]. Another key observation is that different viewpoints of the same scene, or different scenes containing similar objects, may appear contextually alike yet differ substantially in spatial structure. Conversely, visually distinct scenes can share similar structural layouts. Such misalignment suggests that entangling contextual and structural features within a single representation pathway may lead to representational conflict and collapse. To address this issue, we decouple contextual and structural representations using two specialized feature projectors.

Bidirectional interaction architecture. As shown in Fig. 3, we first split the ViT blocks of the VFM encoder into three subsets and extract an intermediate token set $\mathcal{T}^V = \{\boldsymbol{T}_i^V\}_{i=1}^3$, which consists of D-dimensional feature tokens $\boldsymbol{T}_i^V \in \mathbb{R}^{\frac{HW}{16^2} \times D}$. Meanwhile, the depth encoder [21] extracts a multi-scale feature pyramid $\mathcal{F}^E = \{\boldsymbol{F}_i^E\}_{i=1}^3$, corresponding to spatial resolutions of $1/8, 1/16, 1/32$, respectively. We apply 1×1 convolution to unify the channel dimensions, followed by flattening and concatenation, to obtain D-dimensional feature tokens $\boldsymbol{T}_1^E \in \mathbb{R}^{\left(\sum_i \frac{HW}{S_i^2}\right) \times D}$, where $S_i = 2^{i+2}$. The two branches exchange complementary information through the interaction module introduced in [13]. Concretely, the VFM tokens are updated via cross-attention ($\mathrm{Attn}(\cdot)$) with the depth tokens:

$$\hat{\boldsymbol{T}}_i^V = \boldsymbol{T}_i^V + \mathrm{Attn}(\mathrm{LN}(\boldsymbol{T}_i^V), \mathrm{LN}(\boldsymbol{T}_i^E)), \tag{1}$$

where $\mathrm{LN}(\cdot)$ is LayerNorm, while the depth tokens are updated by attending to the VFM tokens, followed by a feed-forward network ($\mathrm{FFN}(\cdot)$):

$$\hat{\boldsymbol{T}}_{i+1}^E = \hat{\boldsymbol{T}}_i^E + \mathrm{FFN}(\mathrm{LN}(\hat{\boldsymbol{T}}_i^E)), \quad \hat{\boldsymbol{T}}_i^E = \boldsymbol{T}_i^E + \mathrm{Attn}(\mathrm{LN}(\boldsymbol{T}_i^E), \mathrm{LN}(\boldsymbol{T}_{i+1}^V)) \tag{2}$$

This bidirectional interaction aligns heterogeneous features within a shared latent space while injecting learnable geometric cues from the depth encoder into the frozen VFM representations. Notably, unlike its common use as an auxiliary module in other fields [44, 45], this architecture plays an indispensable role in enabling structural representation learning and scalable self-supervised training in our work, as demonstrated by the ablation results in Sect. 4.4.

Anchored-decoupled feature projectors. Next, we introduce two feature projectors that take the interacted VFM features as input and progressively upsample them to produce decoder-compatible feature pyramids $\mathcal{F}^C = \{\boldsymbol{F}_i^C\}_{i=1}^3$ and $\mathcal{F}^S = \{\boldsymbol{F}_i^S\}_{i=1}^3$, corresponding to contextual and structural

representations, respectively. Concretely, $\mathcal{T}^V$ are reshaped from sequence form into spatial feature maps $\mathcal{F}^V = \{\hat{\boldsymbol{F}}_i^V\}_{i=1}^3$, which are used to compute feature pyramids as

$$\boldsymbol{F}_i^k = \mathrm{ReLU}\left(\mathrm{BN}^k\left(\mathrm{TransConv}^k_{2^{i-1}\times 2^{i-1}}\left(\hat{\boldsymbol{F}}_i^V\right)\right)\right), \quad k \in \{C, S\},\ i \in \{1, 2, 3\}. \tag{3}$$

where $\mathrm{TransConv}^{\mathrm{k}}(\cdot)$ denotes transposed convolution, $\mathrm{BN}^{\mathrm{k}}(\cdot)$ represents batch normalization, and $\mathrm{ReLU}(\cdot)$ is the ReLU activation. To enforce robust and generalizable adaptation of the contextual representation while encouraging the structural pathway to focus on the complementary geometric cues, we introduce an anchoring-decoupling constraint. Specifically, we aggregate the highest-level feature maps $\boldsymbol{F}_3^C$, $\boldsymbol{F}_3^S$, and $\boldsymbol{F}_3^V$ from the VFM into latent descriptors via global average pooling followed by normalization. For a batch of size B, this yields three sets of normalized descriptors: $\{\boldsymbol{z}_i^C\}_{i=1}^B$, $\{\boldsymbol{z}_i^S\}_{i=1}^B$, and $\{\boldsymbol{z}_i^V\}_{i=1}^B$. The introduced constraint can then be formulated as follows:

$$\mathcal{L}_{\text{ac}} = \sum_{i=1}^{B}\sum_{j=1, j\neq i}^{B}(\boldsymbol{z}_i^{C\top}\boldsymbol{z}_j^C - \boldsymbol{z}_i^{V\top}\boldsymbol{z}_j^V)^2 + \sum_{i=1}^{B}(\boldsymbol{z}_i^{C\top}\boldsymbol{z}_i^S)^2, \tag{4}$$

where the first term anchors the contextual features to the semantic relationships captured by the VFM across samples, ensuring that semantic context is preserved during task-specific adaptation. The second term decouples the structural features from contextual ones, encouraging them to focus solely on the complementary geometric cues that reflect scene layout and viewpoint variations, rather than redundant semantic content.

The above two architectural designs jointly establish a geometrically enriched and collapse-free structural pathway for representation learning, for which the corresponding learning objective is introduced in the following section.

3.2 MaSS Learning Objective

Existing self-supervision signals mainly penalize local photometric inconsistencies [4, 9], which predominantly supervise contextual features and fail to guide the learning of global scene structure. Consequently, an additional objective is required to formulate structural representation learning. However, directly supervising structural representations is infeasible, as the mapping between intermediate feature distributions and the underlying depth distribution is not analytically tractable. We find an alternative solution drawing inspiration from contrastive learning [46, 47], which organizes the feature space through relational supervision rather than explicit labels. Unlike prior works [30, 34, 37] that construct contrastive pairs in a generic instance-level manner [46], we reformulate a dedicated scene-structure discrimination task, where positive and negative pairs are explicitly defined based on a proposed structural discrepancy metric.

Structural discrepancy metric. Given a training batch denoted as $\{\boldsymbol{I}_i, \hat{\boldsymbol{D}}_i\}_{i=1}^B$ of size B, where $\hat{\boldsymbol{D}}_i \in \mathbb{R}^{H\times W}$ represents the predicted depth map corresponding to image $\boldsymbol{I}_i$, the structural discrepancy metric aims to reflect the similarity between samples in the batch in terms of their global depth distribution. A straightforward approach is to compare their depth histograms. However, such histogram-based comparisons are highly sensitive to local structural variations and high-frequency artifacts [48, 49]. To emphasize global distributional differences rather than local details, we first downsample each depth map $\hat{\boldsymbol{D}}_i$ by a factor of s to obtain a coarse proxy map $\boldsymbol{M}_i \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}}$, where $s = 8$ in our implementation. We then model the depth distribution of $\boldsymbol{M}_i$ as a discrete probability distribution over K bins, with $\{c_i\}_{i=1}^K$ being the bin centers and Δ denoting the interval width. K is set to 20. To further mitigate the effect of high-frequency artifacts, we adopt a soft binning strategy inspired by [50] when computing the depth histogram. Instead of hard assignments, each depth value contributes to nearby bins with continuous weights. Concretely, for a depth value $d \in \boldsymbol{M}_i$, its contribution to the k-th bin is defined as the following function:

$$\Psi_k(d) = \sigma\left(\left(d - c_k + \Delta/2\right)/\tau_h\right) - \sigma\left(\left(d - c_k - \Delta/2\right)/\tau_h\right), \tag{5}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and τ_h is a temperature parameter that controls the smoothness of the bin assignment and suppresses high-frequency perturbations. Based on this formulation, we define a histogram-like probability vector that summarizes the global depth distribution of the i-th sample: $\boldsymbol{h}_i \in \mathbb{R}^K$, obtained by marginalizing over all pixels as: $\boldsymbol{h}_i = [h_{i,1}, \ldots, h_{i,K}]^\top$, where $h_{i,k} = \frac{s^2}{HW}\sum_{\boldsymbol{p}}\Psi_k(\boldsymbol{M}_i(\boldsymbol{p}))$. By construction, $\boldsymbol{h}_i$ satisfies $\sum_{k=1}^K h_{i,k} \approx 1$. We then define

the structural discrepancy metric as the symmetric Kullback–Leibler (KL) divergence between two given samples i and j, which is formulated as follows:

$$\mathcal{D}_{\text{kl}}(\boldsymbol{h}_i||\boldsymbol{h}_j) = \frac{1}{2}(\sum_{k=1}^{K} h_{i,k} \log\left(h_{i,k}/h_{j,k}\right) + \sum_{k=1}^{K} h_{j,k} \log\left(h_{j,k}/h_{i,k}\right)). \tag{6}$$

