3D Distillation:
Improving Self-Supervised Monocular Depth Estimation on Reflective Surfaces
— Supplementary Material —

1. Metric Definitions

We present the definitions of the evaluation metrics in Tab. 1.

2. Relationship with Model Distillation

Model distillation [3] can use the knowledge from stronger and heavier networks to improve the prediction accuracy of weaker and smaller networks. In contrast, our proposed 3D distillation utilizes the multi-view 3D information aggregated from the predicted depth of multiple video frames to improve the depth prediction accuracy on reflective surfaces.

In Tab. 2, we show the results of different distillation methods on the ScanNet val and test sets [1]. We can observe that: (i) Both 3D distillation and model distillation can improve the overall depth prediction accuracy; (ii) Combining 3D distillation and model distillation can achieve the best accuracy, which supports that 3D distillation and model distillation are complementary.

In Tab. 3, we show the results of different distillation methods on the ScanNet-Reflection val and test sets [1]. We can observe that 3D distillation can significantly improve the accuracy on reflective surfaces but model distillation can not. This supports that our 3D distillation can better improve the accuracy on reflective surfaces.

3. Mesh Results

To show the superiority of our 3D distillation under 3D mesh metrics, we use TSDF-fusion [5] to reconstruct the scenes in the ScanNet val and test set [1] and evaluate the meshes [7]. In the TSDF-fusion [5] for this evaluation, the voxel size is 0.05m and the truncation distance is 0.2m, and we only integrate every 10^{th} frame during fusion to speed up the reconstruction. We evaluate using the best network architecture, *i.e.*, MonoViT architecture [8]. The results are shown in Tab. 4. We can observe: (i) On the val set, 3D distillation model achieves the best result for Acc, Prec, and F-score, and achieves the second best result for Recal; (ii) On the test set, 3D distillation model achieves the best result for Acc, Prec, and F-score, and achieves the second best result for Comp and Recal.

4. Sensitivity of the Thresholds

Tab. 5a shows the results of 3D distillation models using different truncation distance in TSDF-fusion [5] during 3D distillation training. We can see ‘1.0’ is slightly better than ‘0.4’ and ‘0.2’. Reconstructed meshes with larger truncation distance can be more complete, resulting in better projected depth for reflective surfaces.

Tab. 5b shows the results of 3D distillation models using different uncertainty thresholds. We can see ‘0.4’ is the best. If the threshold is too large, the recall rate of reflective surfaces during the uncertainty-guided depth fusion will be too low; while if the threshold is too small, there will be many false positives of reflective surfaces during the uncertainty-guided depth fusion.

5. Predicted Depth Screening

In Tab. 6, we conduct an evaluation on screening the predicted depth used to reconstruct meshes. We observe that depth screening can lead to more accurate but sparser projected depth (Coverage% on uncertain regions: $89 \rightarrow 82$). Introducing depth screening is a variant of our 3D distillation training, which may further improve the accuracy of our 3D distillation model.

Depth Metric		Mesh Metric	
Abs Rel	$\frac{1}{n}\sum \lvert d-d^*\rvert / d^*$	Acc	$\text{mean}_{p\in P}(\min_{p^*\in P^*} \lvert\lvert p-p^*\rvert\rvert)$
Sq Rel	$\frac{1}{n}\sum \lvert d-d^*\rvert^2 / d^*$	Comp	$\text{mean}_{p^*\in P^*}(\min_{p\in P} \lvert\lvert p-p^*\rvert\rvert)$
RMSE	$\sqrt{\frac{1}{n}\sum \lvert d-d^*\rvert^2}$	Prec	$\text{mean}_{p\in P}(\min_{p^*\in P^*} \lvert\lvert p-p^*\rvert\rvert < .05)$
RMSE log	$\sqrt{\frac{1}{n}\sum \lvert \log d-\log d^*\rvert^2}$	Recal	$\text{mean}_{p^*\in P^*}(\min_{p\in P} \lvert\lvert p-p^*\rvert\rvert < .05)$
$\delta < 1.25^i$	$\frac{1}{n}\sum (\max(\frac{d}{d^*}, \frac{d^*}{d}) < 1.25^i)$	F-score	$\frac{2\times\text{Prec}\times\text{Recal}}{\text{Prec}+\text{Recal}}$

Table 1: Definitions of metrics: n is the number of pixels with both valid predictions and ground truth; d and d^* are the predicted and ground truth depth, respectively; p and p^* are the predicted and ground truth point clouds, respectively.

