Computational
Photography Lab.

Find the project web page here:
https://yaksoy.github.io/sidepth/

Scale-Invariant Monocular Depth Estimation via SSI Depth Supplementary Material

S. Mahdi H. Miangoleh
Simon Fraser University
Canada

Mahesh Reddy
Simon Fraser University
Canada

Yağız Aksoy
Simon Fraser University
Canada

In this supplementary document, we provide:

(i) Additional qualitative examples showcasing the performance of our scale-invariant (SI) depth estimation model in Section A. Please refer to Figure 1, 4, and 3, which serve as an extension to Figures 1, 2, 3, and 6 of the main document.

(ii) Further qualitative examples highlighting the capabilities of our scale and shift-invariant (SSI) depth estimation model in Section B.1. For additional insights, refer to Figure 5, 6, and 7, extending the content presented in Figure 5 of the main document.

(iii) Elaboration on the training details of our SSI depth estimation network in Section B.2.

(iv) An expanded discussion on the limitations of our work in Section C.

(v) A detailed network complexity and runtime analysis comparing our method to state of the art in Section D.

(vi) An extension to Table 2 of the main paper providing extra metrics in Table 1.

A SI DEPTH ESTIMATION - ADDITIONAL RESULTS

We extend Figures 1, 2, 3, and 6 of the main document and present additional qualitative results of our scale invariant depth estimation in Figure 1, 4, and 3.

In Figure 4, we present visualizations of the estimated depth maps produced by our method alongside those of other state-of-the-art baselines. MD [Li and Snavely 2018], MC [Li et al. 2019], and VNI [Yin et al. 2019] attempt to directly estimate SI depth from the RGB input. As depicted in Figure 4, these approaches are limited to providing an overall scene structure with blurry object boundaries and significant details are missing. LeRes [Yin et al. 2021], leveraging the generalizability of SSI depth, achieves more robust results. However, our method employs a second high-resolution forward path from the SSI network, supplying local details that contribute to highly detailed SI-depth estimations.

For a clearer assessment of the quality of the estimated depth maps generated by our method compared to LeReS, we present projected point clouds using the estimated depth in Figure 3, alongside the ground truth 3D point cloud. This figure illustrates that our method adeptly captures the shape of the scene and successfully recovers intricate details, showcasing its ability to faithfully reproduce the scene's geometry.

Furthermore, we extend our comparison to complex and diverse in-the-wild scenes provided in Figure 6 of the main paper in Figure 1 presenting both the depth maps and projected 3D point clouds in comparison to LeReS. Our depth maps accurately capture the intricate geometry of the scenes, exhibiting precise depth discontinuities and placements. In contrast, LeReS struggles to faithfully reconstruct the shape of these complex scenes due to numerous missing details.

B SSI DEPTH ESTIMATION

B.1 Extra results

Figures 6 and 7 serve as extensions to Figure 5 in the main document, depicting the quality of our SSI depth estimation in comparison to state-of-the-art baselines. Our approach, leveraging a combination of dense SSI and ordinal sparse training, produces finer details compared to all other baselines, including the larger transformer-based MiDaS DPT [Ranftl et al. 2021].

Transformer-based SSI MDE methods lack the boostability observed in CNN-based backbones using the boosting framework of BMD [Miangoleh et al. 2021]. The attention mechanism in transformers enables them to process and relate local features in the images independent of the resolution and spatial distance of the features. This eliminates the limitation posed by the receptive field of CNN-based approaches. As a result, as also demonstrated by others [Miangoleh 2022], transformer performance cannot be boosted by optimizing their input resolution, which is how BMD [Miangoleh et al. 2021] proposes to boost SSI MDE methods. This is while, the utilized CNN backbone of our SSI depth model, enables boostability. Figure 6 and 7 also illustrate that our method produces significantly finer details when boosted, outperforming even boosted versions of other CNN-based baselines like MiDaS Resnext101 [Ranftl et al. 2020] and SGR [Xian et al. 2020].