Learning to discriminate scene structures. To explicitly guide the structural feature projector to learn generalized, transferable structural representations, we introduce a scene-structure discrimination objective via contrastive learning. Specifically, given a sample i within a training batch, the structural descriptor $\boldsymbol{z}_i^S$ is encouraged to be similar to descriptors of samples with similar scene structures, while being dissimilar to those of structurally different samples. This objective naturally leads to a contrastive learning formulation in the structural representation space. The positive sample set $\mathcal{P}_i$ for anchor i is determined based on the previously defined structural discrepancy metric. We transform the discrepancy into a similarity measure and then threshold it with a hyperparameter γ:

$$\mathcal{P}_i = \{j \mid \exp\left(-\mathcal{D}_{\text{kl}}(\boldsymbol{h}_i|\boldsymbol{h}_j)\right) > \gamma\}, \tag{7}$$

where we set γ to 0.95 based on empirical tuning. Following common practice in contrastive learning [46, 30], each sample is also treated as a positive to itself. Based on the resulting positive set, we adopt a multi-positive InfoNCE objective to optimize the structural representations, formulated as:

$$\mathcal{L}_{\textbf{scl}} = -\frac{1}{B}\sum_{i=1}^{B} \log(\sum_{j\in\mathcal{P}_i} \exp(\boldsymbol{z}_i^{S\top} \boldsymbol{z}_j^S/\tau_c)/\sum_{k=1}^{B} \exp(\boldsymbol{z}_i^{S\top} \boldsymbol{z}_k^S/\tau_c)), \tag{8}$$

where τ_c is the contrastive temperature controlling the concentration of the embedding space, and is set to 0.1 in our implementation. We provide several quantitative examples in the *Supplementary Materials* to demonstrate that the selected hyperparameters effectively prevent this approach from degenerating into discrimination based on coarse scene categories (*e.g.*, indoor and outdoor).

3.3 Depth Prediction from Learned Representations

The contextual and structural representations learned by MaSS are visualized in Fig. 2(c)(d). The contextual pathway effectively adapts the VFM representations into depth-related context, while the structural pathway captures generalized scene geometry. Their corresponding feature pyramids are then concatenated along the channel dimension and passed through a 1×1 convolution for dimension reduction, and finally fed into the depth decoder.

4 Experiments

4.1 Experimental Setups

Implementation details. The MaSS model utilizes the pre-trained DINOv2 ViT-B/14 architecture [24] as its VFM encoder and employs the depth encoder [21]. The interaction module follows the design of ViT-Adapter [13]. During training, the DINOv2 encoder remains frozen while all other parameters are updated. The MaSS model can be trained on a single RTX 4090 GPU with a batch size of 6, which takes approximately 2.5 days to complete. Data augmentation strategy and optimizer settings adhere to the previous studies [1, 21, 4] without additional tuning. The proposed scene-structure discrimination objective operates on the depth predictions to formulate supervision for representation learning. Consequently, it is infeasible to train the depth model from scratch with this objective, as unreliable early predictions would adversely affect representation learning. Therefore, an additional initialization stage for one epoch is required to obtain reasonable depth estimates before applying this objective. During training, we follow the prior work [3] to handle dynamic scenes and extreme conditions that may compromise training stability [50, 5].

Datasets. Experiments are conducted on a total of 13 public datasets. The KITTI [40], NYUv2 [39], DDAD [51], and nuScenes [52] datasets are used for both training and evaluation, covering outdoor, indoor, dynamic, and adverse weather scenarios. Eight additional datasets are used to evaluate the generalizability of MaSS to unseen data, including three driving datasets, Waymo Open [53], DrivingStereo [54], and Cityscapes [55]; three indoor datasets, iBims-1 [56], SUN-RGBD [57], and 7-Scenes [58]; and two out-of-distribution datasets, DIML [59] and DIODE [60]. Finally, the

Table 1: Quantitative results on the training-evaluation datasets. † indicates training under the mixed-dataset setting. For single-dataset training, results are reported at the image resolutions of 192×640, 384×640, 256×320, and 288×512, respectively, while mixed-dataset training results are evaluated at a unified resolution of 384×640. The best results are marked in bold type.

Method	Type	Align.	KITTI		DDAD		NYUv2		nuScenes	
			Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑
SC-DepthV2 [61] TPAMI'22	SI	Median	0.118	0.866	0.164	0.775	0.138	0.820	–	–
Lite-Mono [21] CVPR'23	SI	Median	0.101	0.897	0.175	0.799	0.198	0.688	0.419	0.720
Dynamo-Depth [17] NeurIPS'23	SI	Median	0.112	0.873	0.150	0.798	–	–	0.179	0.787
Gas-Mono [2] ICCV'23	SI	Median	0.111	0.894	0.178	0.756	0.113	0.871	–	–
SC-DepthV3 [3] TPAMI'24	SI	Median	0.118	0.864	0.142	0.813	0.123	0.848	0.167	0.784
DCPI-Depth [4] TIP'25	SI	Median	0.095	0.902	0.140	0.815	0.151	0.785	0.160	0.793
GeoDepth [63] CVPR'25	SI	Median	0.100	0.897	–	–	0.134	0.833	–	–
Jasmine [9] NeurIPS'25	SI	Median	0.097	0.907	–	–	–	–	–	–
DiMoDE [5] TPAMI'26	SI	Median	0.097	0.899	0.134	0.831	–	–	0.139	0.836
MaSS (Ours)	SI	Median	**0.094**	**0.909**	**0.128**	**0.845**	**0.109**	**0.868**	**0.139**	**0.851**
Lite-Mono† [21] CVPR'23	SI	Median	0.111	0.886	0.202	0.758	0.198	0.672	1.147	0.816
Dynamo-Depth† [17] NeurIPS'23	SI	Median	0.126	0.844	0.165	0.764	0.312	0.479	0.181	0.710
SC-DepthV3† [3] TPAMI'24	SI	Median	0.132	0.832	0.158	0.795	0.156	0.773	0.226	0.685
DCPI-Depth† [4] TIP'25	SI	Median	0.114	0.870	0.162	0.777	0.331	0.472	0.160	0.776
MaSS† (Ours)	SI	Median	**0.092**	**0.913**	**0.125**	**0.862**	**0.104**	**0.890**	**0.134**	**0.852**

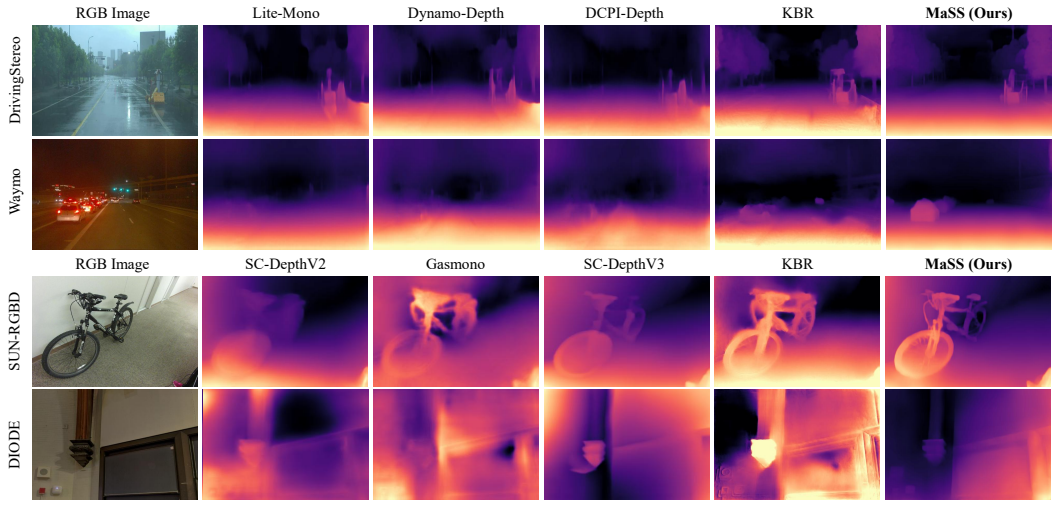


Figure 4: Qualitative results on zero-shot generalization datasets. Compared with representative SSMDE models specialized for either indoor or outdoor scenes and the model scaled with large-scale data and pretrained backbones, MaSS demonstrates consistent superiority across both scenarios.

TartanAir dataset is used to expand the size and diversity of the training data to assess the scalability of MaSS. The characteristics of each dataset and the adopted data splits, following those established in prior works [14, 17, 9, 61, 3, 62], are detailed in the *Supplementary Materials*.

Evaluation protocol. In addition to self-supervised methods that inherently predict scale-invariant (SI) depth, we also include several supervised depth foundation models that produce affine-invariant (SSI) predictions for comparison. For scale-invariant evaluation, predicted depths are aligned to ground truth using median scaling (Median). For affine-invariant evaluation, predictions are aligned to ground-truth depth via least-squares fitting (LSQ). We report two widely used metrics: Absolute Mean Relative (Abs Rel) error and δ_1 accuracy.

4.2 Compare with SoTA Self-supervised Methods

Individual dataset training. Table 1 shows that MaSS outperforms all previous SoTA methods [61, 17, 2, 3, 63, 4, 9] across all four datasets when trained and evaluated individually.