Self-Supervised	Student Network	Training Label	Distillation	ScanNet Val Set						
				Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Monodepth2 [2]	Monodepth2 [2]	pred. depth	None	0.160	0.090	0.365	0.193	0.780	0.941	0.983
Monodepth2 [2]		pred. depth + proj. depth	3D	0.157	0.083	0.357	0.190	0.782	0.943	**0.985**
MonoViT [8]		pred. depth	Model	0.159	0.092	0.361	0.192	0.788	0.941	0.983
MonoViT [8]		pred. depth + proj. depth	3D + Model	**0.153**	**0.080**	**0.346**	**0.185**	**0.791**	**0.945**	**0.985**
HR-Depth [4]	HR-Depth [4]	pred. depth	None	0.159	0.090	0.360	0.190	0.785	0.943	0.984
HR-Depth [4]		pred. depth + proj. depth	3D	0.154	0.080	0.349	0.186	0.788	0.945	0.986
MonoViT [8]		pred. depth	Model	0.155	0.088	0.354	0.187	0.796	0.944	0.984
MonoViT [8]		pred. depth + proj. depth	3D + Model	**0.149**	**0.075**	**0.335**	**0.180**	**0.801**	**0.949**	**0.987**
Self-Supervised	Student Network	Training Label	Distillation	ScanNet Test Set						
				Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Monodepth2 [2]	Monodepth2 [2]	pred. depth	None	0.184	0.109	0.392	0.210	0.742	0.925	0.976
Monodepth2 [2]		pred. depth + proj. depth	3D	0.181	0.105	0.388	0.208	0.746	0.927	0.976
MonoViT [8]		pred. depth	Model	0.181	0.105	0.382	0.207	0.752	0.928	0.976
MonoViT [8]		pred. depth + proj. depth	3D + Model	**0.178**	**0.101**	**0.378**	**0.205**	**0.754**	**0.929**	**0.977**
HR-Depth [4]	HR-Depth [4]	pred. depth	None	0.178	0.102	0.381	0.204	0.752	0.931	**0.979**
HR-Depth [4]		pred. depth + proj. depth	3D	0.176	0.098	0.378	0.202	0.754	0.932	**0.979**
MonoViT [8]		pred. depth	Model	0.175	0.099	0.372	0.201	0.763	0.933	0.978
MonoViT [8]		pred. depth + proj. depth	3D + Model	**0.172**	**0.095**	**0.367**	**0.198**	**0.766**	**0.934**	**0.979**

Table 2: Results of different distillation methods on the ScanNet val and test sets [1]. We can observe that both 3D distillation and model distillation [3] can improve the overall depth accuracy, and combining 3D distillation and model distillation [3] can achieve the best accuracy. **Bold** and underline indicate the best and second results of a student network, respectively.

Self-Supervised	Student Network	Training Label	Distillation	ScanNet-Reflection Val Set						
				Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Monodepth2 [2]	Monodepth2 [2]	pred. depth	None	0.192	0.188	0.548	0.233	0.764	0.920	0.967
Monodepth2 [2]		pred. depth + proj. depth	3D	0.156	0.093	0.442	0.191	0.786	0.943	**0.987**
MonoViT [8]		pred. depth	Model	0.196	0.206	0.561	0.241	0.778	0.917	0.961
MonoViT [8]		pred. depth + proj. depth	3D + Model	**0.147**	**0.086**	**0.416**	**0.182**	**0.808**	**0.950**	**0.987**
HR-Depth [4]	HR-Depth [4]	pred. depth	None	0.202	0.208	0.565	0.243	0.756	0.914	0.964
HR-Depth [4]		pred. depth + proj. depth	3D	0.153	0.090	0.430	0.188	0.789	0.948	**0.989**
MonoViT [8]		pred. depth	Model	0.194	0.207	0.562	0.239	0.780	0.919	0.961
MonoViT [8]		pred. depth + proj. depth	3D + Model	**0.145**	**0.083**	**0.407**	**0.179**	**0.807**	**0.953**	**0.989**
Self-Supervised	Student Network	Training Label	Distillation	ScanNet-Reflection Test Set						
				Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Monodepth2 [2]	Monodepth2 [2]	pred. depth	None	0.179	0.146	0.502	0.218	0.750	0.938	0.980
Monodepth2 [2]		pred. depth + proj. depth	3D	0.156	0.096	0.459	0.195	0.766	0.945	0.988
MonoViT [8]		pred. depth	Model	0.175	0.146	0.490	0.216	0.771	0.935	0.976
MonoViT [8]		pred. depth + proj. depth	3D + Model	**0.155**	**0.092**	**0.435**	**0.190**	**0.778**	**0.948**	**0.990**
HR-Depth [4]	HR-Depth [4]	pred. depth	None	0.175	0.145	0.492	0.215	0.757	0.936	0.982
HR-Depth [4]		pred. depth + proj. depth	3D	**0.152**	**0.089**	0.451	0.190	0.771	**0.956**	**0.990**
MonoViT [8]		pred. depth	Model	0.175	0.148	0.485	0.215	0.771	0.940	0.978
MonoViT [8]		pred. depth + proj. depth	3D + Model	**0.152**	0.090	0.**435**	**0.188**	**0.789**	0.950	0.988

