Scale-Invariant Monocular Depth Estimation via SSI Depth

S. Mahdi H. Miangoleh
Simon Fraser University
Canada

Mahesh Reddy
Simon Fraser University
Canada

Yağız Aksoy
Simon Fraser University
Canada

Figure 1: (top) We propose a framework to generate high resolution scale-invariant (SI) depth from a single image that can be projected to geometrically accurate point clouds of complex scenes. Our generalization ability comes from formulating SI depth estimation with SSI inputs. (bottom) For this purpose, we introduce a novel scale and shift invariant (SSI) depth estimation formulation that excels in generating intricate details. Image credits: @Alka Jha, @Joel Muniz

ABSTRACT

Existing methods for scale-invariant monocular depth estimation (SI MDE) often struggle due to the complexity of the task, and limited and non-diverse datasets, hindering generalizability in real-world scenarios. This is while shift-and-scale-invariant (SSI) depth estimation, simplifying the task and enabling training with abundant stereo datasets achieves high performance. We present a novel approach that leverages SSI inputs to enhance SI depth estimation, streamlining the network's role and facilitating in-the-wild generalization for SI depth estimation while only using a synthetic dataset for training. Emphasizing the generation of high-resolution details, we introduce a novel sparse ordinal loss that substantially improves detail generation in SSI MDE, addressing critical limitations in existing approaches. Through in-the-wild qualitative examples and zero-shot evaluation we substantiate the practical utility of our approach in computational photography applications, showcasing its ability to generate highly detailed SI depth maps and achieve generalization in diverse scenarios.

CCS CONCEPTS

• **Computing methodologies** → **Reconstruction**; *Computational photography*.

KEYWORDS

monocular depth estimation, 3D geometry estimation, mid-level vision

ACM Reference Format:
S. Mahdi H. Miangoleh, Mahesh Reddy, and Yağız Aksoy. 2024. Scale-Invariant Monocular Depth Estimation via SSI Depth. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27-August 1, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3641519.3657523

SIGGRAPH Conference Papers '24, July 27-August 1, 2024, Denver, CO, USA

ACM ISBN 979-8-4007-0525-0/24/07…$15.00
https://doi.org/10.1145/3641519.3657523

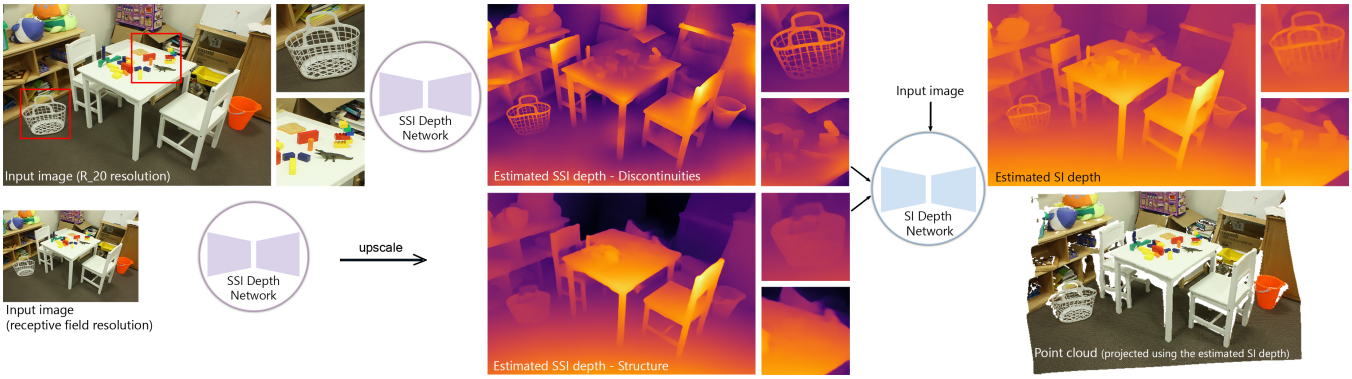

Figure 2: In our framework, we employ a low-resolution SSI depth estimation to capture the rough scene structure, and a high-resolution SSI depth estimation representing sharp depth discontinuities. Feeding this rich structural information to the SI network, we regress the high-resolution scale-invariant monocular depth that can be projected into geometrically accurate point clouds. Image credits: Middlebury dataset [Scharstein et al. 2014]

1 INTRODUCTION

Monocular depth estimation (MDE) is a fundamental mid-level computer vision problem and a critical part of computational photography pipelines such as 3D photography, free view-point rendering, and depth-based editing on individual photographs [Niklaus et al. 2019; Peng et al. 2022; Shih et al. 2020; Wadhwa et al. 2018]. Lacking geometric cues available in multi-view reconstruction formulations, MDE is a challenging high-level problem that requires reasoning about monocular depth cues such as occlusions, relative object size, and converging lines. The challenge is further enhanced for computer graphics applications with the requirement of high-resolution estimations and in-the-wild generalization.

MDE can be defined as the estimation of the physical distance of every pixel to the camera, which is referred to as *metric depth* that requires the focal length of the camera as well as semantic knowledge of the size of objects. The scene geometry, however, can be captured up to a scale with an unknown focal length reflecting the inherent scale invariance in image formation. The estimation of this geometric depth is referred to as *scale-invariant (SI) MDE*. While the metric scale is required for robotics applications such as autonomous driving, computational photography applications only require the geometric SI depth for rendering. In this work, we focus on achieving high-resolution SI MDE in in-the-wild and complex scenes.

Due to the lack of high-resolution, large-scale, and diverse training datasets for SI depth, earlier methods have failed to achieve the boundary accuracy and generalizability demanded by photography applications. To address the generalizability challenge, several works [Ranftl et al. 2020; Yin et al. 2019, 2021b] define the MDE problem in the disparity space coming from stereo pairs with unknown baselines. This stereo MDE is referred to as *scale-and-shift-invariant (SSI) depth*, reflecting the arbitrary shift from the true geometry inherent in stereo pair disparities. With the abundance of stereo training datasets, SSI depth is shown to have better generalization compared to SI MDE. SSI MDE also generates better details at high resolutions [Miangoleh et al. 2021], but is insufficient for computer graphics applications due to the loss of geometric accuracy.

In this work, we propose an SI MDE pipeline, visualized in Figure 2, that makes use of abundant stereo datasets for in-the-wild high-resolution geometric depth estimation. Our pipeline consists of an initial SSI depth estimation, the results of which are fed to a second SI depth estimation network. We first develop a novel sparse loss to improve SSI MDE performance in detail generation and boundary accuracy. We show that our SSI depth estimation outperforms the current state-of-the-art, allowing highly detailed estimation of depth discontinuities even in complex scenes as Figure 1 demonstrates. We use our SSI MDE network to generate an overall scene structure and high-resolution depth discontinuities to be given to our SI network as input.

Given rich structural information in the form of SSI depth, the task of our SI MDE network gets simplified into the enforcement of geometric constraints. This simplified task definition narrows the domain gap between synthetic datasets and in-the-wild images. We show that with the generalizable SSI depth used as input, in-the-wild geometric depth estimation can be achieved using only synthetic SI depth datasets for training. Effectively, our two-step pipeline allows us to harness the advantages of SSI MDE to generate highly detailed geometry from a single image in a wide variety of scenes as shown in Figure 3. We demonstrate the practical use of our methodology through qualitative examples and 3D computational photography applications in the supplementary material.