B.2 Training details

Since SSI loss involves a least-square fitting it can destabilize the training in the first iterations as the number of outlier is high. To stabilize the training we warm-start the ordinal depth network with $\sim 5K$ images by randomly sampling 20 images per scene in the Hypersim [Roberts et al. 2021] dataset by optimizing the L1 loss for one epoch. Next, we continue our training with all the depth datasets (Hypersim [Roberts et al. 2021], OpenRooms [Li et al. 2021], Replica [Straub et al. 2019], Replica [Straub et al. 2019]+GSO [Choi et al. 2016], FSVG [Krähenbühl 2018], TartanAir [Wang et al. 2020], HRWSI [Xian et al. 2020], and Holopix50K [Hua et al. 2020]) and using the combination of our ordinal and SSI loss defined in Section 4 of the main document. We construct a batch size by uniformly sampling images from every dataset and set the batch size to 16. To crop the input image for training, we follow the setup proposed in Miangoleh et al. [2021] to compute the $\mathcal{R}_0$ size for the input image to ensure no pixel in the image is far away from the contextual cues (e.g., depth edges). We then randomly crop from between the receptive size and $\mathcal{R}_0$ and resize to 384×384, to match the receptive field of the network. We randomly apply horizontal flip, color jitter, Gaussian blur, and grayscale data augmentation operations for

Table 1: Extension to Table 2 of the main document. Quantitative evaluation of metric depth estimation methods. These networks often inaccurately estimate depth due to focal length mismatch without scale matching. Accurate results are achieved only after scale adjustment (denoted by †).

Methods	Middlebury						iBims-1							
	Structure and Shape			Surface Normal		Edges	Structure and Shape			Surface Normal		Edges		
	RMSE ↓	Abs. ↓	δ_1 ↑	∠ Dist ↓	% wtn $t°$ ↑	D^{3}R ↓	RMSE ↓	Abs. ↓	δ_1 ↑	∠ Dist ↓	% wtn $t°$ ↑	D^{3}R ↓	$\varepsilon^{\text{acc}}_{\text{DBE}}$ ↓	$\varepsilon^{\text{comp}}_{\text{DBE}}$ ↓
Metric3D	218.6	186.8	58.9	53.5	26.1	0.443	0.60	17.5	79.5	19.3	57.4	0.463	2.32	15.8
Zoedepth	229.8	169.4	22.1	52.1	27.4	0.245	0.80	16.8	71.6	24.7	41.9	0.368	2.29	15.3
PatchFusion	223.7	150.3	22.4	53.8	25.6	0.076	0.86	20.9	58.4	29.7	29.0	0.230	1.87	19.63
Metric3D †	51.7	45.6	50.8	52.2	26.9	0.400	**0.46**	8.35	**92.5**	**19.3**	**57.7**	0.440	2.34	14.1
Zoedepth †	47.2	43.4	56.8	**50.4**	**28.9**	0.239	0.51	**7.96**	92.4	22.9	47.6	0.369	2.36	**13.8**
PatchFusion †	42.9	40.4	**58.1**	53.8	25.6	**0.076**	0.57	9.24	91.2	27.8	33.7	**0.248**	1.87	19.6
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

better generalization. During training, we match the scale and shift of the predicted ordinal estimate with the ground truth disparity using the least-squares criterion. In addition, we set the disparity to zero for sky regions in outdoor datasets.

C LIMITATION DISCUSSION

Our method focuses on generating highly-detailed, high-resolution scale-invariant depth estimations. The quality of our estimations, however, depends on the quality of the input images. For images at very low-resolutions, or for 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 demonstrate this in Figure 2, where our method starts to fail when a high amount of noise is introduced to the image.

D NETWORK COMPLEXITY AND RUNTIME ANALYSIS

For our SSI-network, we adopt the same CNN-based architecture and pretraining weights as MiDaS [Ranftl et al. 2020] and LeReS [Yin et al. 2021]. This results in evaluation consistency and allows us to exploit boostable properties of CNNs as shown by Miangoleh et al. [2021]. Our objective is to generate high-resolution results. Hence, we need to train the SI-network at high-resolutions. Despite the potential advantages of a transformer architecture for our SI-network, the resource-intensive nature of training transformers at high-resolutions led us to choose a CNN with a large receptive field to avoid global structural issues.

The high-resolution SSI-depth we generate relies on the context-aware resolution selection process of R_{20}, designed by Miangoleh et al. [2021]. As this value depends on the image content, our SSI-depth estimation inference happens at different resolutions, hence different runtimes for each image. Consequently, we will report all the runtimes as an average on the high-resolution Middlebury dataset in Table 2 computed on a single Nvidia RTX 2080 GPU.

