GasMono: Geometry-Aided Self-Supervised Monocular Depth Estimation for Indoor Scenes – Supplementary Material

[1]Chaoqiang Zhao [2]Matteo Poggi [2]Fabio Tosi
[1]Lei Zhou [1]Qiyu Sun [1,*]Yang Tang [2]Stefano Mattoccia

[1]East China University of Science and Technology [2]University of Bologna

This document provides additional details for ICCV paper "GasMono: Geometry-Aided Self-Supervised monocular depth estimation for indoor scenes".

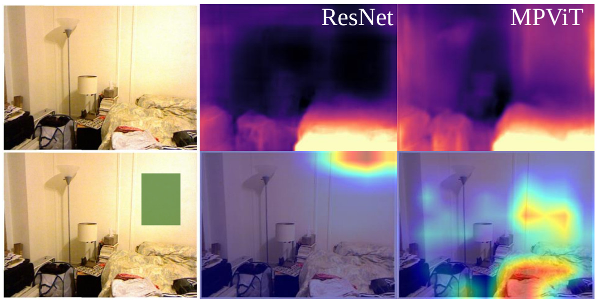

Figure 1. **Attention maps by different encoders.** On top: RGB image and depth map by models using CNN-based (ResNet) and transformer-based (MPViT) encoders. At bottom: untextured region highlighted on RGB image and attention maps by ResNet and MPViT based models.

1. Ablation studies

For the sake of space, we complement here the ablation studies reported in the main paper.

Encoder: Benefiting from the global feature extraction of the transformer, the model can infer the depth of low-texture regions more accurately text to global attention. Fig. 1 visually shows an intuition of this mechanism: on left, we show an RGB image with a large, untextured wall, highlighted in green on the second row. A model with a ResNet encoder cannot estimate an accurate depth map (second column), as the attention of the last features is mainly focused on the low-texture area on top of the image. Conversely, the output of a model using MPVit as a backbone results coherent, thanks to the global guidance it can recover from the bottom regions of the image showing richer texture. Furthermore, we report an ablation study in Tab. 1 comparing the results yielded by MPViT variants [5] (tiny, xsmall, small, base). We identify a good trade-off between accuracy and complexity in "MPViT-small", which we use as the encoder backbone for all our experiments in the main paper.

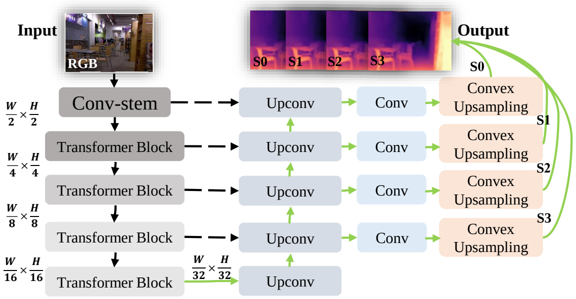

Figure 2. **Architecture of Depth Network.** The network takes a single RGB image as input and outputs four full-resolution disparity maps [S0, S1, S2, S3].

Low texture regions: To further verify the effectiveness of different encoders on low-texture regions, we evaluate the depth estimation performance of models on low-texture regions. We use image gradient information to select and locate low-texture regions, as shown in Fig.4. As shown in Table 1 (V-X), with the global modeling ability of the transformer, the MPViT encoder shows outstanding performance on low-texture regions than CNN encoders.

Decoder: For the depth decoder, we follow the architecture design proposed by Monodepth2 [3], yet we implement convex upsampling as proposed by RAFT [8], to output four full resolution disparity maps, i.e. depth maps [S0, S1, S2, S3] in Fig. 2. We compare the accuracy yielded by the standard decoders proposed by Monodepth2 [3] and ours in Table 1, rows (E-H). We can notice that our decoder allows for improving the results by both Monodepth2 itself and our network architecture.