2 RELATED WORK

Monocular depth estimation (MDE) is a high-level task that requires reasoning about monocular depth cues such as occlusions, perspective, and relative size of objects. Hence, modern MDE approaches are overwhelmingly data-driven to implicitly learn the depth cues [Eigen et al. 2014; Godard et al. 2017; Ramamonjisoa et al. 2020; Wang et al. 2020a; Wong and Soatto 2019; Zheng et al. 2018].

Scale-and-shift-invariant MDE. In-the-wild MDE requires large training datasets for better generalization to in-the-wild images. However, due to the difficulty in capturing metric depth ground-truth at high-resolution at scale, there only exist a few real-world datasets for this task. In order to use datasets with stereo image pairs to

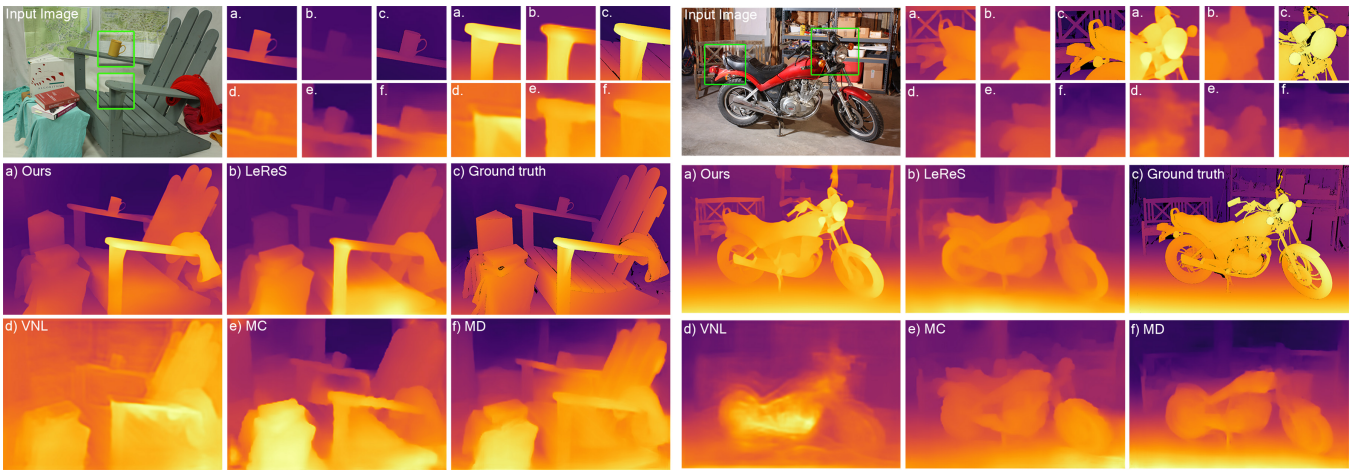

Figure 3: Qualitative comparison of scale-invariant networks on the Middlebury dataset [Scharstein et al. 2014]. Our scale-invariant network exhibits superior performance in capturing intricate objects with higher levels of depth details compared to the state-of-the-art.

extend the available training data, Ranftl et al. [2020] develop a scale-and-shift invariant (SSI) depth formulation and train their network in the disparity space. The relaxed formulation and the extended set of image-ground truth pairs with novel loss functions improve the accuracy of the estimated SSI depth and allow generalization to in-the-wild images, but the unknown shift still needs to be recovered for geometric reconstruction. Miangoleh et al. [2021] propose a boosting framework and demonstrate that high-resolution SSI depth with rich details can be achieved using CNN-based SSI models through inference at two different resolutions. Yin et al. [2021a, 2019] define a virtual plane and define a loss on its surface normal and combine it with an SSI loss in the depth space. Yin et al. [2021b] resolves the SSI ambiguity by normalizing the depth distribution per image with a Gaussian assumption. Ranftl et al. [2021] propose a novel transformer architecture for dense depth estimation that benefits from the higher learning ability of transformer architectures. Yang et al. [2024] exploit unlabeled data and pretrained depth models to generate pseudo ground truth to train a student model that is more effective than the original teacher model. We introduce a novel sparse ordinal loss and demonstrate that combining our sparse loss with the dense SSI loss enhances detail generation for our SSI model.

Ordinal MDE. As Zoran et al. [2015] explore, estimating whether a pixel is closer to the camera than the other without enforcing geometric constraints leads to better performance in estimating depth discontinuities. Several works [Chen et al. 2016, 2019b; Xian et al. 2018, 2020] aim for dense estimation of ordinal depth using sparse ranking loss functions that only enforce the correct ordering of pixels. Xian et al. [2020] shows that they generate MDE with more high-resolution details when compared to MiDaS [Ranftl et al. 2020] which still encodes the geometry. As discussed in Section 3, traditional sparse ranking loss cannot be combined with SSI loss. However, our sparse loss is compatible with SSI loss, thereby enabling the detail-generating capability of the sparse loss for SSI depth estimation.

Scale-invariant MDE. The scale-invariant (SI) depth estimation networks require geometrically consistent depth maps that are a scale away from the true depth. The earlier data-driven methods approached MDE geometrically through scale-invariant loss definition [Eigen and Fergus 2015; Eigen et al. 2014; Li and Snavely 2018]. Also, the geometrically consistent scale-invariant depth can be used to reliably estimate the surface normals as well. This connection has also been exploited to train MDE for scale-invariant estimation using surface normals [Chen et al. 2017]. Due to the complex nature of SI depth estimation and the constrained capacity of neural networks, models attempting to directly estimate SI depth struggle to generate fine details. Additionally, these models often face limitations due to dataset constraints, being typically trained on a single dataset, which reduces their ability to generalize to in-the-wild images.

Yin et al. [2021b] proposes to estimate SI depth by estimating the unknown shift in SSI depth using a network trained on point clouds. This allows benefiting from the SSI depth to achieve better generalization for SI depth estimation. However, as shown in Figure 1, their method fails to generate detailed geometry for complex scenes. This deficiency stems from its reliance on the geometry estimated from a single pass through an SSI network.

To achieve a high level of details and leverage the generalizability of SSI for SI MDE, we utilize an SSI depth network with a CNN backbone and propose feeding SSI depth at both low and high resolutions as input to a dense SI depth estimation model. By incorporating the local details generated from high-resolution SSI depth, in addition to the structure of the low-resolution SSI depth, our method demonstrates the capability to generate highly detailed SI depth. We utilize a publicly available synthetic dataset to train our SI model, demonstrating that, owing to the simplified task of SI depth estimation by feeding SSI depth to it, our model can generalize effectively to diverse scenes despite being trained on a single dataset.

Metric MDE. Metric MDE directly regresses metric depth. To enhance this process, significant efforts have been made to refine network architectures [Chen et al. 2019a; Eigen et al. 2014], incorporate CRFs [Yuan et al. 2022], or reformulate continuous depth regression as a classification task [Bhat et al. 2021, 2022; Fu et al. 2018]. To simplify the task, Jun et al. [2022]; Lee and Kim [2019] decompose metric depth into ordinal features and aim to estimate the full metric depth using these features. Bhat et al. [2023] utilize low-resolution SSI depth to estimate metric depth. Li et al. [2024] employed a brute-force, patch-based approach to estimate depth, succeeding in generating highly detailed depth maps. However, this iterative approach results in slower runtime compared to the state of the art. Ultimately, when in-the-wild images taken with varying focal lengths cause the depth regressed by these methods to differ from the actual metric depth by a scale factor. This discrepancy reduces their effectiveness to that of a scale-invariant (SI) depth estimation model, despite significant network capacity being dedicated to training for metric depth estimation.