LeReS [Yin et al. 2021] and Zoedepth [Bhat et al. 2023] generate results at a smaller resolution of 448 and 512 respectively compared to the 1024 for our method. This contributes to their faster runtimes as they are processing smaller images. Metric3D [Yin et al. 2023] achieves a more efficient runtime despite having an inference resolution of 1088, thanks to its single forward pass. But this comes with the loss of many details and poor depth discontinuity performance. Patchfusion [Li et al. 2024], employs a brute-force patch-based approach, which contributes to its huge runtime.

Table 2: Network Architecture and runtime analysis. Runtimes are reported as the average time that each model takes to process images from Middlebury2014 dataset on a single Nvidia RTX 2080 GPU.

Methods	Runtime (seconds)	Number of parameters (millions)
Metric3D [Yin et al. 2023]	0.6	200
Zoedepth [Bhat et al. 2023]	1	300
PatchFusion [Li et al. 2024]	180	700
LeReS [Yin et al. 2021]	2.5	130
Ours	3	180

Table 2 also summarizes that the CNN based network architecture utilized in Our method, LeReS, and Metric3D needs less parameters compared to the much larger transformer-based architectures of Zoedepth and PatchFusion.

REFERENCES

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).

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

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*.

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*.

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*.

Seyed Mahdi Hosseini Miangoleh. 2022. *Boosting Monocular Depth Estimation to High Resolution.* Master's thesis. Simon Fraser University.