ISD: As shown in Table 1 (K-R), with the help of the iterative self-distillation method we proposed in this paper, depth maps are accurate at any scale. Note that the depth map **S3** is as accurate as the depth map **S0**, which means that the final features extracted by the encoder include the

	Details	**Params**	lower is better				higher is better		
			Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log ↓	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$
			Different ResNet and MPViT [5] variants						
(A)	ResNet–18	11.6M	0.141	0.107	0.533	0.183	0.812	0.957	0.989
(B)	ResNet–34	21.8M	0.133	0.100	0.517	0.174	0.828	0.963	0.991
(C)	MPViT-**tiny**	5.8M	0.122	0.087	0.478	0.162	0.854	0.969	0.992
(D)	MPViT-**xsmall**	10.5M	0.117	0.085	0.472	0.157	0.863	0.972	0.992
(E)	MPViT-**small**	22.8M	**0.113**	**0.083**	**0.459**	**0.153**	0.871	**0.973**	0.992
(F)	MPViT-**base**	74.8M	**0.113**	0.085	0.466	**0.153**	**0.872**	**0.973**	**0.993**
	Trainig framework	**Decoder**	**Different decoders**						
(G)	Monodepth2 [3]	Monodepth2 [3]	0.167	0.137	0.603	0.208	0.754	0.944	0.985
(H)	Monodepth2 [3]	ours	0.161	0.130	0.586	0.202	0.767	0.949	0.988
(I)	Ours	Monodepth2 [3]	0.114	0.086	0.465	0.154	0.869	0.972	0.992
(J)	Ours	ours	0.113	0.083	0.459	0.153	0.871	0.973	0.992
	Details	**Scale**	**Disparity maps from different scale**						
(K)	Ours	S0	0.113	0.083	0.459	0.153	0.871	0.973	0.992
(L)	Ours(w/o **ISD**)	S0	0.114	0.085	0.469	0.156	0.867	0.972	0.992
(M)	Ours	S1	0.114	0.082	0.459	0.153	0.870	0.973	0.993
(N)	Ours(w/o **ISD**)	S1	0.139	0.099	0.522	0.177	0.822	0.968	0.992
(O)	Ours	S2	0.114	0.082	0.458	0.153	0.869	0.973	0.993
(P)	Ours(w/o **ISD**)	S2	0.175	0.299	0.918	0.231	0.791	0.947	0.980
(Q)	Ours	S3	0.114	0.082	0.460	0.153	0.868	0.973	0.993
(R)	Ours(w/o **ISD**)	S3	0.203	0.534	1.122	0.272	0.787	0.943	0.972
	Details	**n | training time**	**Results under different number of iterations**						
(S)	Ours	**1** | 26.2h	0.115	0.084	0.467	0.155	0.867	0.973	0.992
(T)	Ours	**2** | 44.3 h	0.113	0.083	**0.459**	0.153	0.871	**0.973**	**0.992**
(U)	Ours	**3** | 64.1h	**0.112**	**0.082**	0.460	**0.152**	**0.873**	**0.973**	**0.992**
	Details	**test regions**	**Results of different encoders on low texture regions**						
(V)	R18 (GasMono training framework)	Low Texture	0.140	0.106	0.527	0.181	0.816	0.958	0.990
(W)	R34 (GasMono training framework)		0.131	0.099	0.512	0.172	0.834	0.963	0.991
(X)	MPViT-small (GasMono training framework)		0.110	0.082	0.453	0.149	0.877	0.974	0.993

Table 1. **Ablation Studies on NYUv2**. We test the depth encoder based on different MPViT variants and different depth decoders on NYUv2 datasets. The effects of **ISD** on the accuracy of each scale depth map are also evaluated. We also test the performance of different encoders on low-texture regions.