3 HIGH-RESOLUTION SSI DEPTH ESTIMATION

Seeking high-resolution scale-invariant monocular depth estimation (SI MDE), our approach unfolds in two key steps. The initial stage involves extracting the overall structure and high-resolution depth discontinuities through the application of the scale-and-shift invariant (SSI) formulation. This extracted information serves as the input to our subsequent scale-invariant depth estimation network.

The concept of SSI monocular depth estimation was originally introduced by Ranftl et al. [2020] as a more flexible alternative to traditional SI MDE. It allows for training SSI networks on stereo datasets with unknown baselines while also utilizing standard SI ground-truth. Miangoleh et al. [2021] further demonstrated the feasibility of achieving high-resolution SSI estimations by combining multi-resolution outputs. Exploiting these dual advantages –generalizability and accurate depth discontinuities– our focus lies in enhancing the high-resolution accuracy of the SSI formulation.

To achieve this goal, we introduce a novel sparse ordinal loss for SSI training. This loss contributes to the improved high-resolution performance of our SSI formulation, a critical component in generating detailed SI depth estimations in the subsequent stages of our pipeline.

3.1 Sparse ordinal loss

SSI depth estimation is characterized by its scale and shift-invariant loss [Ranftl et al. 2020] defined in the disparity space:

$$\mathcal{L}_{ssi} = \frac{1}{N}\sum_{i}^{N}(f(O_i) - D_i^*)^2, \tag{1}$$

where O is the estimated disparity, D^* is the ground-truth, and

$$f(x) = ax + b \quad (a,b) = \arg\min_{a,b}\sum_{i}(f(O_i) - D_i^*)^2, \quad a > 0 \tag{2}$$

is a linear function parameters of which are estimated for each individual estimation during training. This formulation is particularly useful as it allows the use of both geometric ground-truth as well as disparities estimated from stereo pairs with unknown baseline distance for training.

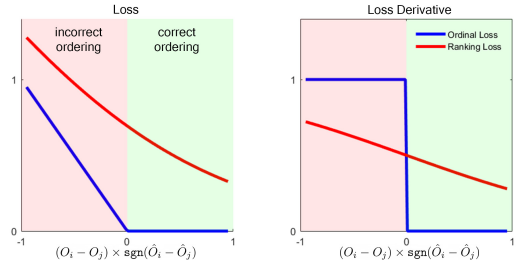

Figure 4: The plot of our ordinal loss and the ranking loss [Chen et al. 2016]. The ranking loss assigns a high penalty for correctly ordered pairs, while we only apply a penalty for incorrectly ordered pairs.

The SSI loss is a global function, instilling coherence in the depth estimation structure. However, the sole use of the SSI loss does not allow the network to generate sharp depth discontinuities when compared to sparse ordinal formulations. To enhance the emphasis on sharp depth discontinuities, we introduce a sparse ordinal loss, working in tandem with the dense SSI loss, to enforce the correct ordering of pixel pairs in the depth space. For a given pixel pair (i, j), we define our ordinal loss as:

$$\mathcal{L}_o(i,j) = \begin{cases} (\Delta O_{ij})^2 & \text{if } |\Delta\hat{O}_{ij}| < \delta \\ \mathrm{ReLU}\left(-\Delta O_{ij} \times \mathrm{sgn}(\Delta\hat{O}_{ij})\right) & \text{otherwise} \end{cases} \tag{3}$$

Here, O and $\hat{O}$ represent the estimated and ground-truth disparity, respectively, and $\Delta O_{ij} = O_i - O_j$. The term $\delta = 0.01$ is a small threshold, defining when two points are considered to be at the same depth. For pixels at different depths, we apply a linear loss only when the estimated ordering of the pair diverges from the ground-truth. Conversely, for pixels at similar ground-truth depths, we apply an L_2 loss, encouraging estimations to be similar.

As discussed in Section 5.3.1, our sparse loss significantly enhances the edge accuracy of SSI estimations. This improvement aligns with the advantages of other sparse losses observed in the realm of relative depth literature, particularly the ranking loss by Chen et al. [2016] and its subsequent use by others [Chen et al. 2019b; Xian et al. 2018, 2020]. A drawback of the sparse ranking loss, as illustrated in Figure 4, is their non-zero contribution even when the pixel ordering is correct. This characteristic leads to a conflict when combined with the SSI loss, rendering their joint use impractical. In contrast, our ordinal loss is carefully defined to circumvent such conflicts and enabling seamless integration with the SSI loss.

Following Chen et al. [2016], we compute our sparse ordinal loss over 2500 randomly sampled pixel pairs over the image, $\mathcal{L}_{so} = \sum_{\forall(i,j)}\mathcal{L}_o(i,j)$. We define our final loss with the SSI and sparse ordinal losses, as well as the multi-scale gradient loss $\mathcal{L}_{ssig}$ [Li and Snavely 2018] as an edge-aware smoothness metric:

$$\mathcal{L}_{ssiNet} = \lambda_{ssi}\mathcal{L}_{ssi} + \lambda_{so}\mathcal{L}_{so} + \lambda_{ssig}\mathcal{L}_{ssig}, \tag{4}$$

where $\lambda_{ssi} = 3$, $\lambda_{so} = 1$, and $\lambda_{ssig} = 0.1$.

3.2 Training details

We follow Ranftl et al. [2020] and adapt the network architecture from Xian et al. [2018] with a ResNeXt101 [Xie et al. 2017] feature extractor with weakly supervised learning weights [Mahajan et al. 2018] as initialization. We use the sigmoid activation to predict the ordinal inverse-depth, i.e. disparity, in $[0, 1]$ and train the network with Adam optimizer for 30000 iterations with a learning rate of 10^{-3}. More details are provided in the supplementary material.

We train our ordinal network on a diverse set of datasets for better generalization. From the Omnidata framework [Eftekhar et al. 2021], we use the Hypersim [Roberts et al. 2021], Replica [Straub et al. 2019] and Replica+GSO [Choi et al. 2016] datasets. We also use the synthetic OpenRooms [Li et al. 2021], TartanAir [Wang et al. 2020b], and FSVG [Krähenbühl 2018] datasets. In addition, we use the real-world stereo datasets HRWSI [Xian et al. 2020] and Holopix50k [Hua et al. 2020] where we compute the disparity maps using RAFT [Teed and Deng 2020] and use the sky segments from Mask2Former [Cheng et al. 2022]. This wide range of real-world datasets ensures that the SSI inputs we generate for our SI network are reliable in a wide range of real-world scenarios.