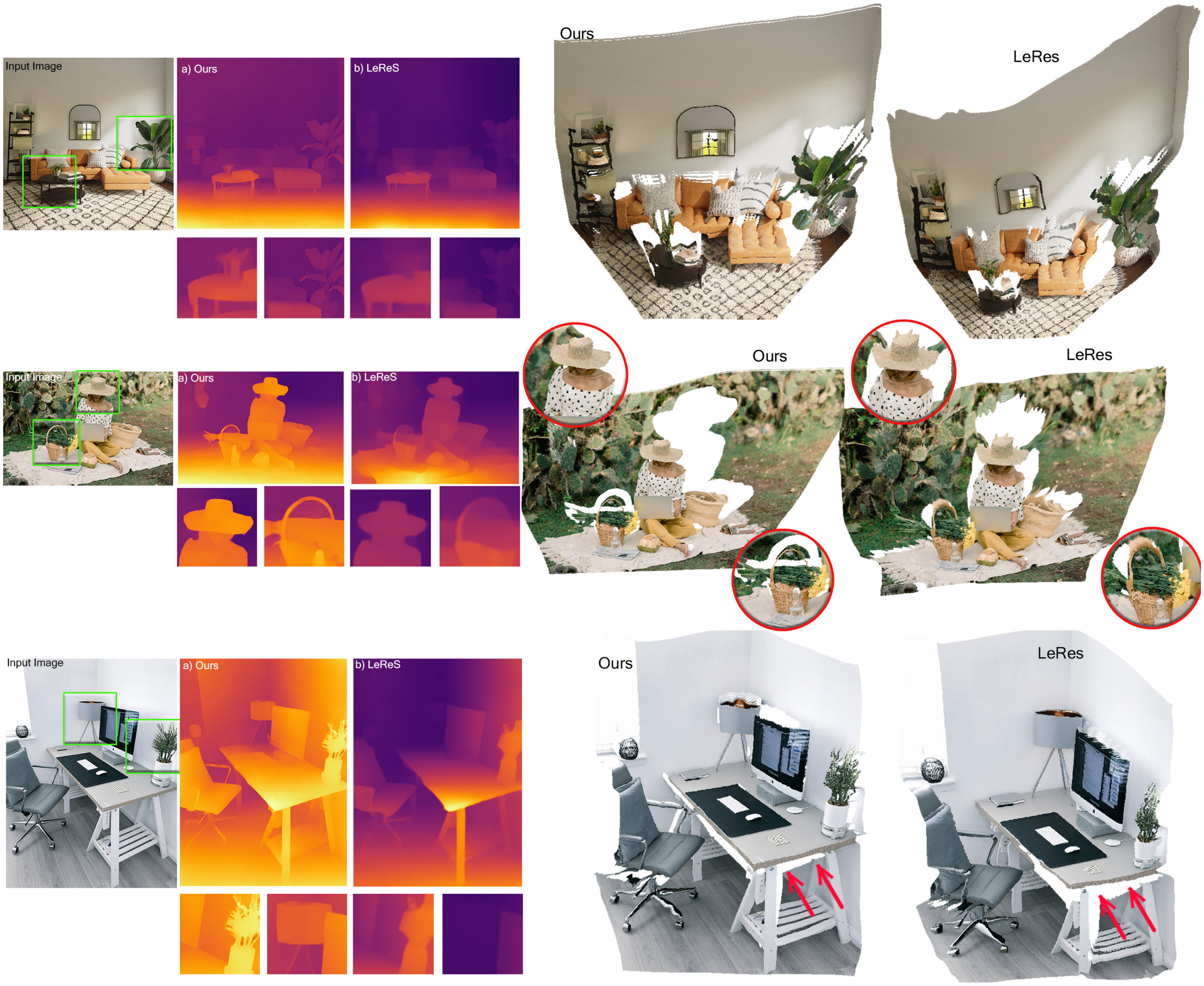

Figure 1: Extension to Figure 6 of the main paper. The figure depicts the in-the-wild performance of our model in accurately modeling the scene compared to LeRes [Yin et al. 2021]. Our model is able to model the 3D shape of various scenes with different depth distributions at a high resolution and with precise boundary accuracy.

Image credits: @Spacejoy, Death to the Stock Photo, @James McDonald

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

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).

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

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*.

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. 2021. Learning to recover 3d scene shape from a single image. In *Proc. CVPR*.

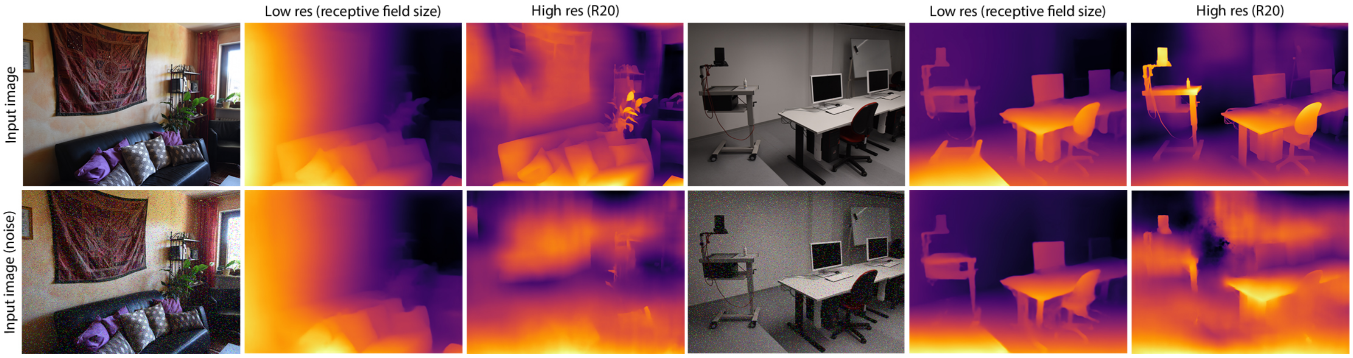
Low res (receptive field size)
High res (R20)
Low res (receptive field size)
High res (R20)
Input image
Input image (noise)

Figure 2: Our high-resolution SSI estimate suffers from artifacts on images with noise as seen in the "High-res (R20)" estimate for both noisy images. This sensitivity to image quality and noise for high-resolution estimation leads to a noisy input for the SI network affecting its performance in recovering details. Image credits: iBims-1 dataset [Koch et al. 2018].