	Depth Encoder	**Pose Estimation**	lower is better				higher is better		
			Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log ↓	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$
(i1)	ResNet–18	Monoindoor++ [6]	0.149	0.116	0.551	0.190	0.800	0.953	0.988
(i2)	ResNet–18	GasMono(ours)	0.141	0.107	0.533	0.183	0.812	0.957	0.989
(ii1)	MPViT-**small**	Monoindoor++ [6]	0.124	0.092	0.488	0.163	0.851	0.970	0.992
(ii2)	MPViT-**small**	GasMono(ours)	0.113	0.083	0.459	0.153	0.871	0.973	0.992

Table 2. **Ablation Studies of different pose estimation methods on NYUv2 dataset**. To compare with the SoTA method – monoindoor++, we replace the pose estimation method proposed in GasMono with that proposed in monoindoor++. Except for the pose estimation part, the settings of the training process are all the same.

key information for depth inferring, and with the help of skip connection, the depth is optimized as information increases. Besides, we also do an ablation study on the number of iterations **n**, as shown in Table 1 (S-U). When $\mathbf{n} = 1$, the self-distillation method cannot improve the accuracy of the depth labels already obtained by the framework itself, having a negative impact on the performance. When using $\mathbf{n} > 1$, ISD allows for getting better pseudo labels, thereby yielding a much better depth model after training. In the main paper, we select $\mathbf{n} = 2$, since further increases do not yield substantial improvement, yet increasing notably the total training time.

2. Coarse Pose analysis

We now further inquire about the use of coarse poses obtained through COLMAP.

COLMAP for coarse poses: COLMAP is executed offline, so it does not affect training/inference time directly. It only delays the start of the training (on NYUv2, a single sequence takes $\sim$ 10 min, with sequences being processed in parallel), a delay we can tolerate given the SoTA accuracy it yields. Since some sequences in NYUv2 are very short, and some scenes (like the bathroom) expose many low-texture regions, COLMAP may fail on some of these sequences, which are excluded from the training set training.

One of the most important and valuable points in this paper is: the scale ambiguity observed across the hundred se-