4 SCALE-INVARIANT DEPTH WITH SSI INPUTS

We redefine the scale-invariant monocular depth estimation (SI MDE) problem by incorporating scale-and-shift invariant (SSI) inputs. The network receives two SSI inputs concatenated with the input image, forming an input of dimensions $h \times w \times 5$, where h and w represent the height and width of the input image, respectively. The first SSI input, denoted as O^L, is computed at the receptive field size of the SSI network, offering an overall depiction of the scene's structure. The second SSI input, O^H, is generated at a higher resolution, capturing intricate depth discontinuities. The resolution of O^H estimation is selected based on the image content using the $\mathcal{R}_{20}$ measurement by Miangoleh et al. [2021] using the local edge density. This formulation streamlines the task for the SI MDE network, focusing on enforcing geometric constraints using the provided overall structure and high-resolution depth information.

Benefiting from a diverse training dataset for our SSI network, the SSI inputs exhibit robust generalization to in-the-wild imagery. By simplifying the task definition for the SI network and leveraging generalizable SSI inputs, we demonstrate the feasibility of achieving in-the-wild high-resolution SI depth estimation via synthetic-only training.

4.1 Scale ambiguity

The inherent scale ambiguity in SI depth formulations necessitates reliance on scale-invariant losses during network training. These losses, enforcing a least squares fit between network estimations and ground truth, face challenges in determining scale during the initial training phases due to inherent inaccuracies in under-trained networks.

In our framework with scale-and-shift invariant (SSI) inputs, the globally consistent low-resolution SSI estimation, denoted as O^L, acts as a stable reference for our SI MDE network. Sequential training, starting with the SSI network, ensures a stable least-squares fit between O^L and ground truth. Utilizing this low-resolution input establishes the ground truth's arbitrary scale with stability during early training, addressing challenges posed by scale ambiguity.

In accordance with the SSI estimation, we formulate SI depth estimation in the inverse depth space. To maintain stability, we fix the ground truth's arbitrary scale throughout training using:

$$c = \arg\min_s \sum_i \left(s\hat{D}_i^* - O_i^L \right)^2, \quad \hat{D} = c\hat{D}^*, \tag{5}$$

Here, $\hat{D}^*$ and $\hat{D}$ represent the original and scale-adjusted ground truth inverse depth. Simultaneously, we align the average scale of the high-resolution SSI input O^H with that of O^L, ensuring consistent scales for input and output variables. Fixing the arbitrary scale provides a foundation for defining dense losses without requiring scale invariance, enhancing stability and effectiveness in training.

4.2 Loss functions

Setting the arbitrary scale using O^L, we employ a straightforward L_1 loss ($\mathcal{L}_d$) on estimated depth values and scale-adjusted ground truth as well as a multi-scale gradient loss [Li and Snavely 2018] ($\mathcal{L}_{dg}$) for spatial coherency.

Additionally, following Chen et al. [2017], we include a surface normal loss ($\mathcal{L}_n$), defined by the cosine-similarity between normals computed from estimated depth (n) and ground-truth normals ($\hat{n}$). We also incorporate a multi-scale gradient loss on surface normals, defined effectively on the second derivative of SI depth, promoting better curvature in estimations. This enhances the geometric representation and spatially coherent surface normals:

$$\mathcal{L}_{ng} = \frac{1}{NM} \sum_m \sum_i \left(\nabla \hat{n}_i^m - \nabla n_i^m \right)^2. \tag{6}$$

where ∇n^m is dimensions-wise gradients of the surface normal and M is the number of scales. Our overall loss, combines each component with appropriate weights:

$$\mathcal{L}_{siNet} = \lambda_d \mathcal{L}_d + \lambda_{dg} \mathcal{L}_{dg} + \lambda_n \mathcal{L}_n + \lambda_{ng} \mathcal{L}_{ng}, \tag{7}$$

where $\lambda_d = 1$, $\lambda_{dg} = 0.5$, $\lambda_n = 0.1$, and $\lambda_{ng} = 0.01$.

4.3 Training details

We adopt the architecture from Xian et al. [2018] with EfficientNet-b7 [Tan and Le 2019] as the backbone for our scale-invariant depth estimation network. Training resolution is 1024×1024, and during inference, we resize to a maximum dimension of 1024 pixels while maintaining aspect ratio. We use the ADAM optimizer (learning rate of $1e-3$) for 30 epochs with a batch size of 2 to train the network.

Given the scarcity of high-resolution real-world datasets with scale-invariant (SI) depth ground truth, we exclusively train on the synthetic dataset Hypersim [Roberts et al. 2021]. This dataset offers high-resolution ground truth for both SI depth and surface normals. To avoid over-fitting to the intrinsic parameters of this dataset we use diverse crop augmentations as described in supplementary material. The simplified task for the SI network and generalization from SSI inputs enable in-the-wild SI monocular depth estimation by training solely on this synthetic indoor dataset, as demonstrated in qualitative evaluations.

Figure 5: Figure depicts the in-the-wild performance of our model in accurately modeling the scene compared to LeRes [Yin et al. 2021b]. Our model can model the 3D shape of various scenes with different depth distributions at a high resolution and with precise boundary accuracy. As highlighted by the insets, the absence of details in LeReS causes geometrical distortions in the projected point clouds. Our accurate boundary localization enables precise shape representation, even for complex in-the-wild scenes.

Image credits: Death to the Stock Photo

Table 1: Quantitative evaluation for scale-invariant depth estimation. We are reporting surface normal metric with $t = 11.25°$.

Methods	Middlebury						iBims-1								DIODE					
	Structure and Shape			Surface Normal		Edges	Structure and Shape			Surface Normal			Edges		Structure and Shape			Surface Normal		Edges
	RMSE ↓	Abs. ↓	δ_1 ↑	∠ Dist ↓	% wtn $t°$ ↑	D^3R ↓	RMSE ↓	Abs. ↓	δ_1 ↑	∠ Dist ↓	% wtn $t°$ ↑	D^3R ↓	ϵ^{acc}_{DBE} ↓	ϵ^{comp}_{DBE} ↓	RMSE ↓	Abs. ↓	δ_1 ↑	∠ Dist ↓	% wtn $t°$ ↑	D^3R ↓
MD [Li and Snavely 2018]	58.4	39.4	44.2	74.9	16.6	0.552	2.20	47.2	48.6	51.1	12.1	0.596	3.29	58.4	2.92	57.2	42.3	53.3	7.80	0.857
MC [Li et al. 2019]	61.4	50.6	42.8	73.3	16.2	0.694	1.07	22.7	60.6	48.0	11.7	0.724	4.08	57.4	1.63	30.5	52.1	48.6	10.30	0.901
VN ICCV [Yin et al. 2019]	64.4	49.3	41.5	75.8	15.9	0.698	0.74	13.7	80.4	39.9	22.8	0.707	4.27	30.9	0.97	16.0	77.2	39.1	21.2	0.910
LeReS [Yin et al. 2021b]	42.6	34.3	**56.0**	65.1	22.7	0.415	0.88	20.2	68.7	**25.3**	**42.1**	0.431	2.25	20.1	1.49	27.3	56.1	28.9	32.0	0.745
Ours SI	**41.3**	**34.0**	55.4	**58.4**	**24.1**	**0.215**	**0.69**	**11.7**	**86.7**	26.9	35.1	**0.342**	**1.69**	**16.0**	**0.89**	**15.7**	**80.1**	**26.0**	**36.8**	**0.742**

5 EXPERIMENTS AND EVALUATION

We present an evaluation of our method using datasets that were not included in the training, namely Middlebury2014 [Scharstein et al. 2014], iBims1 [Koch et al. 2018], and DIODE [Vasiljevic et al. 2019]-(indoor).