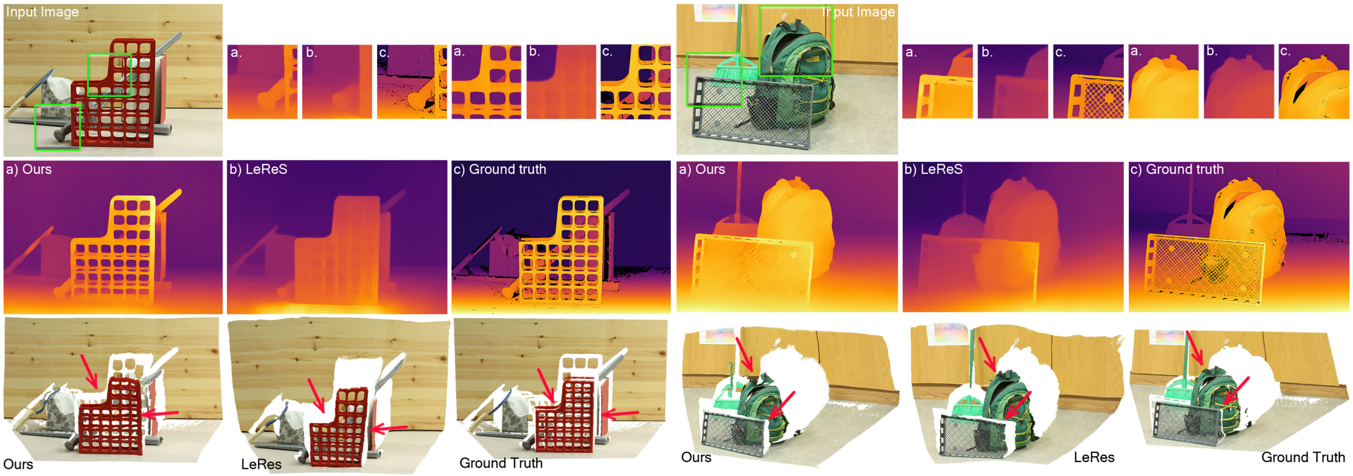

Figure 3: Figure shows samples of our scale-invariant depth estimation and projected point clouds in comparison to LeRes [Yin et al. 2021]. Our method is able to capture the small holes in the object as emphasized in the figure while LeRes generates limited details and fails to predict a correct scene shape. Image credits: Middlebury dataset [Scharstein et al. 2014]