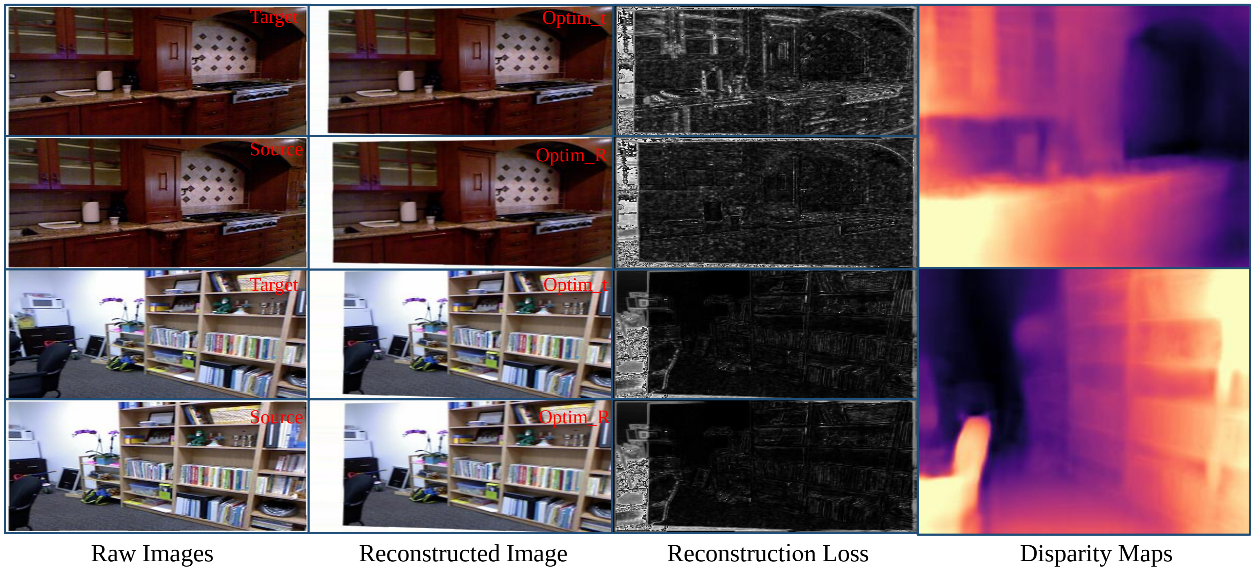

Figure 3. **Visualization of the coarse rotation optimization.** The two samples are based on inaccurate and accurate coarse rotations, and both the rotations between the source and target images are large, resulting in large unreconstructable regions. With the accurate coarse rotation, the coarse pose with translation optimization ($Optim\_t$) can get a satisfactory reconstruction, see 2nd sample. While the coarse rotation is inaccurate, we need to improve the rotation further, otherwise, it will introduce noise into training, see the reconstruction loss of 1st sample. Besides, as shown in the reconstruction loss, the rotation from the coarse pose can eliminate most rotation components, which makes it possible to directly estimate the residual rotation using a pose network.

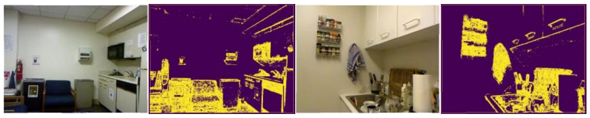

Figure 4. Low texture regions for depth evaluation.

quences used for training makes COLMAP use not as simple/effective as expected, avoiding network convergence – see Tab.1 (a-g) in the paper. Indeed, using COLMAP poses alone is, counter-intuitively, harmful to indoor SMDE.

Translation rescaling and refinement: By its nature, our AlignNet learns to regress a scale value that is meaningful for training images only – i.e., it aligns the scale of the translation component of COLMAP poses on the training set, to allow for DepthNet convergence (Tab. 1). As such, it is a training optimization tool only, and it would be ineffective on test images. Therefore, considering the tightly coupled relationship between the depth estimation part and the pose estimation part, we use the performance of the depth network to verify the effectiveness of pose optimization – see Tab.1 (a-g) in the paper. After **translation rescaling and refinement**, our GasMono has achieved SoTA performance – see Tab.1 (n) in the paper.

Rotation optimization: The above part only focuses on translation optimization, though the rotation of coarse poses can also be inaccurate and noisy, so rotation optimization is proposed. To further show the effect of rotation optimization within the training, in Fig. 3, we report two samples with inaccurate coarse rotation (top sample) and accurate coarse rotation (bottom sample) respectively. For both, we compute the reconstruction losses based on “Optim_t” and “Optim_R” and show them in column 3, based on estimated depth (column 4). For the first sample, because of the inaccurate coarse rotation, optimizing only for translation (“Optim_t”, row 1) cannot compensate for the wrong rotation and thus yield a high reprojection error. After refining the rotation, the reconstruction based on “Optim_R” (row 2) results in a much lower photometric error. On the contrary, in case of accurate coarse pose, shown as the second sample in the figure, the reconstruction based on “Optim_t” can already achieve a reasonable reconstructed image.

Comparison with different pose estimation strategies: To further prove the effectiveness of the pose estimation part in our GasMono, we compare it with the SoTA self-supervised indoor framework – monoindoor++ [6]. Since the SoTA Monoindoor series [4, 6] do not open source their code, we cannot construct experiments on the Monoindoor framework. Therefore, based on our GasMono, we replace the pose estimation method with the one proposed in monoindoor++, which is purely the learning-based pose estimation method. Except for the pose estimation part, the settings of the training process and loss functions are all the same. Experimental results in Tab. 2 show that our proposed method yields more accurate results, which means that the proposed geometric-assisted