5.1 SI depth evaluation

To perform a comprehensive numerical evaluation of our SI depth, we have selected three categories of metrics. (I) RMSE, Absolute relative difference (Abs.) and $\delta_1 = \max(\frac{z}{z^*}, \frac{z^*}{z}) < 1.25$ assess shape and structure of the scene. (II) Angle Distance ($\angle\ Dist$) [Chen et al. 2017] and % $wtn\ t$ [Chen et al. 2017] focus on surface orientation accuracy. (III) D^3R [Miangoleh et al. 2021], ϵ^{comp}_{DBE} and ϵ^{acc}_{DBE} [Koch et al. 2018] measure depth discontinuities, edge completeness and location accuracy, respectively.

The results, presented in Table 1, demonstrate that our method significantly enhances the accuracy of scale-invariant depth estimation across various metrics. Compared to the current state-of-the-art (SOTA) techniques, our method consistently produces superior structures, depth distribution, and boundary accuracy. Moreover, the evaluation of surface orientation metrics indicates that our method outperforms existing approaches when applied to high-resolution datasets like Middlebury and DIODE, while achieving competitive performance on the iBims1 dataset.

We also present qualitative comparisons of our method to SOTA in Figure 1, 3, 6, and 5. Our results exhibit significantly improved boundary localization compared to the competing networks. To visualize the reconstructed shape and structure of the scene, we project the images into 3D point clouds using our depth estimations and compare them to the results obtained by LeReS for a variety of in-the-wild images in Figure 1, 6, and 5. We use the focal length values estimated by LeReS for projection. Our results in Figure 6 are provided in various angles to demonstrate the precise scene shape and structure generated by our method in addition to the accurately captured object boundaries. In contrast, LeReS by only relying on a low-resolution SSI depth as input, fails to accurately detect many object boundaries, resulting in an inadequate representation of the complex scenes. Our method's high level of detail and accurate geometry enables the use of SI depth in 3D photography applications. We provide results for the 3D Photography task [Shih et al. 2020] with comparisons to other SI MDE methods as well as further qualitative examples in the supplementary material.

5.1.1 Metric Depth Estimation Methods. Metric depth estimation models aim to go beyond SI depth by removing the arbitrary scale and defining depth in meters. However, when the focal length is unknown or not accounted for in the metric setup, these models tend to generate results with a scale mismatch. We summarize our comparison against metric depth estimation models in Table 2. PatchFusion [Li et al. 2024] and Zoedepth [Bhat et al. 2023] perform poorly when evaluated on unseen datasets of iBims1 and Middlebury2014 due to the mismatch between their training focal length and images from these datasets. Metric3D [Yin et al. 2023], on the other hand, takes the camera parameters into account and generates comparable results on iBims1 dataset. However, it fails to faithfully recover metric depth for high-resolution and complex dataset of Middlebury2014. To ensure a fair comparison, we match the scale outputs to that of the ground truth and include these scale invariant evaluations in Table 2 as well.

Table 2: Quantitative evaluation of metric depth estimation methods. These networks often inaccurately estimate depth due to focal length mismatch. Accurate results are achieved only after scale adjustment (marked with †).

Methods	Middlebury				iBims-1			
	RMSE ↓	Abs. ↓	δ_1 ↑	D^3R ↓	RMSE ↓	Abs. ↓	δ_1 ↑	D^3R ↓
Metric3D	218.6	186.8	58.9	0.443	0.60	17.5	79.5	19.3
Zoedepth	229.8	169.4	22.1	0.245	0.80	16.8	71.6	0.368
PatchFusion	223.7	150.3	22.4	0.076	0.86	20.9	58.4	0.230
Metric3D †	51.7	45.6	50.8	0.400	**0.46**	8.35	**92.5**	0.440
Zoedepth †	47.2	43.4	56.8	0.239	0.51	**7.96**	92.4	0.369
PatchFusion †	42.9	40.4	**58.1**	**0.076**	0.57	9.24	91.2	**0.248**
Ours SI	**41.3**	**34.0**	55.4	0.215	0.69	11.7	86.7	0.342

ZoeDepth [Bhat et al. 2023], which also employs SSI depth as input to estimate metric depth, fails to match our performance in detail generation. We believe this is due to our SSI-network's superior detail generation and our adaptation of multi-resolution processing. Metric3D [Yin et al. 2023] achieves the low edge metric by estimating SI-depth in a single forward pass, reflecting the task's complexity. PatchFusion shows better edge scores than ours, while performing worse in overall structure. As we will discuss in the supplementary material, PatchFusion uses a model with 700M parameters, compared to our 180M parameters, and takes 60 times longer to process at 3 minutes vs. 3 seconds average, due to its patch-based iterative approach. Our approach recovers high-quality details with only 3 forward passes across our pipeline in a faster runtime that is comparable to other state-of-the-art methods.

As Table 2 shows, metric MDE models show a significant drop in performance in the Middlebury dataset when compared to their results on iBims-1. With its high-resolution ground-truth in complex environments, zero-shot evaluation on the Middlebury dataset is more challenging for metric or SI formulations where the training data is limited. Our superior performance on this dataset demonstrates the effectiveness of the SSI input in our formulation, which allows us to generalize to complex scenes even with synthetic-only training of SI MDE.