Mixed dataset training. While our models achieve strong results when trained on individual datasets, such settings do not reflect the scalability challenges posed by large-scale and diverse training

Table 2: Quantitative zero-shot generalization results. Models under the single dataset training are evaluated at the image resolutions of 192×640 (outdoor) and 256×320 (indoor). Mixed-dataset models (†) are evaluated for the image resolutions of 384×640. Supervised foundation models are reported at their pretrained resolutions (518×518 for [41] and 768×768 for [27, 28]). Results on the DIML [59] dataset are not reported for method [41] that was pretrained on this dataset. The best results for self-supervised and supervised models are marked in bold type, respectively. SSI* and LSQ* denote the predicted depth and its alignment with the ground-truth in the inverse depth domain.

Method	Type	Align.	Waymo		DrivingStereo		Cityscapes		DIML	
			Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑
DA V2 [41] NeurIPS'24	SSI*	LSQ*	**0.084**	**0.912**	**0.103**	**0.908**	**0.088**	**0.920**	–	–
Marigold [27] CVPR'24	SSI	LSQ	0.128	0.868	0.128	0.858	0.118	0.886	0.123	0.863
Lotus [28] ICLR'25	SSI	LSQ	0.123	0.881	0.123	0.877	0.121	0.878	**0.099**	**0.908**
Dynamo-Depth [17] NeurIPS'23	SI	Median	0.167	0.784	0.164	0.788	0.233	0.564	0.175	0.748
DCPI-Depth [4] TIP'25	SI	Median	0.178	0.771	0.161	0.787	0.241	0.549	0.174	0.733
Jasmine [9] NeurIPS'25	SI	Median	0.124	0.864	0.148	0.826	0.119	0.855	0.199	0.739
MaSS (Ours, KITTI)	SI	Median	0.141	0.836	0.131	0.850	0.132	0.846	0.115	0.875
KBR [6] ICCV'23	SI	Median	0.156	0.806	0.139	0.825	0.111	0.864	0.105	**0.902**
KBR++ [7] Arxiv'24	SI	Median	0.160	0.803	0.126	0.863	0.105	0.882	0.108	0.901
MaSS† (Ours)	SI	Median	**0.123**	**0.867**	**0.100**	**0.905**	**0.103**	**0.896**	**0.104**	0.890

Method	Type	Align.	iBims		SUN-RGBD		7-Scenes		DIODE	
			Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑
DA V2 [41] NeurIPS'24	SSI*	LSQ*	**0.045**	**0.981**	**0.062**	**0.961**	**0.074**	**0.940**	0.265	0.734
Marigold [27] CVPR'24	SSI	LSQ	0.060	0.958	0.081	0.943	0.086	0.921	0.308	**0.773**
Lotus [28] ICLR'25	SSI	LSQ	0.053	0.969	0.080	0.951	0.082	0.929	**0.229**	0.729
Gas-Mono [2] ICCV'23	SI	Median	0.153	0.775	0.161	0.795	0.137	0.825	0.369	0.624
SC-DepthV3 [3] TPAMI'24	SI	Median	0.136	0.812	0.150	0.822	0.140	0.799	0.405	0.637
MaSS (Ours, NYUv2)	SI	Median	0.113	0.869	0.139	0.865	0.144	0.792	0.312	0.711
KBR [6] ICCV'23	SI	Median	0.139	0.828	0.163	0.816	0.170	0.758	0.328	0.665
KBR++ [7] Arxiv'24	SI	Median	0.121	0.871	0.154	0.840	0.179	0.763	0.431	0.627
MaSS† (Ours)	SI	Median	**0.107**	**0.883**	**0.127**	**0.893**	**0.120**	**0.861**	**0.289**	**0.736**

Table 3: Quantitative comparisons on the models' overhead and their required training data.

Method	Marigold [27]	DA V2 [41]	Lite-Mono [21]	DCPI-Depth [4]	SC-DepthV3 [3]	Jasmine [9]	KBR [6]	KBR++ [7]	**MaSS† (Ours)**
Params. (M)	1290.0	97.5	3.1	8.8	14.8	1290.0	92.5	266.6	93.2
FLOPs (G)	339.3	89.5	6.2	13.7	9.8	339.3	88.7	381.7	81.6
Training Data	74K	62M	40K	40K	20K	68K	1.9M	2.2M	150K

data, which constitute the primary focus of this work. To this end, we further conduct comparative experiments in a mixed-dataset training setting. Specifically, each training batch is assembled by randomly sampling data from four datasets, thereby simulating a realistic, large-scale training scenario with substantial data diversity. To maintain a comparable number of training iterations to the individual-dataset setting, the number of training epochs is reduced to one-third. Under this setting, we reproduce several SoTA methods [21, 17, 3, 4] with publicly available training code and compare them with our approach. As shown in Table 1, although the compared SoTA methods are robust across different datasets in individual dataset training, meaning no single dataset causes training to fail, their performance still drops significantly in mixed dataset training. In contrast, our method consistently achieves clear performance improvements.

4.3 Zero-Shot Generalizability Evaluation

Unlike prior works [21, 3, 4] that often treat generalizability evaluation as merely supplementary to in-domain results, we conduct a comprehensive zero-shot generalization study, as shown in Table 2. First, to validate the advantage of learning generalizable structural representations, we compare MaSS with representative outdoor and indoor SSMDE methods [21, 17, 4, 9, 61, 2, 3]. All models are trained on KITTI for outdoor evaluation and on NYUv2 for indoor evaluation. Second, to demonstrate the benefit of scaling training with diverse data, we further include our mixed-dataset trained model (MaSS†) in comparisons across all zero-shot evaluation datasets. Furthermore, we include the prior works that directly train SSMDE on large-scale data [6, 7], as well as several supervised depth

Table 4: Quantitative results for ablation studies and model design analyses.

(ID) Method	Training Sets				Zero-Shot			
	nuScenes		NYUv2		Waymo		7-Scenes	
	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑
MaSS†	0.134	0.852	0.104	0.890	0.123	0.867	0.120	0.855
(a1) w/o Contrastive Learning	0.145	0.850	0.106	0.891	0.163	0.805	0.135	0.820
(a2) w/o Representation Decoupling	0.144	0.840	0.116	0.862	0.148	0.831	0.128	0.823
(a3) w/o Interaction Architecture	0.147	0.837	0.119	0.859	0.190	0.750	0.139	0.805
(b1) DINOv2 [24] (Frozen) + DPT [64]	0.186	0.719	0.129	0.832	0.163	0.757	0.126	0.839
(b2) DINOv2 [24] (FFT) + DPT [64]	0.197	0.709	0.167	0.747	0.226	0.681	0.215	0.600
(b3) ViT-Adapter [13] + DPT [64]	0.166	0.767	0.123	0.856	0.147	0.822	0.163	0.757
(c1) w/ DINOv3 [33]	0.136	0.855	0.100	0.897	0.152	0.826	0.128	0.834

foundation models [41, 27, 28] in the generalization comparison. Finally, the comparisons of the models' overhead and their required training data are included for a more comprehensive discussion on their achieved performance, as provided in Table 3.

First, even on datasets [55, 56] that are fairly similar to the training data, previous approaches do not necessarily achieve satisfactory performance. In contrast, our method maintains strong generalization across all zero-shot datasets, including those with more complex and challenging environments [54, 53, 57], and especially on out-of-distribution scenes [59, 60]. This result highlights the advantage of MaSS in learning transferable structural representations rather than relying on dataset-specific biases. Moreover, MaSS† consistently outperforms MaSS trained on individual datasets, indicating that our method enables mutual enhancement across diverse training data through learning to discriminate scene structure, demonstrating the effectiveness of scalable self-supervised depth learning. Furthermore, MaSS† outperforms the KBR series [6, 7] while requiring less overhead and significantly less training data, highlighting that our method delivers substantial gains through improved learning scalability rather than merely increasing data scale. Finally, MaSS† achieves competitive performance compared to supervised depth foundation models on several outdoor and out-of-distribution datasets. It is worth noting that MaSS still underperforms supervised models on standard indoor datasets. This is likely because these datasets have relatively simple and homogeneous depth distributions, making them particularly favorable for supervised models trained on large-scale indoor ground-truth data (*e.g.*, Hypersim [65] and NYUv2). This analysis is supported by results on DIODE and DIML, where the performance gap becomes much smaller on indoor scenes with broader depth ranges and on outdoor scenes that exhibit indoor-style depth distributions.

Notably, part of the performance gap compared to DA v2 [41], as well as the inconsistencies with the originally reported results of the KBR series[6, 7], stem from differences in alignment between Median and SSI*. Additional results and analysis of this issue are provided in the *Supplementary Material*.

Qualitative results are presented in Figs. 4. Compared with previous methods, MaSS produces sharper object boundaries and finer-grained details (row 3). It also remains robust in complex environments and preserves more accurate scene geometry in these unseen scenarios (rows 1, 2, 4), benefiting from the explicitly learned structural representation.

4.4 Ablation Study

Efficacy of each innovative component. To verify that the three key components of MaSS for learning generalizable structural representations are all indispensable, we conduct ablation studies in Table 4(a). Removing (a1) the contrastive learning objective eliminates explicit supervision for structural representation learning, leading to a significant performance drop, particularly on zero-shot datasets. Removing (a2) the specialized structural pathway for representation decoupling prevents the model from preserving contextual information while learning structure, resulting in consistent degradation across all datasets. Removing (a3) the interaction architecture forces the model to learn structural representations directly from DINOv2 features, which lack sufficient geometric cues, and thus causes a substantial performance decline.