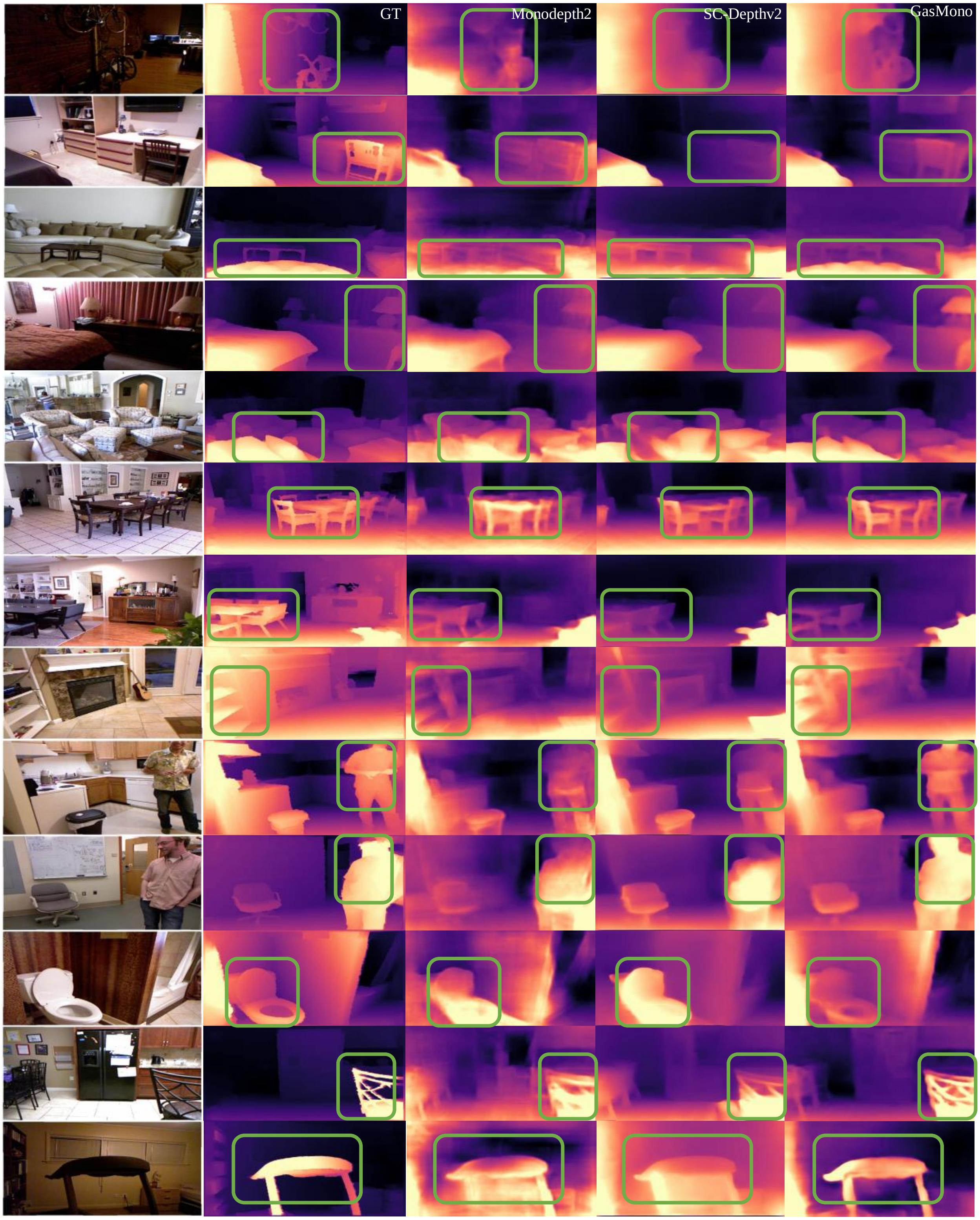

Figure 5. **Visualization on NYUv2.** Our **GasMono** achieved finer-grained depth estimation than the baseline Monodepth2 [3] and recent work SC-Depthv2 [1].