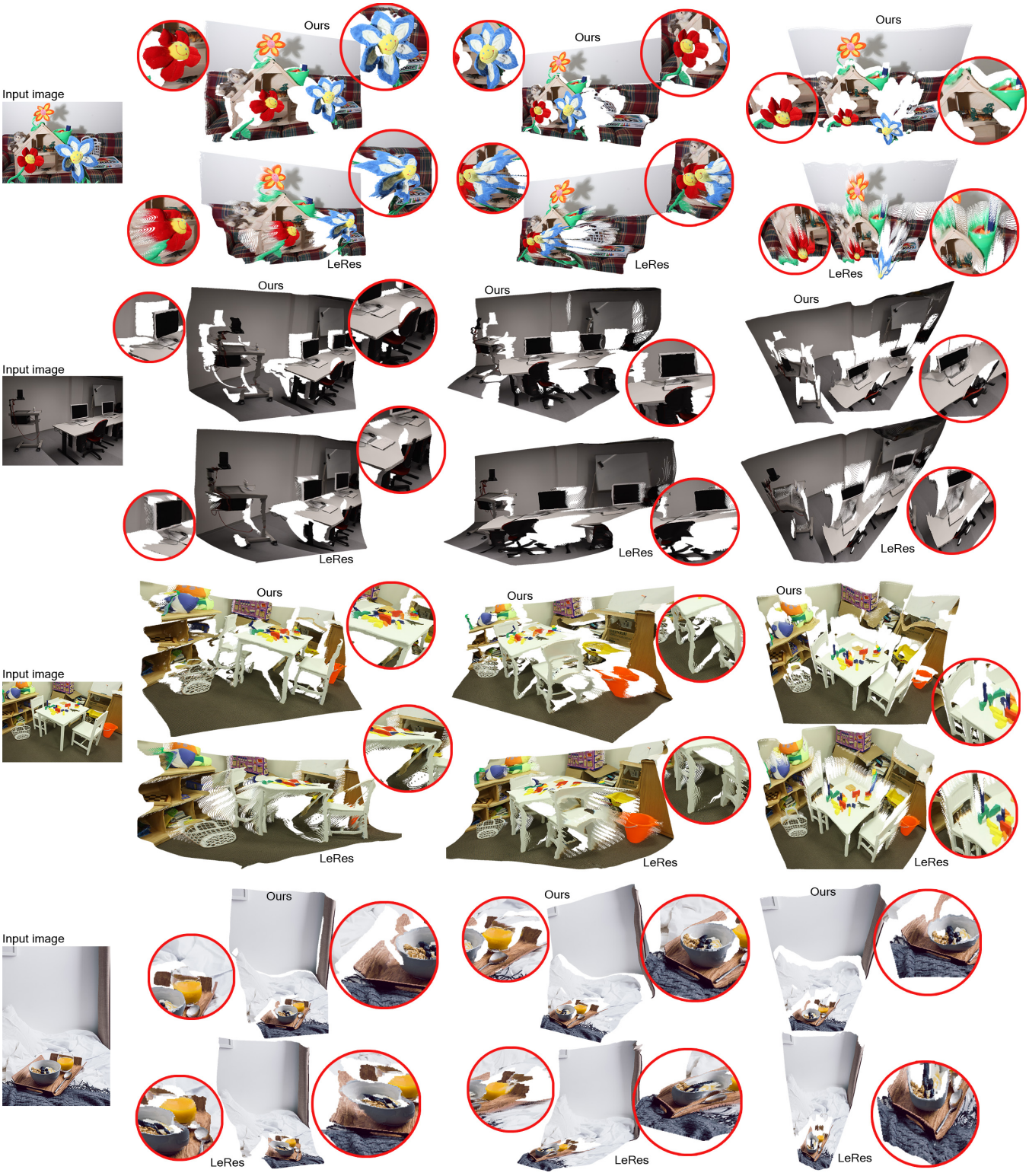

Figure 6: 3D point clouds generated by our SI-depth and LeRes [Yin et al. 2021b] from various views shows leveraging our crisp SSI depth, our SI depth produces finer details. This results in a more precise representation of shape compared to the less detailed and inaccurate results of LeRes. The missing details in LeReS leads to distortion and blending of the details into the background. (see flowers in the first row, monitors in the 2nd row, objects on the table in the 3rd row and the tray in the last row as emphasized by the insets.) Image credits: [Scharstein et al. 2014], [Koch et al. 2018], Death to the Stock Photo

Table 3: Overview of the ordinal depth quantitative comparison. "bmd" indicates boosted using [Miangoleh et al. 2021].

Methods	Middlebury		iBims-1			DIODE	
	Ord. ↓	D^3R ↓	Ord. ↓	D^3R ↓	ϵ_{DBE}^{comp} (ϵ_{DBE}^{acc}) ↓	Ord. ↓	D^3R ↓
VN TPAMI [Yin et al. 2021a]	0.213	0.613	0.140	0.623	52.9(4.68)	0.167	0.935
SGR [Xian et al. 2020]	0.221	0.507	0.200	0.522	34.6(2.37)	0.288	0.831
Ken Burns [Niklaus et al. 2019]	0.221	0.453	0.125	0.487	22.8(2.19)	0.226	0.883
LeReS SSI [Yin et al. 2021b]	0.199	0.444	0.108	0.459	23.9(2.40)	0.143	0.820
MDS [Ranftl et al. 2020]	0.176	0.449	0.128	0.458	28.4(2.22)	0.167	0.846
DPT [Ranftl et al. 2021]	0.162	0.369	0.101	0.403	23.2(2.20)	0.134	0.818
DepthAnything [Yang et al. 2024]	**0.092**	0.155	**0.051**	0.334	12.5(1.94)	**0.074**	**0.771**
Our SSI	0.190	0.339	0.112	0.345	22.5(2.16)	0.147	0.817
SGR-bmd	0.210	0.280	0.196	0.411	22.3(2.29)	0.287	0.804
MDS-bmd	0.162	0.201	0.126	0.368	23.3(2.15)	0.165	0.828
Ours bmd	0.174	**0.120**	0.116	**0.255**	**11.5(2.23)**	0.150	0.790

5.2 SSI depth evaluation

We evaluate the performance of our SSI depth estimation module against SOTA methods, as detailed in Table 3. The assessment includes metrics such as D^3R [Miangoleh et al. 2021], $\epsilon_{\text{DBE}}^{\text{comp}}$, and $\epsilon_{\text{DBE}}^{\text{acc}}$ [Koch et al. 2018] to gauge the quality of depth discontinuities, which is the primary focus of our work. Additionally, to measure the structural coherence of the estimated depth map, we employ the ordinal relation metric (ORD) proposed by Xian et al. [2020].

Table 3 demonstrates that our method consistently outperforms other baselines in generating detailed depth maps, benefiting from our novel loss combination, except when compared to DepthAnything [Yang et al. 2024]. However, employing a CNN backbone allows our method to be boosted by Miangoleh et al. [2021]'s boosting framework which is not applicable to transformer-based DepthAnything and DPT [Ranftl et al. 2021]. Results indicate that our boosted method generates substantial amount of details, outperforming every baseline by a significant margin. The qualitative examples presented in Figure 1 and 7 also illustrate the significant improvement of our SSI depth over other baselines in generating details.

However, this improvement in details comes with a slight degradation in the ORD metric representing overall structure. We believe this is the result of the limited network capacity which makes it harder for CNNs to maintain global coherency when generating details [Miangoleh et al. 2021]. When SSI inputs are utilized for SI depth, however, we see that our network with accurate details is more effective in providing important information for the SI network. We demonstrate in Section 5.3.2 that the final SI-depth experiences a performance loss when our SSI model is substituted with another SOTA method. This underscores the high level of details fed to the SI depth model by our SSI method, making it more effective in simplifying the task of the SI network, leading to superior performance.

5.3 Ablation studies

5.3.1 SSI MDE ablation. To assess the impact of our relaxed ranking loss on enabling mixed dense-sparse training, we conduct an ablation study. Our study utilizes a subset of the Hypersim dataset [Roberts et al. 2021] consisting of 10,000 images (20 per scene) randomly sampled from the train split. We train multiple networks for 20 epochs, employing different loss functions as outlined in Table 5. Throughout all setups, we use $\mathcal{L}_{ssig}$ as a default component due to its crucial role in generating spatially coherent estimations. The results support our discussion in Section 4 that a naive combination of ranking and SSI loss yields inferior performance compared to utilizing either of them individually. However, with our novel definition of the ordinal loss a higher performance is achieved when two losses are combined.

Table 4: Overview of the influence of SSI depth in improving the performance of SI depth estimation.

Methods	Middlebury				iBims-1			
	Abs. ↓	δ_1 ↑	∠ Dist ↓	D^3R ↓	Abs. ↓	δ_1 ↑	∠ Dist ↓	D^3R ↓
Only RGB	53.5	38.3	74.9	0.728	26.4	65.4	46.9	0.715
+ MiDaS $O^{L,H}$	48.3	50.2	63.4	0.335	15.4	74.9	38.0	0.536
+ Our SSI $O^{L,H}$ (Ours-si)	**36.2**	**58.8**	**61.1**	**0.285**	**11.7**	**86.1**	**30.7**	**0.379**
w/o O^H	37.2	57.7	**61.1**	0.383	12.4	83.2	32.5	0.437
w/o Normal loss	42.4	53.8	66.9	0.290	12.4	83.2	36.8	0.409

Table 5: Our ordinal loss can be combined with SSI yielding superior performance as opposed to ranking loss [Chen et al. 2016].