Superiority over alternative designs. MaSS builds upon well-established components such as

Table 5: Quantitative results for the training scalability analyses.

Method	KITTI		DDAD		NYUv2		nuScenes	
	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑
MaSS† (K+D+NY+NU, 150K)	**0.092**	**0.913**	0.125	**0.862**	0.104	0.890	**0.134**	**0.852**
MaSS†* (K+D+NY+NU+T, 200K)	0.093	0.912	**0.122**	0.857	**0.097**	**0.905**	0.136	0.850

Method	Waymo		DrivingStereo		Cityscapes		DIML	
	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑
MaSS† (K+D+NY+NU, 150K)	0.123	0.867	0.100	0.905	**0.103**	**0.896**	0.104	0.890
MaSS†* (K+D+NY+NU+T, 200K)	**0.119**	**0.878**	**0.097**	**0.908**	0.106	0.892	**0.087**	**0.929**

Method	iBims		SUN-RGBD		7-Scenes		DIODE	
	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑
MaSS† (K+D+NY+NU, 150K)	0.107	0.883	**0.127**	**0.893**	**0.120**	**0.861**	0.289	0.736
MaSS†* (K+D+NY+NU+T, 200K)	**0.105**	**0.886**	0.129	0.891	0.126	0.840	**0.289**	**0.746**

DINOv2 [24] and ViT-Adapter [13], making it important to verify that the superior performance stems from our design rather than from simply adopting these components. To this end, we compare MaSS with several plausible alternative configurations in Table 4(b): (b1) directly fine-tuning a DPT [64] head with a frozen pretrained DINOv2 backbone, (b2) fully fine-tuning the entire DINOv2–DPT architecture, and (b3) incorporating ViT-Adapter while fine-tuning both the adapter and the DPT head. Results show that MaSS outperforms all these alternative designs by a large margin. This confirms our analysis that effective scaling of self-supervised depth learning requires explicitly modeling scene structure, rather than relying solely on stronger pretrained backbones.

Effect of VFM backbone. As demonstrated in Table 4(c), replacing the VFM backbone with a stronger self-supervised pretrained model (DINOv3 [33]) leads to only marginal performance changes. This result corroborates our earlier analysis that pretrained VFMs in our framework mainly provide generalizable semantic priors and are not directly responsible for depth decoding. This naturally raises a follow-up question: since pretrained depth or geometry foundation models encode rich geometric priors [66], can MaSS benefit from them, or even further enhance such models? We investigate this question in the *Supplementary Material*.

4.5 Scalability Analyses

To further evaluate the scalability of MaSS as training data grow in size and diversity, 50K additional samples from five representative scenes in the TartanAir dataset: carwelding, hospital, neighborhood, office, and seasonforest. This expands the training set from 150K to 200K samples and introduces a broader range of scene types. As shown in Table 5, the average performance changes in Abs Rel and δ_1 are $+1.5\%$ and $+0.2\%$ on training domains, -0.3% and -0.2% on unseen domains similar to the training domains, and $+4.3\%$ and $+3.0\%$ on more challenging out-of-distribution domains (DIML and DIODE datasets). These results suggest that MaSS benefits from larger and more diverse training data, with performance on the original training domains remaining largely stable and generalization improving significantly on the most challenging unseen domains. Notably, these results provide only preliminary evidence of the scalability of MaSS, as our evaluation remains limited to a training scale far smaller than those used by recent supervised depth foundation models, such as the Depth Anything series [26, 41, 67] and MoGe [68]. How SSMDE approaches would perform when scaled to such data regimes remains largely unexplored and is left for future research.

5 Conclusion

This study revisited self-supervised depth estimation from the perspective of large-scale, mix-domain training and prediction, recognizing that the model's tendency to capture dataset-specific structural bias is the key barrier. MaSS addressed this by coupling a depth encoder with a pretrained VFM and explicitly decoupling the structural representations from the contextual ones to learn scene-generalized structural representations via contrastive learning. MaSS achieved SoTA accuracy and remarkable out-of-distribution generalization, demonstrating improved scalability with larger and more diverse data.

Acknowledgments and Disclosure of Funding

This research was supported in part by the National Key R&D Program of China under Grant 2025YFE0202900, and by the National Natural Science Foundation of China under Grant 62473288 and Grant 62233013, the Explorers Program of Shanghai (Basic Research Funding) under Grant 26TS1404200, the State Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University (No. HMHAI-202406), the Fundamental Research Funds for the Central Universities, NIO University Programme (NIO UP), and the Xiaomi Young Talents Program.

The authors declare no competing financial interests beyond the funding disclosed above.

References

[1] Clément Godard et al. Digging into self-supervised monocular depth estimation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 3828–3838, 2019.

[2] Chaoqiang Zhao et al. GasMono: Geometry-aided self-supervised monocular depth estimation for indoor scenes. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 16209–16220, 2023.

[3] Libo Sun et al. SC-DepthV3: Robust self-supervised monocular depth estimation for dynamic scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(1):497–508, January 2024.

[4] Mengtan Zhang et al. DCPI-Depth: Explicitly infusing dense correspondence prior to unsupervised monocular depth estimation. *IEEE Trans. Image Process.*, 34:4258–4272, 2025.

[5] Mengtan Zhang et al. Discriminately treating motion components evolves joint depth and ego-motion learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, September 8, 2026. doi: 10.1109/TPAMI.2026.3732069.

[6] Jaime Spencer et al. Kick back & relax: Learning to reconstruct the world by watching SlowTV. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 15768–15779, 2023.

[7] Jaime Spencer et al. Kick back & relax++: Scaling beyond ground-truth depth with SlowTV & CribsTV. *arXiv preprint arXiv:2403.01569*, 2024.

[8] Shangshu Yu et al. VFM-Depth: Leveraging vision foundation model for self-supervised monocular depth estimation. *IEEE Trans. Circuits Syst. Video Technol.*, 35(5):5078–5091, May 2025.

[9] Jiyuan Wang et al. Jasmine: Harnessing diffusion prior for self-supervised depth estimation. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 38, pages 88638–88667, 2025.

[10] Zizhan Guo et al. Rebenchmarking unsupervised monocular 3D occupancy prediction. *arXiv preprint arXiv:2602.06488*, 2026.

[11] Chuang-Wei Liu et al. Disparity arbiter: Pseudo label mining with student deficiency awareness. 138, February 2027. Art. no. 104692.

[12] Zhiyu Zhou et al. AERGS-SLAM: Auto-exposure-robust stereo 3D Gaussian splatting SLAM. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 40929–40938, 2026.

[13] Zhe Chen et al. Vision transformer adapter for dense predictions. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.

[14] Tinghui Zhou et al. Unsupervised learning of depth and ego-motion from video. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1851–1858, 2017.

[15] Jia-Wang Bian et al. Unsupervised scale-consistent depth learning from video. *Int. J. Comput. Vis.*, 129(9):2548–2564, September 2021.

[16] Yuliang Zou et al. DF-Net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, volume 11209, pages 38–55, 2018.

[17] Yihong Sun et al. Dynamo-Depth: Fixing unsupervised depth estimation for dynamical scenes. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 36, pages 54987–55005, 2023.

[18] Shuwei Shao et al. MonoDiffusion: Self-supervised monocular depth estimation using diffusion model. *IEEE Trans. Circuits Syst. Video Technol.*, 35(4):3664–3678, April 2025.

[19] Chaoqiang Zhao et al. MonoViT: Self-supervised monocular depth estimation with a vision transformer. In *Proc. Int. Conf. 3D Vis. (3DV)*, pages 668–678, 2022.

[20] Mu He et al. RA-Depth: Resolution adaptive self-supervised monocular depth estimation. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, volume 13687, pages 565–581, 2022.

[21] Ning Zhang et al. Lite-Mono: A lightweight CNN and transformer architecture for self-supervised monocular depth estimation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 18537–18546, 2023.

[22] Zhongkai Zhou et al. Recurrent multiscale feature modulation for geometry consistent depth learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):9551–9566, December 2024.

[23] Ariel Gordon et al. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 8977–8986, 2019.

[24] Maxime Oquab et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024.

[25] Robin Rombach et al. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 10684–10695, 2022.

[26] Lihe Yang et al. Depth Anything: Unleashing the power of large-scale unlabeled data. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 10371–10381, 2024.

[27] Bingxin Ke et al. Repurposing diffusion-based image generators for monocular depth estimation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 9492–9502, 2024.

[28] Jing He et al. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025.

[29] Jiyuan Wang et al. FE2E: From editor to dense geometry estimator. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 19844–19853, 2026.

[30] Wenyao Zhang et al. Hybrid-grained feature aggregation with coarse-to-fine language guidance for self-supervised monocular depth estimation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 6678–6692, 2025.

[31] Vitor Guizilini et al. Semantically-guided representation learning for self-supervised monocular depth. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[32] Gangwei Xu et al. Pixel-perfect depth with semantics-prompted diffusion transformers. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 38, pages 193525–193549, 2025.