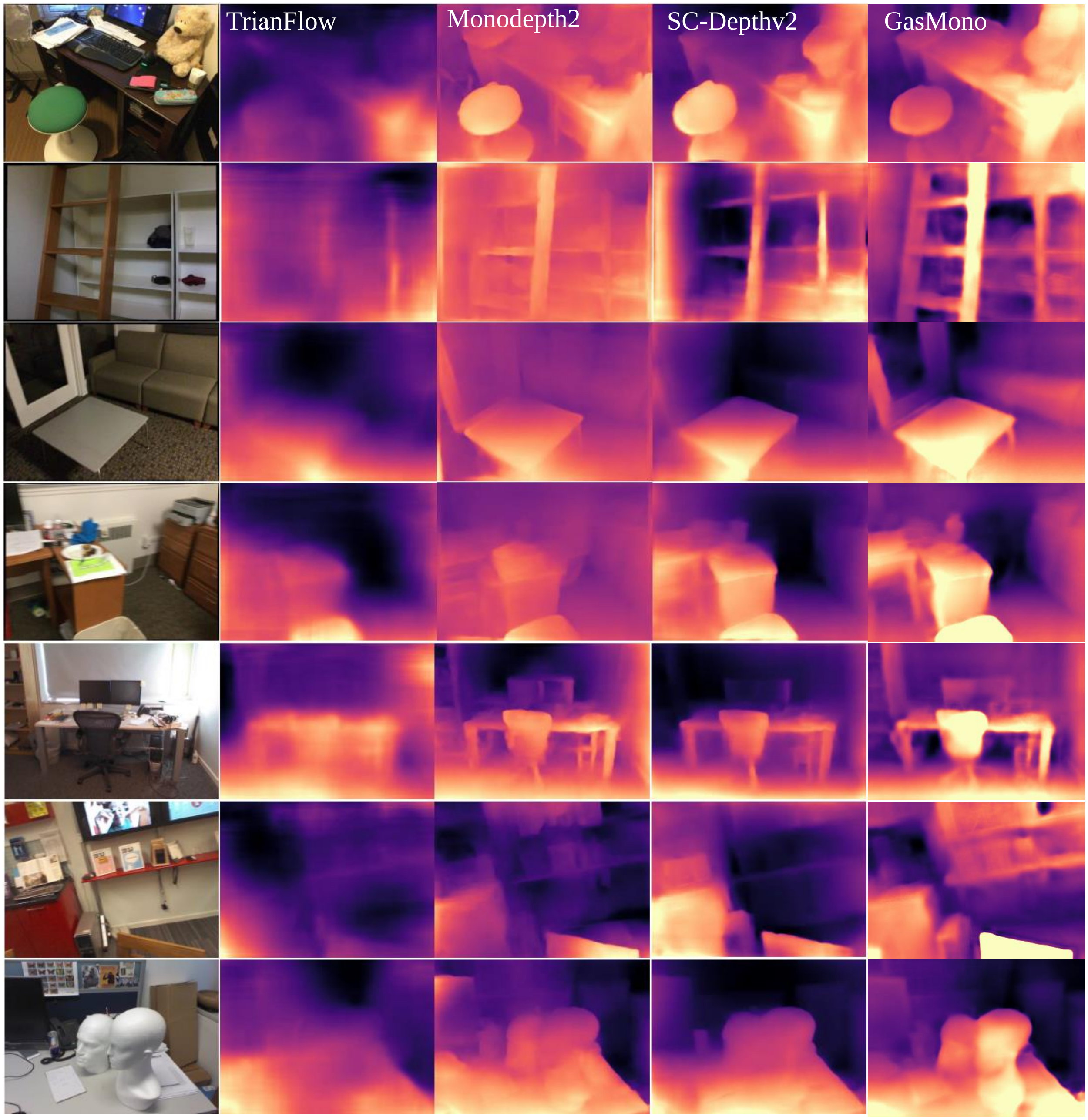

Figure 6. **Visualization on ScanNet [2] and 7scenes [7].** Generalization comparison with the Trianflow [9],Monodepth2 [3] and SC-Depthv2 [1]. GasMono shows more accurate and finer-grained depth estimation on new scenes.

training framework as well as the proposed coarse pose optimization method are more practical, effective, and beneficial for the indoor self-supervised monocular depth estimation task.