Methods	Hypersim		iBims-1	
	Ord.* ↓	D^3R ↓	Ord. ↓	D^3R ↓
+ $\mathcal{L}_{ssi}$	0.185	0.496	0.156	0.520
+ $\mathcal{L}_{ranking}$	0.238	0.570	0.235	0.573
+ $\mathcal{L}_{ssi}$ + $\mathcal{L}_{ranking}$	0.227	0.562	0.213	0.565
+ $\mathcal{L}_{ssi}$ + $\mathcal{L}_{so}$ (*ours*)	**0.184**	**0.486**	**0.147**	**0.497**

5.3.2 SI MDE ablation. In order to assess the impact of using SSI depth estimations in the training of SI depth, we conduct an isolated test as summarized in Table 4. For this experiment, we train our scale-invariant network for 4 epochs with various settings and evaluated its performance. Only using RGB as input demonstrates very low performance. This shows a direct regression of SI-depth leads to a low performance due to the complexity of the task and limitations of neural networks.

Feeding SSI depth estimations alongside RGB simplifies the SI network's task and enhances the SI depth model's capabilities, improving structure, surface normal, and detail metrics as reported in Table 4. Additionally, our model's SSI estimations outperform those from MiDaS, indicating their superiority in providing depth discontinuities and easing the scale-invariant depth estimation network's task. Despite MiDaS showing better ORD performance in Table 3, this ablation underscores that a higher level of details does indeed simplify the task more effectively.

The results of the variant with omitted high-resolution depth estimations exhibit reduced performance in details, assessed by D^3R. Interestingly, it also indicates a decreased ability to estimate the structure and shape of the scene, evaluated by *Abs.* and δ_1. We believe as the network allocates its capacity to recover details, it compromises its ability to accurately recover the structure.

Finally, Table 4 indicates that the surface normal loss plays a crucial role in faithfully recovering the shape and structure of the scene. A surface normal loss promotes better shape and structure by penalizing incorrect surface orientations, especially on flat regions. The D^3R evaluation shows that the network's ability to recover details is not heavily affected by the surface normal loss, as it only decreases marginally. This can be attributed to the dominance of flat surfaces in the surface normal loss, as they constitute the majority of the images.

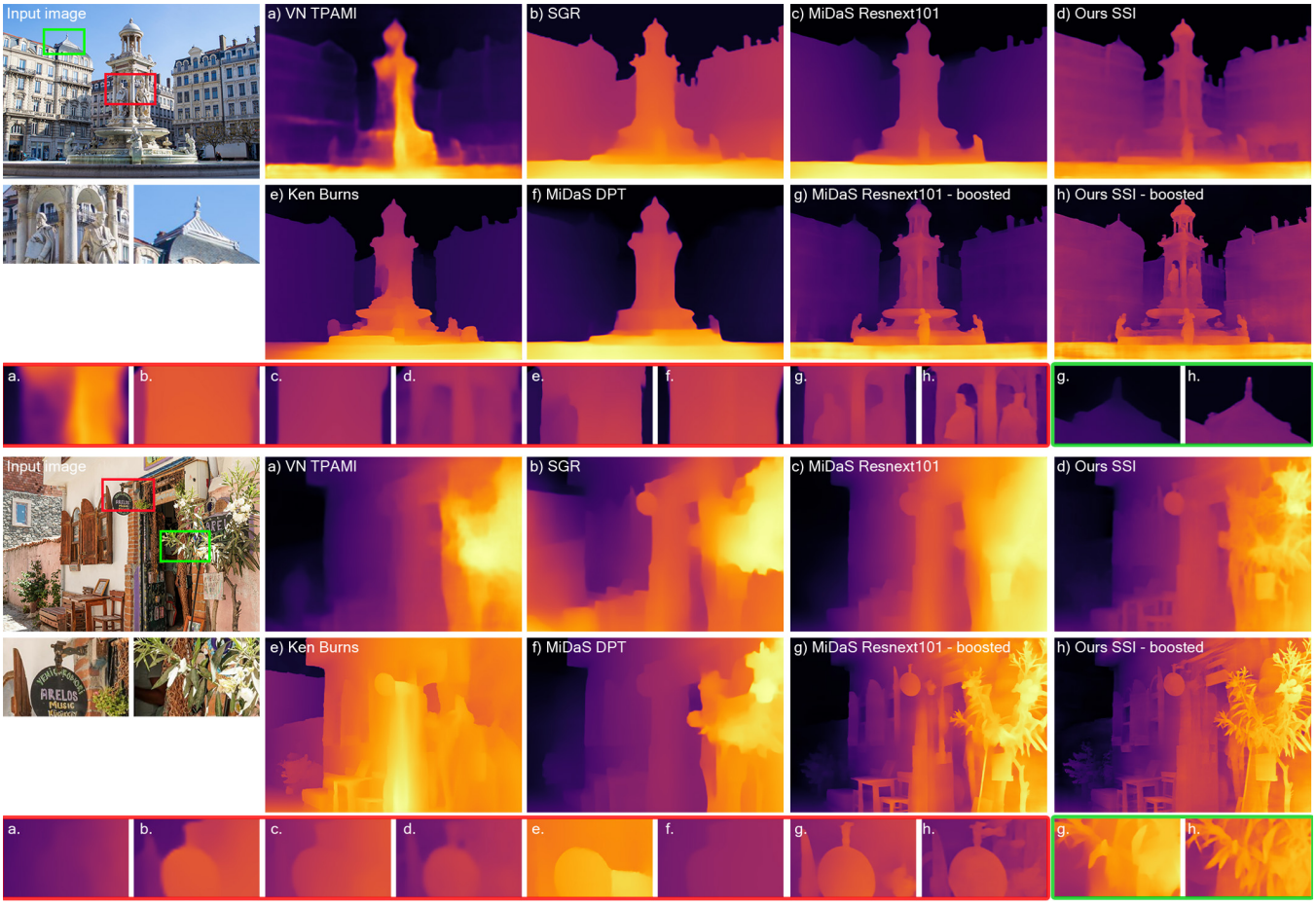

Figure 7: Qualitative comparison of scale and shift invariant networks in-the-wild reveals that our SSI network produces crisp depth boundaries compared to other methods. The results of our high-resolution boosted model exhibit even more refined depth boundaries.

Image credits: @Diogo Nunes, @Mert Kahveci

6 LIMITATIONS

Our method focuses on generating highly detailed, SI depth estimations. The quality of our estimations, however, depends on the quality of the input images. For low-resolutions, or noisy images, our method may fail to generate sharp results. This mainly comes from our high-resolution ordinal input failing to give accurate depth discontinuities in the case of image noise. We present further analysis and discussion on this in the supplementary material.