[33] Oriane Siméoni et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

[34] Jiyuan Wang et al. WeatherDepth: Curriculum contrastive learning for self-supervised depth estimation under adverse weather conditions. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 4976–4982, 2024.

[35] Jiyuan Wang et al. Digging into contrastive learning for robust depth estimation with diffusion models. In *Proc. ACM Int. Conf. Multimedia (ACM MM)*, pages 4129–4137, 2024.

[36] Yanbo Gao et al. Color and geometric contrastive learning based intra-frame supervision for self-supervised monocular depth estimation. *IEEE Signal Process. Lett.*, 31:2940–2944, 2024.

[37] Jing Cao et al. Learning depth from past selves: Self-evolution contrast for robust depth estimation. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, volume 40, pages 2580–2588, 2026.

[38] Xiao Fu et al. GeoWizard: Unleashing the diffusion priors for 3D geometry estimation from a single image. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, volume 15080, pages 241–258, 2024.

[39] Nathan Silberman et al. Indoor segmentation and support inference from RGBD images. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, volume 7576, pages 746–760, 2012.

[40] Andreas Geiger et al. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3354–3361, 2012.

[41] Lihe Yang et al. Depth Anything V2. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 37, pages 21875–21911, 2024.

[42] Jianyuan Wang et al. VGGT: Visual geometry grounded transformer. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 5294–5306, 2025.

[43] Weilong Yan et al. ER-LoRA: Effective-rank guided adaptation for weather-generalized depth estimation. *arXiv preprint arXiv:2509.00665*, 2025.

[44] Mengde Xu et al. Side adapter network for open-vocabulary semantic segmentation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2945–2954, 2023.

[45] Sicen Guo et al. Fully exploiting vision foundation model's profound prior knowledge for generalizable RGB-Depth driving scene parsing. *IEEE Trans. Intell. Veh.*, 11(3):381–390, March 2026.

[46] Alec Radford et al. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Mach. Learn. (ICML)*, volume 139, pages 8748–8763, 2021.

[47] Yuhang Zhang et al. Oracle-guided masked contrastive reinforcement learning for visuomotor policies. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(11):14087–14104, November 2026.

[48] Shariq Farooq Bhat et al. LocalBins: Improving depth estimation by learning local distributions. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, volume 13661, pages 480–496, 2022.

[49] Shuwei Shao et al. IEBins: Iterative elastic bins for monocular depth estimation. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 36, pages 53025–53037, 2023.

[50] Weilong Yan et al. Synthetic-to-real self-supervised robust depth estimation via learning with motion and structure priors. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 21880–21890, 2025.

[51] Vitor Guizilini et al. 3D packing for self-supervised monocular depth estimation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2485–2494, 2020.

[52] Holger Caesar et al. nuScenes: A multimodal dataset for autonomous driving. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 11621–11631, 2020.

[53] Jieru Mei et al. Waymo open dataset: Panoramic video panoptic segmentation. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, volume 13689, pages 53–72, 2022.

[54] Guorun Yang et al. DrivingStereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 899–908, 2019.

[55] Marius Cordts et al. The Cityscapes dataset for semantic urban scene understanding. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3213–3223, 2016.

[56] Tobias Koch et al. Comparison of monocular depth estimation methods using geometrically relevant metrics on the IBims-1 dataset. *Comput. Vis. Image Underst.*, 191, February 2020. Art. no. 102877.

[57] Shuran Song et al. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 567–576, 2015.

[58] Jamie Shotton et al. Scene coordinate regression forests for camera relocalization in RGB-D images. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2930–2937, 2013.

[59] Jaehoon Cho et al. DIML/CVL RGB-D dataset: 2M RGB-D images of natural indoor and outdoor scenes. *arXiv preprint arXiv:2110.11590*, 2021.

[60] Igor Vasiljevic et al. DIODE: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019.

[61] Jia-Wang Bian et al. Auto-rectify network for unsupervised indoor depth estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12):9802–9813, December 2022.

[62] Mengtan Zhang et al. The Midas touch for metric depth: Interpretable metricization for generalizable depth perception. *arXiv preprint arXiv:2610.16749*, 2026.

[63] Haifeng Wu et al. GeoDepth: From point-to-depth to plane-to-depth modeling for self-supervised monocular depth estimation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 11525–11535, 2025.

[64] René Ranftl et al. Vision transformers for dense prediction. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 12179–12188, 2021.

[65] Mike Roberts et al. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 10912–10922, 2021.

[66] Jiyuan Wang et al. Edit in 2D, verify in 3D: Reinforcement learning for multi-view consistent scene editing. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, volume 17011, pages 260–278, 2026.

[67] Haotong Lin et al. Depth Anything 3: Recovering the visual space from any views. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2026.

[68] Ruicheng Wang et al. MoGe: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 5261–5271, 2025.

[69] David Eigen et al. Depth map prediction from a single image using a multi-scale deep network. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 27, pages 2366–2374, 2014.

[70] Jamie Watson et al. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1164–1174, 2021.

[71] Yi Feng et al. PanoBench: A panoptic occupancy prediction benchmark for autonomous driving. *arXiv preprint arXiv:2610.23537*, 2026.

[72] Boyuan Sun et al. Depth Anything at any condition. *arXiv preprint arXiv:2507.01634*, 2025.

[73] Jiyuan Wang et al. CaC: Advancing video reward models via hierarchical spatiotemporal concentrating. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2026.

[74] Yuhang Zhang et al. Skytopia: Monocular drone navigation with action-conditioned latent world models. *arXiv preprint arXiv:2609.26007*, 2026.

[75] Zhiyu Zhou et al. Visual localization using 3D Gaussian splatting representation for mobile robots with geometric feature correspondences synthesis. 22:20435–20447, 2025.

Outline

In this Supplementary Material, we provide the following:

- **Appendix A: Additional Technical Details.** Quantitative analyses of the proposed scene-structure discrimination objective, additional training details, and detailed descriptions of the datasets used in this work.
- **Appendix B: Supplementary Experimental Results.** Supplementary quantitative results under the different alignment protocols, and additional experiments evaluating model generalizability and geometric robustness on in-the-wild, out-of-distribution data, compared with SoTA self-supervised depth estimation methods and geometry foundation models.
- **Appendix C: Discussion and Future Work.** A discussion of the limitations of our method and potential directions for extending it to advance monocular depth estimation and related vision tasks.

A Additional Technical Details



Figure S1: Quantitative analyses of the proposed scene-structure discrimination objective for structural representation learning.

A.1 Details of Learning to Discriminate Scene Structures

In the main paper, we define a structural discrepancy metric to determine positive and negative samples within a training batch, thereby formulating a scene-structure discrimination objective for learning structural representations. To illustrate how the metric functions and how the objective guides representation learning, we present several quantitative examples in Fig. S1. The key observations from Fig. S1 can be summarized as follows:
(1) Our method does not trivially treat each sample as the sole positive instance, nor does it degenerate into discriminating samples based on coarse scene categories (*e.g.*, indoor vs. outdoor). As shown in Fig. S1(a), Sample ① is an outdoor scene but is not considered positive relative to other outdoor samples; this is because its structural layout—a narrow passage flanked by cars and buildings—differs from more open outdoor environments. Fig. S1(b) provides an inverse case, where a wide-open outdoor scene (Sample ①) is correctly distinguished from narrow outdoor passages. These observations are further supplemented by Sample ⑥ in Fig. S1(c). Moreover, our method captures shared structural properties across scene categories. For instance, in Fig. S1(d), the outdoor scene in Sample ⑥ forms positive pairs with multiple indoor samples due to their similar structure. Similarly, in Fig. S1(e), the indoor library aisle (Sample ①) is identified as positive relative to outdoor samples ③, ⑤, and ⑥, which share the "narrow corridor" structure. Meanwhile, it remains negative to the room scene of Sample ②. These results demonstrate that our model prioritizes underlying structural similarity over nominal scene labels.
(2) The hyperparameters of the proposed structural discrepancy metric, together with the masking strategy, effectively identify meaningful contrastive relationships. As visualized in the Discrepancy Metric and Pos./Neg. Mask columns, the resulting similarity maps align well with the expected structural relationships described in the above analysis. Meanwhile, the threshold used in our implementation effectively determines the positive and negative sample pairs accordingly.
(3) Finally, within the MaSS framework, the specialized pathway effectively learns structural representations that are discriminative with respect to different scene structures. This is evidenced by the strong consistency between the feature similarities of the learned structural representations and the in-batch structural discrepancy relationships computed from the predicted depth maps.

A.2 Additional Training Details

A.3 Details of Training-Evaluation Datasets

KITTI [40]. Representing standard outdoor driving scenarios, we adopt the widely-used Eigen split [69], comprising 39,180 monocular image triplets for training, 4,424 for validation, and 697 for testing.
DDAD [51]. To capture dynamic urban scenes with dense traffic, we utilize the DDAD dataset, which is split into a training set of 12,650 images and a test set of 3,950 images.
NYUv2 [39]. For indoor environments, the NYUv2 dataset is split into 654 testing images for depth evaluation, while the remaining 335 sequences are utilized for training and validation.
nuScenes [52]. Serving as our source for adverse weather and varied lighting conditions, we use 79,760 image triplets from the nuScenes dataset for training, and evaluate the model's performance on 6,019 front-camera images.