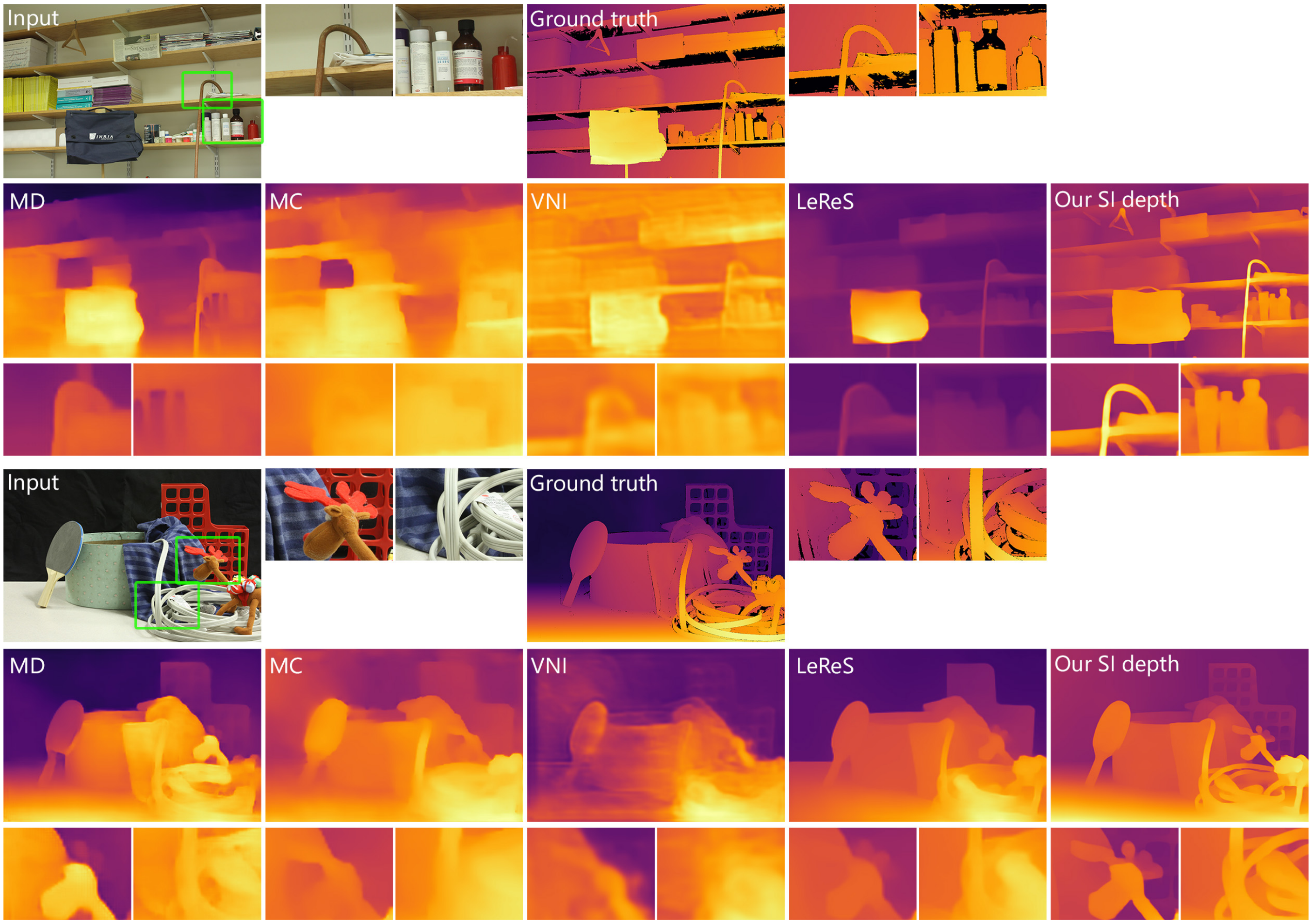

Figure 4: Extension to Figure 3 of the main paper. Qualitative comparison of scale-invariant networks on the Middlebury dataset [Scharstein et al. 2014]. Our scale-invariant network exhibits superior performance in capturing intricate objects with higher levels of depth details compared to the state-of-the-art.

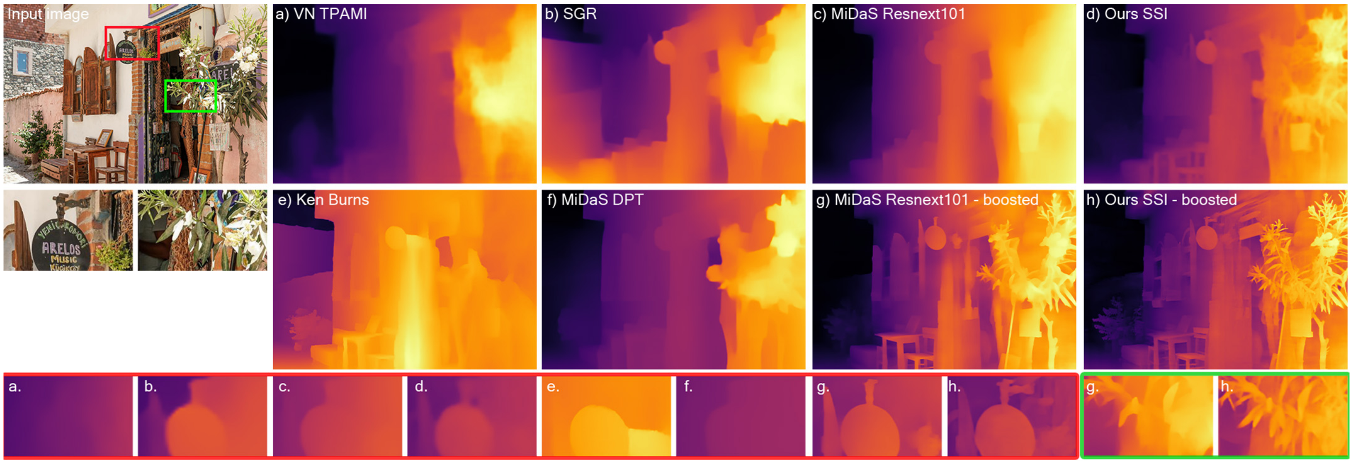

Figure 5: Extension to Figure 5 of the main paper. Qualitative comparison of scale and shift invariant networks in-the-wild reveals that our SSI network produces crisp depth boundaries compared to other methods. The results of our high-resolution boosted model exhibit even more refined depth boundaries.