We utilized a CNN architecture for both our SSI and SI depth estimation networks. This choice allows us to generate estimations at high resolutions with increasing details for our SSI network. However, the SI depth network struggles with increasing resolution due to global scale-invariant constraints, requiring the network to reason about every pixel in the image together. This constraint necessitates the entire input image to fit in our native resolution to generate consistent structures. This makes transformer-based architectures a good candidate for SI depth with SSI inputs. The lack of large SI depth datasets, however, creates a challenge for training a transformer-based architecture than for CNNs.

7 CONCLUSION

We present a geometric monocular depth estimation method that can generate highly detailed and geometrically consistent reconstructions from a single image. To achieve this, we introduce an SSI depth estimation method that can generate sharper depth discontinuities. Using our SSI depth, we formulate SI MDE with SSI inputs, simplifying the SI MDE problem to the enforcement of geometric constraints. We show that through this simplification, in-the-wild generalization of SI task is achievable through only training with a synthetic indoors dataset, inheriting the generalization capability of SSI formulations that can be trained on diverse datasets, Using our estimated SSI depth as input, we show that our novel scale-invariant depth estimation formulation can generate highly detailed results even for complex scenes in the wild. We have demonstrated state-of-the-art performance for scale-invariant depth estimation through zero-shot evaluations.

ACKNOWLEDGMENTS

We would like to thank Long Mai and Obumneme Dukor for their help during the early stages of this work, and Chris Careaga for his feedback on the text and the voice-over of our video. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [RGPIN-2020-05375].

REFERENCES

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. 2021. Adabins: Depth estimation using adaptive bins. In *Proc. CVPR*.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. 2022. LocalBins: Improving Depth Estimation by Learning Local Distributions. In *Proc. ECCV*.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. 2023. ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth. `arXiv:2302.12288 [cs.CV]` (2023).

Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. 2016. Single-image depth perception in the wild. In *Proc. NeurIPS*.

Weifeng Chen, Shengyi Qian, and Jia Deng. 2019b. Learning single-image depth from videos using quality assessment networks. In *Proc. CVPR*.

Weifeng Chen, Donglai Xiang, and Jia Deng. 2017. Surface normals in the wild. In *Proc. ICCV*.

Xiaotian Chen, Xuejin Chen, and Zheng-Jun Zha. 2019a. Structure-Aware Residual Pyramid Network for Monocular Depth Estimation. In *Proc. IJCAI*.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *Proc. CVPR*.

Sungjoon Choi, Qian-Yi Zhou, Stephen Miller, and Vladlen Koltun. 2016. A large dataset of object scans. `arXiv:1602.02481 [cs.CV]` (2016).

Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. 2021. Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets from 3D Scans. In *Proc. ICCV*.

David Eigen and Rob Fergus. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proc. ICCV*.

David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. In *Proc. NeurIPS*.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. 2018. Deep ordinal regression network for monocular depth estimation. In *Proc. CVPR*.

Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. 2017. Unsupervised monocular depth estimation with left-right consistency. In *Proc. CVPR*.

Yiwen Hua, Puneet Kohli, Pritish Uplavikar, Anand Ravi, Saravana Gunaseelan, Jason Orozco, and Edward Li. 2020. Holopix50k: A large-scale in-the-wild stereo image dataset. In *Proc. CVPR Workshops*.

Jinyoung Jun, Jae-Han Lee, Chul Lee, and Chang-Su Kim. 2022. Depth map decomposition for monocular depth estimation. In *Proc. ECCV*.

Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. 2018. Evaluation of CNN-based single-image depth estimation methods. In *Proc. ECCV Workshops*.

Philipp Krähenbühl. 2018. Free supervision from video games. In *Proc. CVPR*.

Jae-Han Lee and Chang-Su Kim. 2019. Monocular Depth Estimation Using Relative Depth Maps. In *Proc. CVPR*.

Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. 2024. PatchFusion: An End-to-End Tile-Based Framework for High-Resolution Monocular Metric Depth Estimation. In *Proc. CVPR*.

Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. 2019. Learning the depths of moving people by watching frozen people. In *Proc. CVPR*.

Zhengqi Li and Noah Snavely. 2018. Megadepth: Learning single-view depth prediction from internet photos. In *Proc. CVPR*.

Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, et al. 2021. OpenRooms: An Open Framework for Photorealistic Indoor Scene Datasets. In *Proc. CVPR*.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *Proc. ECCV*.

S Mahdi H Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yağız Aksoy. 2021. Boosting Monocular Depth Estimation Models to High-Resolution via Content-Adaptive Multi-Resolution Merging. In *Proc. CVPR*.

Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 2019. 3D Ken Burns effect from a single image. *ACM Trans. Graph.* (2019).

Juewen Peng, Zhiguo Cao, Xianrui Luo, Hao Lu, Ke Xian, and Jianming Zhang. 2022. BokehMe: When Neural Rendering Meets Classical Rendering. In *Proc. CVPR*.

Michael Ramamonjisoa, Yuming Du, and Vincent Lepetit. 2020. Predicting Sharp and Accurate Occlusion Boundaries in Monocular Depth Estimation Using Displacement Fields. In *Proc. CVPR*.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision transformers for dense prediction. In *Proc. ICCV*.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* (2020).

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. 2021. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proc. ICCV*.

Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. 2014. High-resolution stereo datasets with subpixel-accurate ground truth. In *Proc. GCPR*.

Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 2020. 3D photography using context-aware layered depth inpainting. In *Proc. CVPR*.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. 2019. The Replica dataset: A digital replica of indoor spaces. `arXiv:1906.05797 [cs.CV]` (2019).

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proc. ICML*.

Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Proc. ECCV*.

Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. 2019. Diode: A dense indoor and outdoor depth dataset. `arXiv:1908.00463 [cs.CV]` (2019).

Neal Wadhwa, Rahul Garg, David E Jacobs, Bryan E Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T Barron, Yael Pritch, and Marc Levoy. 2018. Synthetic depth-of-field with a single-camera mobile phone. *ACM Trans. Graph.* (2018).

Lijun Wang, Jianming Zhang, Yifan Wang, Huchuan Lu, and Xiang Ruan. 2020a. CLIFFNet for Monocular Depth Estimation with Hierarchical Embedding Loss. In *Proc. ECCV*.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. 2020b. Tartanair: A dataset to push the limits of visual SLAM. In *Proc. IROS*.

Alex Wong and Stefano Soatto. 2019. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In *Proc. CVPR*.

Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. 2018. Monocular relative depth perception with web stereo data supervision. In *Proc. CVPR*.

Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. 2020. Structure-guided ranking loss for single image depth prediction. In *Proc. CVPR*.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proc. CVPR*.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *Proc. CVPR*.

Wei Yin, Yifan Liu, and Chunhua Shen. 2021a. Virtual Normal: Enforcing Geometric Constraints for Accurate and Robust Depth Prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* (2021).

Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. 2019. Enforcing geometric constraints of virtual normal for depth prediction. In *Proc. ICCV*.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. 2023. Metric3D: Towards Zero-shot Metric 3D Prediction from A Single Image. In *Proc. ICCV*.

Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. 2021b. Learning to recover 3d scene shape from a single image. In *Proc. CVPR*.

Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. 2022. NeWCRFs: Neural Window Fully-connected CRFs for Monocular Depth Estimation. In *Proc. CVPR*.

Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2018. T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In *Proc. ECCV*.

Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T Freeman. 2015. Learning ordinal relationships for mid-level vision. In *Proc. ICCV*.