A.4 Details of Zero-shot Evaluation Datasets

Waymo Open [53]. Serving as a large-scale outdoor driving benchmark containing diverse urban and suburban scenes, we evaluate models on a test split consisting of 2,216 images, following prior works [17, 4].
DrivingStereo [54]. To evaluate generalization across diverse weather (e.g., rainy and foggy) and lighting conditions in driving scenarios, we use the official test split containing 2,000 images (500 images per weather condition).
Cityscapes [55]. As a benchmark of high-resolution urban street scenes with dense traffic participants, we evaluate on the standard validation set containing 1,525 images, using ground-truth depth from [70].
iBims-1 [56]. Specifically designed to test indoor depth estimation on complex structures (e.g., transparent objects and fine details), we utilize the official test set of 100 images.
SUN-RGBD [57]. A large-scale indoor dataset with high scene diversity. We follow the evaluation

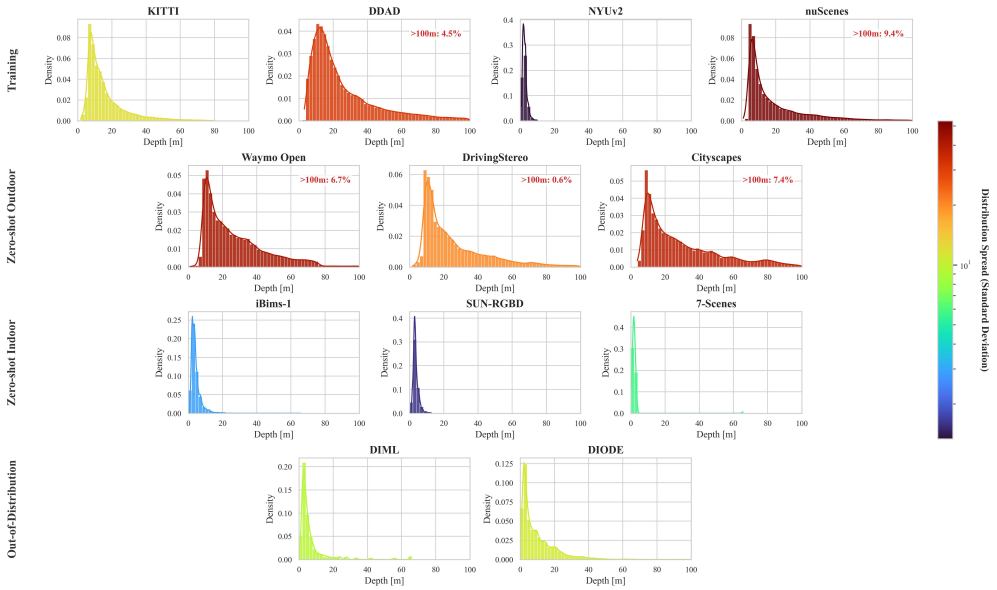


Figure S2: Quantitative visualization of the scene structure (depth) distributions across all training and evaluation datasets.

protocol of [49] using a test split of 5,050 images.

7-Scenes [58]. Collected from a handheld RGB-D camera in various indoor environments, we evaluate on a test split comprising 1,700 images across seven different scenes, following the prior work [61].

DIML [59]. To evaluate robustness on purely out-of-distribution data, we use the DIML test split, which contains 500 outdoor images captured in close-range environments whose depth characteristics differ significantly from typical driving datasets.

DIODE [60]. Featuring diverse high-resolution scenes with highly accurate sensor depth measurements, we evaluate on the official test set of 771 images (325 indoor and 446 outdoor). Compared with commonly used indoor benchmarks, the indoor subset spans a larger depth range, while the outdoor subset also mainly consists of closer-range scenes rather than typical driving environments.

A.5 Data Distribution and Model Generalizability Analyses

We provide a quantitative analysis of the data distributions across all datasets used in this work to better illustrate how our approach improves the scalability of self-supervised depth learning and enhances model generalizability. As shown in Fig. S2, the training datasets (KITTI [40], DDAD [51], and NYUv2 [39]) exhibit distinct depth distributions. These discrepancies arise from differences between indoor and outdoor environments, as well as variations in field of view (FoV) and sensor configurations during data acquisition [71]. In addition, Fig. S3 presents the RGB distributions of the datasets. Since monocular depth estimation maps RGB inputs to depth outputs, diversity in both domains is crucial for robust generalization. Notably, while nuScenes shares a relatively similar depth distribution with KITTI, its RGB distribution differs significantly.

These observations help interpret the zero-shot evaluation results. For the zero-shot outdoor datasets (Waymo Open [53], DrivingStereo [54], and Cityscapes [55]), their depth distributions are similar to those of the driving datasets used for training, whereas their RGB distributions differ noticeably. This suggests that the limited generalization ability of previous methods is not primarily caused by differences in depth ranges. Instead, the performance degradation is largely attributed to appearance shifts, revealing the limited robustness of conventional encoder architectures with restricted pretraining. In contrast, our method learns robust and transferable contextual and structural representations from diverse appearance distributions, substantially improving generalization in depth estimation. For the zero-shot indoor datasets (iBims-1 [56], SUN-RGBD [57], and 7-Scenes [58]), both the depth and RGB distributions closely align with the NYUv2 training set. This relatively homogeneous distribution favors supervised depth foundation models, as similar scene statistics are already well

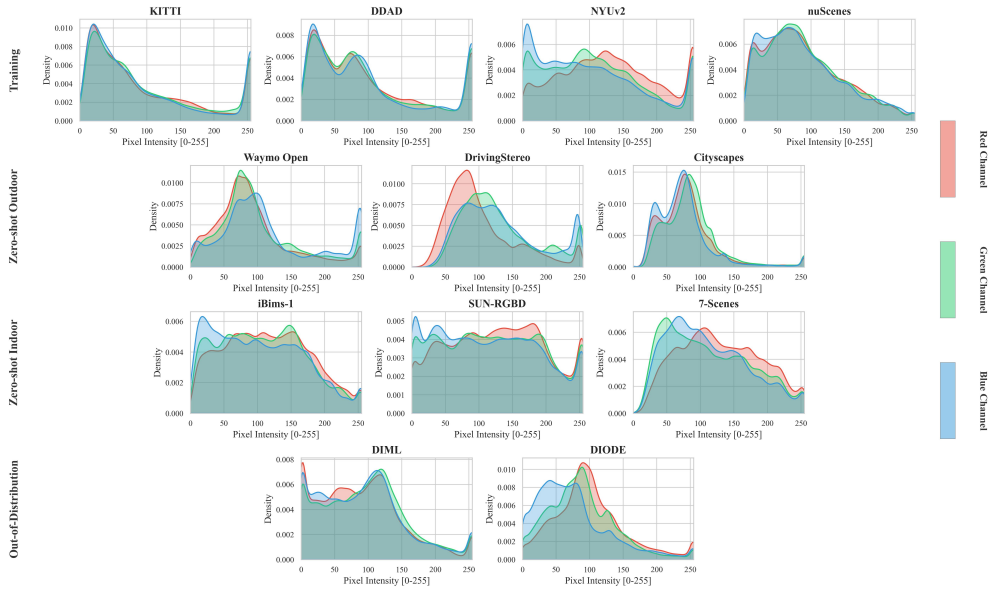


Figure S3: Quantitative visualization of the scene appearance (RGB) distributions across all training and evaluation datasets.

represented in their large-scale training data. Consequently, these models have already learned similar RGB-to-depth mappings under ground-truth supervision, leading to a larger performance gap between supervised and self-supervised approaches on these datasets. Finally, DIML [59] and DIODE [60] represent clear out-of-distribution (OOD) scenarios. As shown in Figs. S2 and S3, they deviate substantially from the training data in both the input (RGB) and output (depth) domains, making them challenging benchmarks for evaluating the OOD generalization capability of depth estimation methods. Notably, on these datasets, our method achieves performance comparable to that of fully supervised depth foundation models.

B Supplementary Experimental Results

B.1 Additional Quantitative Results using Different Alignment Protocols

The results of the KBR series [6, 7] in Table 2 are inconsistent with their originally reported values, as observed on the shared DIODE dataset. This discrepancy arises because they report results using LSQ alignment in the inverse depth domain (LSQ*), whereas we report results using median alignment in the depth domain. However, inherent performance gaps exist when using different alignment strategies, as discussed in prior work [9]. In particular, our study further reveals that the choice of alignment domain (depth or inverse depth) can significantly influence performance. To better understand this effect, we analyze the Least-Squares (LSQ) alignment objective in both domains. Let $\hat{d}_i$ and d_i denote the predicted and reference depth values at pixel i, respectively. In the depth domain, LSQ estimates a scale parameter s and a shift parameter t by minimizing

$$(s^*, t^*) = \arg\min_{s,t} \sum_i \left((s \cdot \hat{d}_i + t) - d_i \right)^2, \tag{S1}$$

which assigns equal importance to absolute depth errors across all pixels. Since depth uncertainty generally increases with distance, distant regions tend to exhibit larger absolute errors and therefore dominate the optimization process, potentially biasing the global alignment. In contrast, LSQ* performs alignment in the inverse depth domain. Let $\hat{v}_i = 1/\hat{d}_i$ and $v_i = 1/d_i$ denote the predicted and reference inverse depth values, respectively. The corresponding objective is formulated as

$$(s'^*, t'^*) = \arg\min_{s',t'} \sum_i \left((s' \cdot \hat{v}_i + t') - v_i \right)^2. \tag{S2}$$

Table S1: Quantitative zero-shot generalization results. Mixed-dataset models (†) are evaluated for the image resolutions of 384×640. Supervised foundation models are reported at their pretrained resolutions (518×518 for [41, 72] and 768×768 for [27, 28]). Results on the DIML [59] dataset are not reported for methods [41, 72] that were pretrained on this dataset. The best results for self-supervised and supervised models are marked in bold type, respectively.