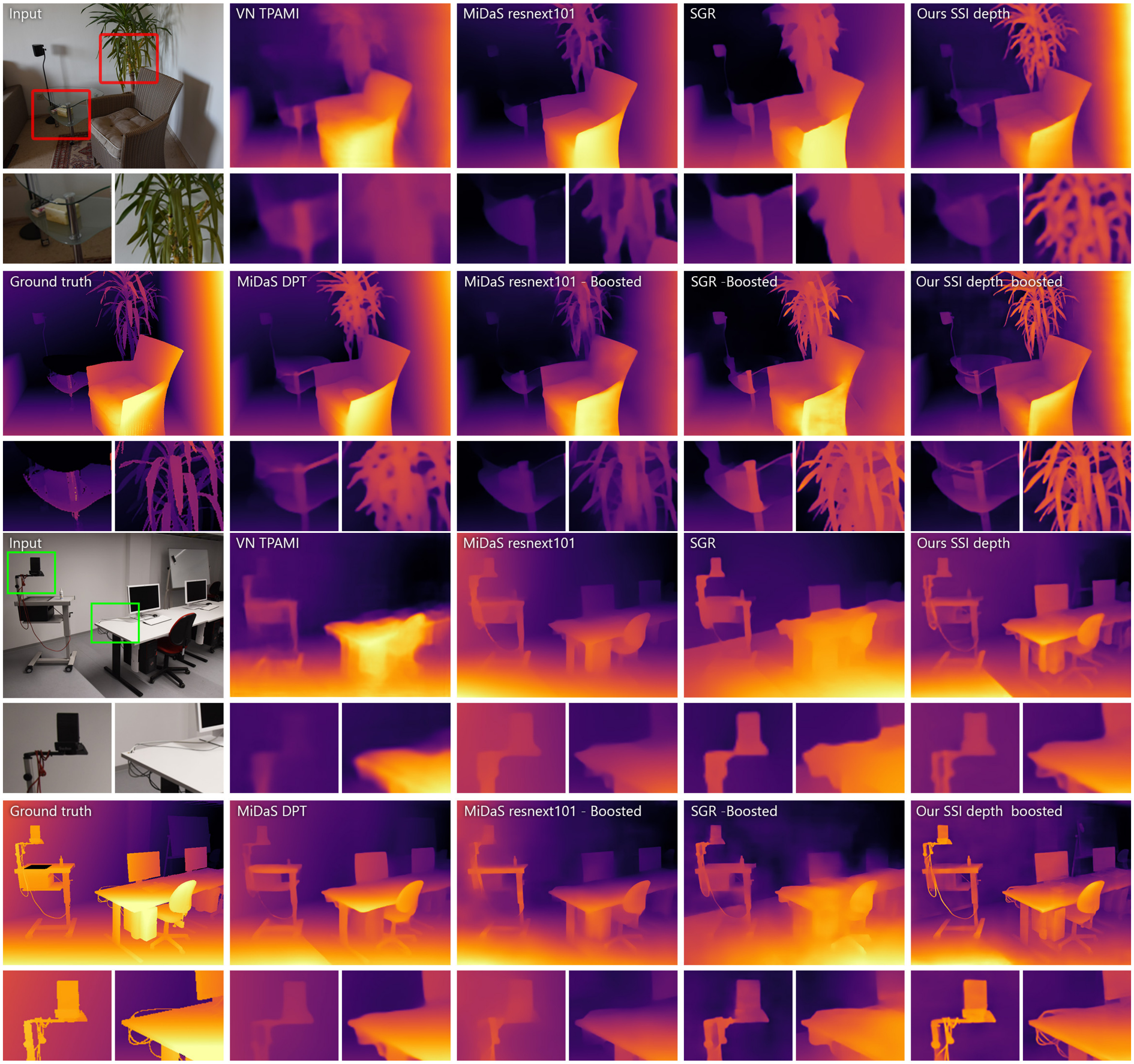

Figure 6: Extension to the Figure 5 of the main paper. Qualitative comparison of scale and shift invariant networks on iBims-1 [Koch et al. 2018] dataset reveals that our SSI network produces crisp depth boundaries compared to other methods. The results of our high-resolution boosted model exhibit even more refined depth boundaries.