3. Qualitative results

We report additional qualitative results to show a comparison with existing self-supervised frameworks.

Comparison with other methods – in-domain: We show some qualitative results in Fig. 5, comparing some existing methods with our GasMono framework. Ours outperforms the baseline Monodepth2 [3] and a more recent work, i.e. SC-Depthv2 [1], on fine-grained depth estimation under different conditions. GasMono shows outstanding performance when estimating depth at object details, such as the bicycle (1st line), chair (2nd and 6-7th lines), desk lamp (4th line), and so on. Besides, for large-scale objects, like the sofa (3rd line), chair (12th), and treadmill (13th), and dynamic objects, like people (9-10th lines), our model shows very accurate results as well.

Comparison with other methods – generalization: We also show more qualitative examples for generalization to unseen datasets in Fig. 6. Compared with recent works [1, 3, 9], GasMono shows higher-quality results in estimating the depth for unseen objects.

4. Limitations

Since our training framework builds on self-supervised losses, GasMono still cannot perform well in regions where reprojection constraints are violated, e.g. specular/transparent surfaces or surfaces with reflected light. As shown in Fig. 7, since those regions do not satisfy the photometric consistency assumption, GasMono fails on highly-specular or transparent surfaces.

Moreover, running COLMAP to obtain the coarse poses – required to start the training – is time-consuming, in particular on datasets counting hundred of sequences made of thousand frames each.

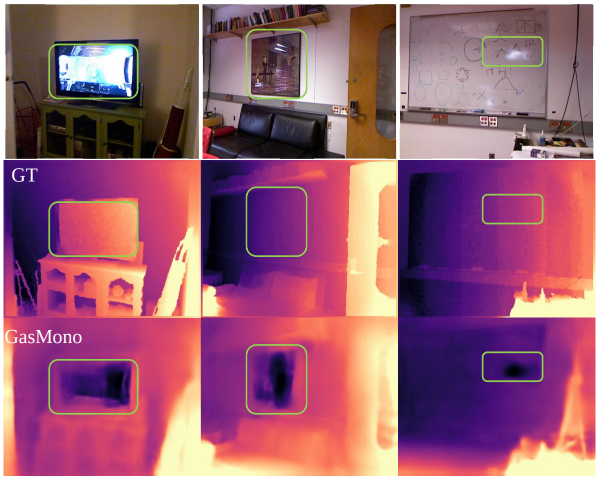

Figure 7. **Limitations of GasMono.** Since GasMono is trained in a self-supervised manner, it cannot estimate a reasonable depth for those regions with photometric changes.

References

[1] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Tat-Jin Chin, Chunhua Shen, and Ian Reid. Auto-rectify network for unsupervised indoor depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 4, 5, 6

[2] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 5

[3] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 4, 5, 6

[4] Pan Ji, Runze Li, Bir Bhanu, and Yi Xu. Monoindoor: Towards good practice of self-supervised monocular depth estimation for indoor environments. In *ICCV*, 2021. 3

[5] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7287–7296, 2022. 1, 2

[6] Runze Li, Pan Ji, Yi Xu, and Bir Bhanu. Monoindoor++: Towards better practice of self-supervised monocular depth estimation for indoor environments. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 2, 3

[7] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013. 5

[8] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 1

[9] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. Towards better generalization: Joint depth-pose learning without posenet. In *CVPR*, 2020. 5, 6