Method	Types	Align.	Waymo		DrivingStereo		Cityscapes		DIML	
			Abs Rel	δ_1	Abs Rel	δ_1	Abs Rel	δ_1	Abs Rel	δ_1
DA V2 [41]	SSI*	LSQ*	0.084	0.912	0.103	0.908	0.088	0.920	–	–
KBR [6]	SI	Median	0.156	0.806	0.139	0.825	0.111	0.864	0.105	0.902
KBR [6]	SI	LSQ*	0.116	0.850	0.115	0.865	0.098	0.896	0.079	0.902
MaSS† (Ours)	SI	Median	0.123	0.867	0.100	0.905	0.103	0.896	0.104	0.890
MaSS† (Ours)	SI	LSQ*	0.088	0.908	0.088	0.916	0.087	0.917	0.089	0.922

Method	Types	Align.	iBims		SUN-RGBD		7-Scenes		DIODE	
			Abs Rel	δ_1	Abs Rel	δ_1	Abs Rel	δ_1	Abs Rel	δ_1
DA V2 [41]	SSI*	LSQ*	0.045	0.981	0.062	0.961	0.074	0.940	0.265	0.734
KBR [6]	SI	Median	0.139	0.828	0.163	0.816	0.170	0.758	0.328	0.665
KBR [6]	SI	LSQ*	0.104	0.890	0.121	0.867	0.117	0.865	0.255	0.679
MaSS† (Ours)	SI	Median	0.107	0.883	0.127	0.893	0.120	0.861	0.289	0.736
MaSS† (Ours)	SI	LSQ*	0.083	0.927	0.100	0.914	0.099	0.903	0.230	0.721

The inverse depth error can be further expressed as

$$(\hat{v}_i - v_i)^2 = \frac{(d_i - \hat{d}_i)^2}{(\hat{d}_i \cdot d_i)^2}, \tag{S3}$$

which shows that LSQ* effectively downweights distant regions through the denominator term. As a result, the influence of high-variance distant pixels is suppressed, preventing the estimated fitting parameters from being dominated by unreliable regions.

As shown in Table S1, we additionally report results for both KBR [6] and our method using LSQ* alignment. Results show that this alignment protocol substantially reduces the error metrics. In particular, the DIODE results of KBR under this setting are consistent with those reported in their original paper. Under the same alignment protocol, our method still significantly outperforms prior methods. In most SSMDE studies, median alignment is a widely adopted and well-justified evaluation protocol. This is because SSMDE inherently learns scale-invariant depth that capture scene geometry without requiring additional shift fitting, whereas LSQ* alignment changes the evaluation objective and no longer faithfully reflects this capability. Therefore, in the main text, we report the results of all self-supervised methods using median alignment.

In addition, the performance gap with DA v2 [41] is substantially reduced when also using LSQ*. Specifically, our method outperforms it on several outdoor and out-of-distribution datasets, while only underperforming on relatively easier indoor scenes. Nevertheless, direct comparisons between models with different invariance properties, such as DA v2 and our method, may not be fully well-defined, even when evaluated under the same LSQ* alignment protocol. This is because DA v2 is explicitly trained to predict scale- and shift-invariant inverse depth, making it inherently compatible with this alignment formulation during both training and inference. As discussed in their original paper, this design facilitates cross-domain depth learning. In contrast, self-supervised methods do not possess such an inherent advantage.

B.2 Adapting Depth/Geometry Foundation Models

As demonstrated in Table S2, when replacing the original VFM encoder with pretrained depth/geometry foundation models, including DepthAnythingV2 [41] and VGGT [42], MaSS achieves substantial performance gains and also outperforms the original depth prediction heads of these models. This highlights that learning structural representations provides non-trivial benefits even on top of strong depth/geometry-supervised pretraining. Moreover, since self-supervised methods require only unlabeled monocular videos and inherently predict scale-invariant depth, MaSS can

Table S2: Quantitative results on adapting depth/geometry foundation models.

Method	Types	Align.	KITTI		nuScenes		DrivingStereo		Cityscapes	
			Abs Rel	δ_1	Abs Rel	δ_1	Abs Rel	δ_1	Abs Rel	δ_1
DA V2 [41]	SSI	LSQ	0.095	0.906	0.171	0.780	0.103	0.908	0.088	0.920
DA V2 + Ours	SI	Median	0.085	0.923	0.125	0.891	0.090	0.924	0.085	0.930
VGGT [42]	SI	Median	0.099	0.885	0.127	0.837	0.094	0.918	0.111	0.862
VGGT + Ours	SI	Median	0.085	0.924	0.123	0.871	0.089	0.932	0.092	0.920

serve as an effective annotation-free adaptation framework for depth foundation models that output affine-invariant depth or disparity, which cannot be directly deployed in many practical scenarios.

B.3 Generalizability Evaluation on In-the-Wild Data.

While the previously discussed OOD datasets demonstrate the strong generalization capability of our model, their diversity and domain shift relative to the training data remain limited. To further evaluate robustness, we collect a set of in-the-wild images from online repositories and mobile photography, covering a wide range of real-world challenges.

As illustrated in Fig. S4, the first three rows (a–c) present unstructured natural scenes. Notably, (a) and (c) contain specular reflections on water surfaces, while (b) includes dense vegetation with high-frequency textures. Rows (d–f) exhibit complex geometries and structures that differ significantly from standard urban or indoor benchmarks, such as the detailed roof tiles in (e) and the repetitive grid structure in (f). Row (g) includes power lines to evaluate the model's ability to estimate thin structures. Rows (h–j) further evaluate model robustness under extreme lighting and motion blur conditions. In particular, row (j) presents a challenging case with strong visual ambiguity at a road junction caused by shadows and lighting effects.

Qualitative comparisons are conducted among SC-DepthV3 [3] (trained on NYUv2), SC-DepthV3†, Jasmine [9] (trained on KITTI), and MaSS†. As shown in Fig. S4, SC-DepthV3 performs unsatisfactorily on complex and unseen structures and tends to interpret scenes with an indoor bias. When trained under the mixed-dataset setting, SC-DepthV3† further deteriorates, suffering from severe structural confusion caused by the diverse training data. Jasmine represents one of the strongest self-supervised generative depth models. Although it produces detailed predictions and exhibits improved generalization due to diffusion priors, extreme RGB domain shifts can cause the diffusion process to become unstable and produce depth predictions with significant errors. In contrast, our method remains robust across these challenging in-the-wild scenes, producing depth predictions with both fine-grained details and consistent structural layouts.

B.4 Geometric Fidelity Evaluation on In-the-Wild Data.

Since ground-truth depth is unavailable for these in-the-wild images, quantitative evaluation is infeasible. Instead, we assess the geometric correctness of the predicted depth through monocular 3D reconstruction.

We compare our method with two representative monocular reconstruction approaches, VGGT [42] and MoGe [68]. To evaluate geometric fidelity, we synthesize novel viewpoints with slight viewpoint changes relative to the original image. As shown in Fig. S5, VGGT produces relatively noisy depth maps on these in-the-wild scenes, leading to noticeable artifacts in the reconstructed geometry. MoGe represents one of the most robust monocular reconstruction foundation models, benefiting from strong geometric supervision during training, and achieves high-quality reconstructions in most cases, with only minor local inconsistencies observed in (b)-(d). Our method, benefiting from the inherent geometric constraints of self-supervised learning as well as effective scaling with diverse training data, achieves reconstruction results comparable to these prior approaches, with significantly less training data. These results highlight the potential of self-supervised methods for building versatile "DepthAnything"-style models from ubiquitous unlabeled data.