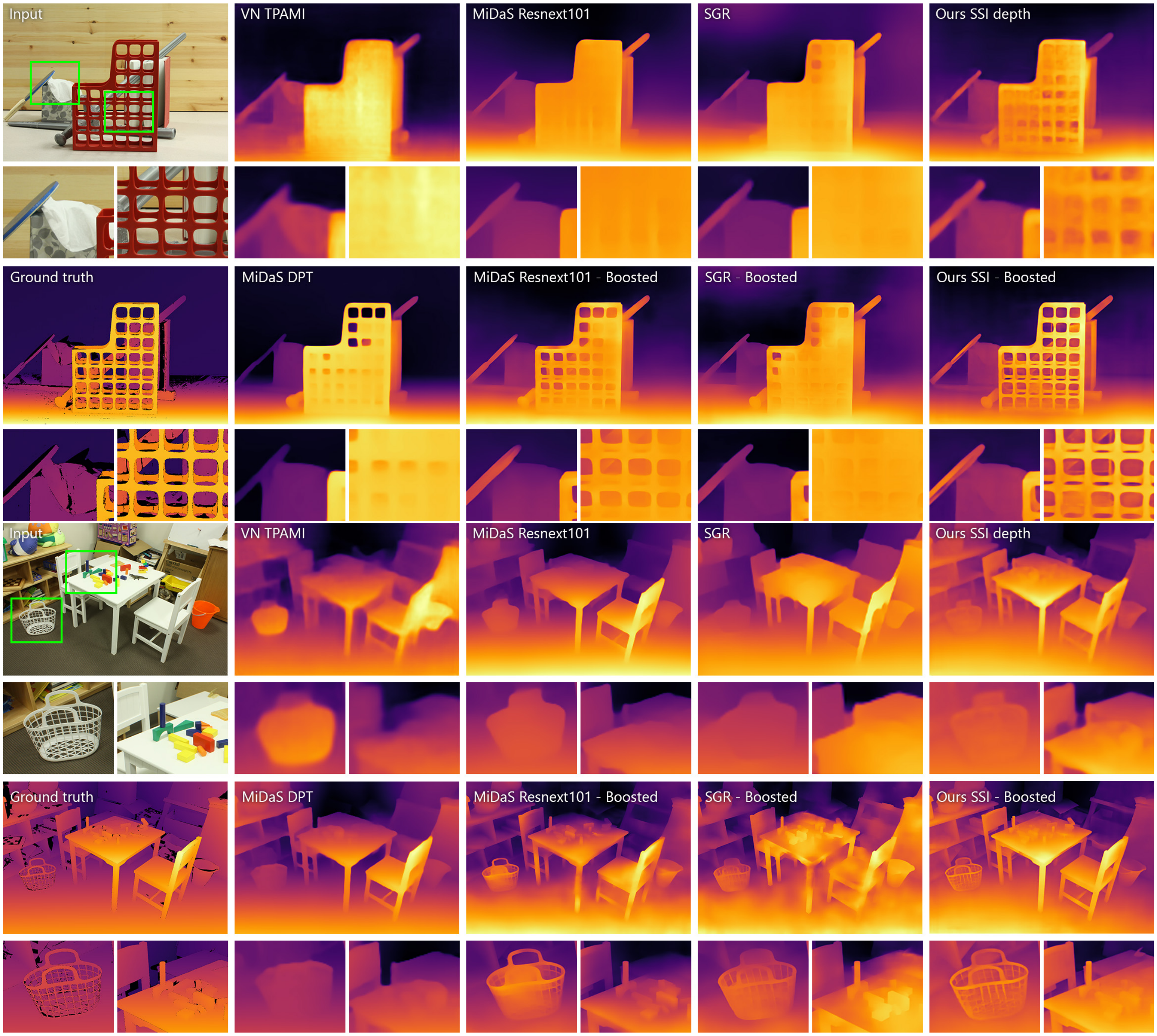

Figure 7: Extension to the Figure 5 of the main paper. Qualitative comparison of scale and shift invariant networks on Middlebury [Scharstein et al. 2014] dataset reveals that our SSI network produces crisp depth boundaries compared to other methods. The results of our high-resolution boosted model exhibit even more refined depth boundaries.