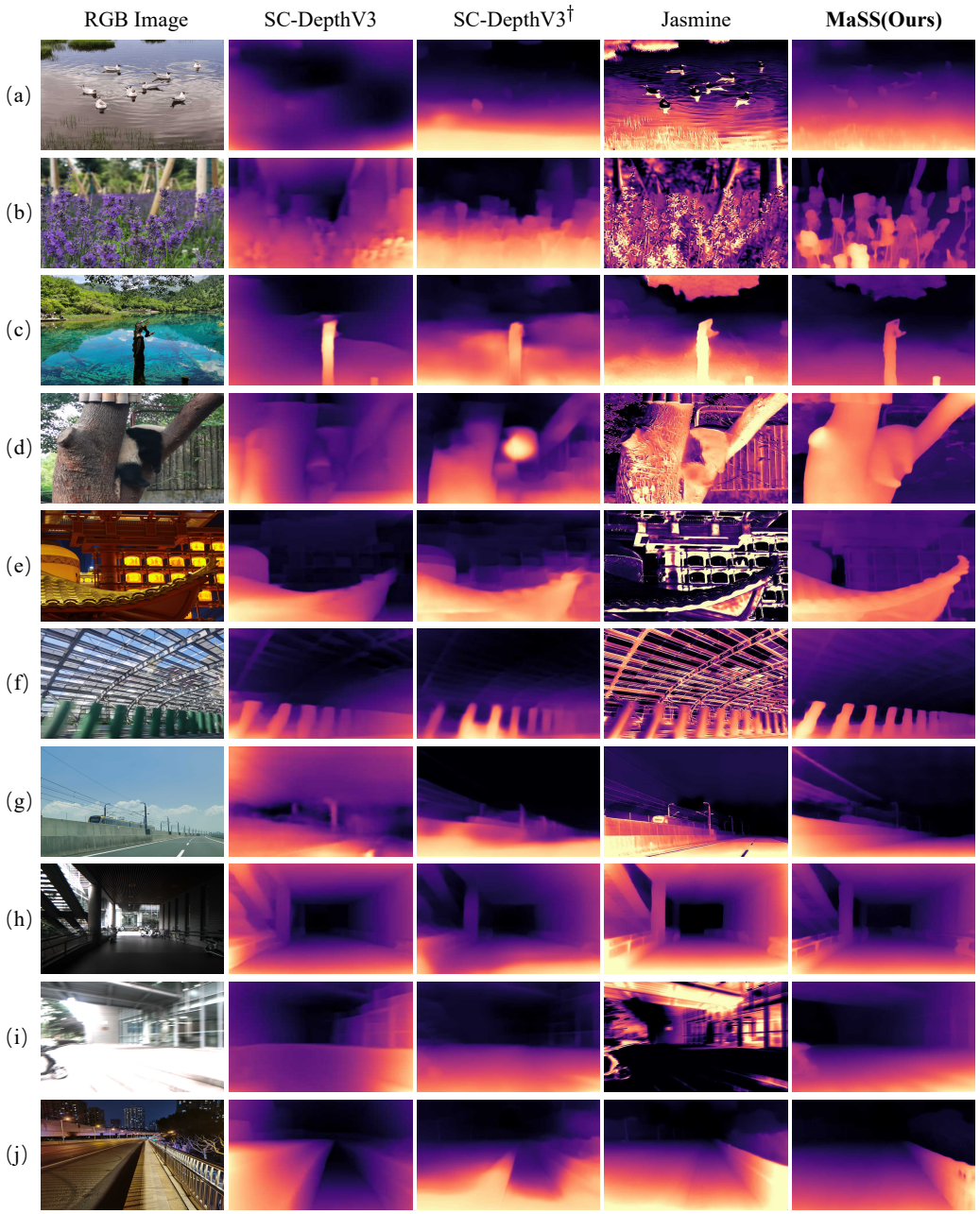


Figure S4: Qualitative monocular depth estimation results on in-the-wild, out-of-distribution data.

C Discussion and Future Work

In addition to increasing the training data scale, as discussed in Sect. 4.5, the scalability of SSMDE also depends on the availability of diverse and suitable training data [73]. Although self-supervision enables learning from imagery collected in the wild, systematically curated datasets covering diverse scenes remain scarce. Existing methods therefore still rely heavily on commonly used datasets originally collected for supervised learning. Constraints imposed by data collection platforms (*e.g.*, vehicles, handheld devices, or UAVs), together with the high cost of sensor calibration and annotation, often limit these datasets to relatively homogeneous scenes, such as typical indoor rooms and common driving environments. This limited diversity may constrain the potential benefits of self-supervision. Such limited diversity may even reduce the potential advantages of self-supervised learning. Systematically curating large and diverse datasets from imagery collected in the wild is

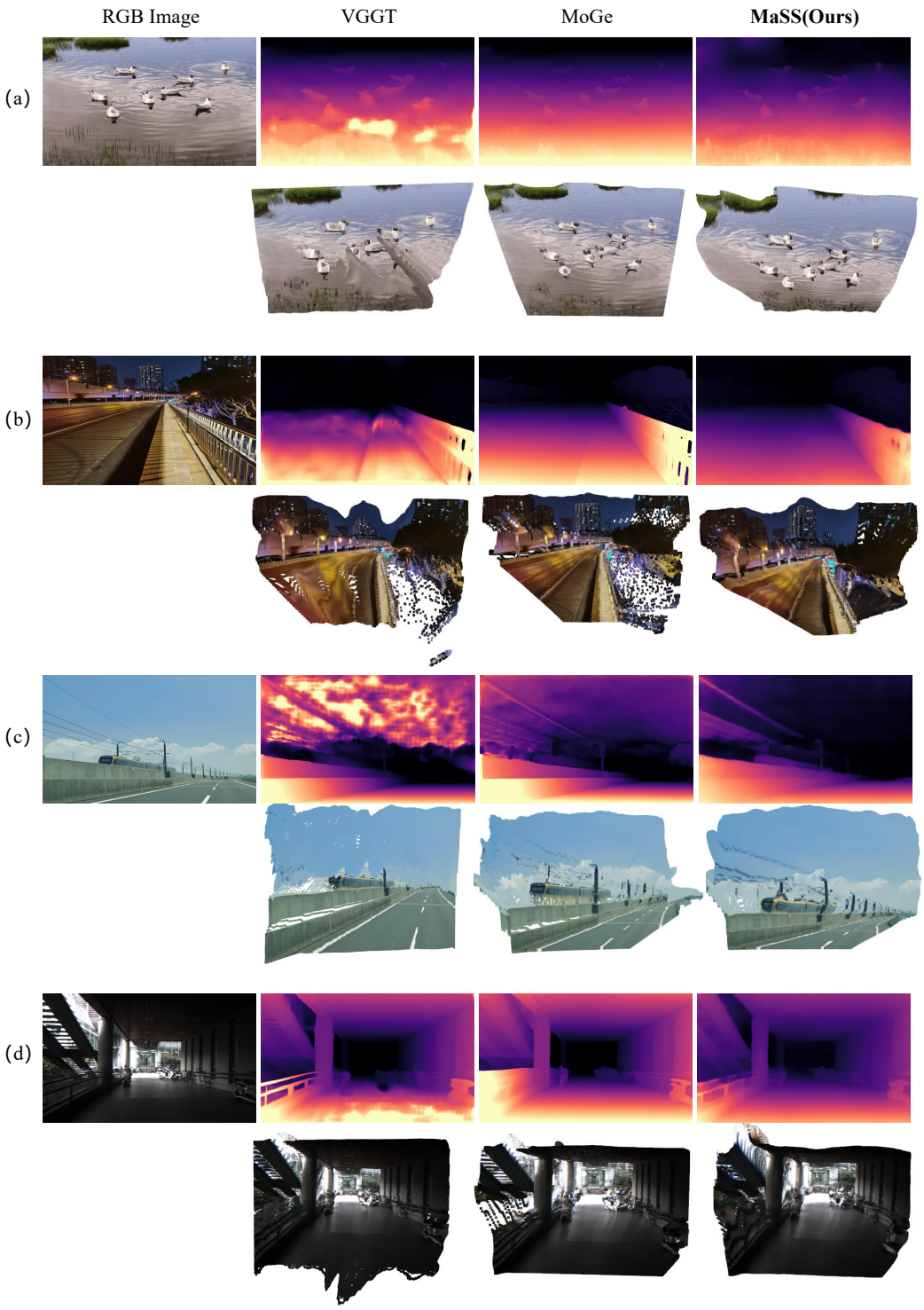


Figure S5: Qualitative monocular reconstruction estimation results on in-the-wild, out-of-distribution data.

therefore a promising direction for advancing SSMDE and extending the benefits of self-supervision to other vision tasks [74, 75].

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly present the core contributions of the paper, which are subsequently supported by experimental results in the main body.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in the "Discussion and Future Work" Section in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not include formal theoretical results, theorems, or mathematical proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides sufficient implementation details to reproduce the main experimental results, including model architecture, training settings, datasets, evaluation metrics, and ablation studies. We also plan to release the code and models to further facilitate reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The datasets used in our experiments are publicly available, ensuring accessibility of the data. Although the code is not provided at submission time, we plan to release the code, models, and detailed instructions to facilitate full reproducibility.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides comprehensive details on the experimental setup, including dataset splits, hyperparameter settings, and optimizer type.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report quantitative metrics for the depth estimation task and provide qualitative visualizations of the depth map, error map, and point clouds to demonstrate the effectiveness of our method. However, the current experiments do not include error bars or formal statistical significance analysis.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the number of GPUs used during training. We provide the time to reproduce the experiments measured on an RTX 4090 GPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm our research fully complies with its principles. No ethical issues arise.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We discuss this aspect in the "Discussion and Future Work" Section.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not involve the release of models or datasets that pose a high risk of misuse. Therefore, no specific safeguards are necessary.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use publicly available datasets and pretrained models, all of which are properly cited in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Although the assets (code and model weights) are not released at submission time, we plan to release them upon acceptance. The released assets will be properly documented, including usage instructions, license information, and known limitations. No new datasets are introduced in this work.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing experiments or research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This paper does not involve research with human subjects or crowdsourcing, and therefore no IRB or equivalent ethical approval is required.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method of this research does not involve any important, original, or non-standard usage of large language models (LLMs). Any use of LLMs was limited to writing assistance and did not impact the scientific contributions of the paper.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.