Table 3: Results of different distillation methods on the ScanNet-Reflection val and test sets [1]. 3D distillation can significantly improve the accuracy on reflective surfaces, but model distillation [3] can not.

Architecture	Model	ScanNet Val Set				
		Acc ↓	Comp ↓	Prec ↑	Recal ↑	F-score ↑
MonoViT [8]	Self-Supervised [2]	0.214	**0.096**	0.253	**0.403**	0.307
	Self-Teaching [6]	0.203	0.098	0.270	0.396	0.317
	3D Distillation (ours)	**0.186**	0.099	**0.284**	0.397	**0.328**
Architecture	**Model**	**ScanNet Test Set**				
		Acc ↓	Comp ↓	Prec ↑	Recal ↑	F-score ↑
MonoViT [8]	Self-Supervised [2]	0.256	**0.126**	0.203	**0.329**	0.246
	Self-Teaching [6]	0.249	0.130	0.214	0.324	0.253
	3D Distillation (ours)	**0.235**	0.128	**0.224**	0.326	**0.260**

Table 4: Mesh results on the ScanNet val and test set [1]. 'Self-Supervised' indicates that the model is trained with the photometric loss [2]. 'Self-Teaching' indicates that the model is supervised by the predicted depth from self-supervised models and trained with the depth loss in the main paper. '3D Distillation' indicates that the model is supervised by the fusion of the predicted depth and project depth and trained with the depth loss in the main paper. **Bold** and underline indicate the best and second results of an architecture, respectively.

Truncation Distance	ScanNet Val Set						
	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
0.2	0.159	0.085	0.358	0.191	0.780	0.942	0.984
0.4	0.158	**0.083**	**0.357**	**0.190**	**0.782**	**0.943**	**0.985**
1.0 (ours)	**0.157**	**0.083**	**0.357**	**0.190**	**0.782**	**0.943**	**0.985**

(a) Experiments with different truncation distance in TSDF-fusion [5].

Uncertainty Threshold	ScanNet Val Set						
	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
0.2	**0.156**	**0.083**	0.364	0.191	**0.782**	0.942	0.984
0.4 (ours)	0.157	**0.083**	**0.357**	**0.190**	**0.782**	**0.943**	**0.985**
0.6	0.158	0.085	0.358	0.191	0.781	0.942	0.984
0.8	0.159	0.087	0.361	0.192	0.780	0.942	0.984

(b) Experiments with different uncertainty thresholds, *i.e.*, α_{uncer}.

Table 5: Experiments on the ScanNet val set [1] using the network architecture of Monodepth2 [2].

Depth Source	ScanNet-Reflection Val Set							
	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑	Coverage% ↑
Self-Supervised Model	0.206	0.227	0.584	0.246	0.750	0.912	0.961	**100**
proj. depth w/o uncer. filt.	0.197	0.172	0.595	0.306	0.677	0.863	0.940	89
proj. depth w uncer. filt.	0.186	0.150	0.548	0.288	0.698	0.875	0.949	82
3D Distillation Model	**0.156**	**0.093**	**0.442**	**0.191**	**0.786**	**0.943**	**0.987**	**100**

Table 6: Projected depth quality with and without uncertainty-driven filtering, using Monodepth2 architecture [2].

References

[1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 1, 3, 4

[2] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 3, 4

[3] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NeurIPSW*, 2014. 1, 3

[4] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. In *AAAI*, 2021. 3

[5] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew W. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, 2011. 1, 4

[6] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *CVPR*, 2020. 4

[7] Mohamed Sayed, John Gibson, Jamie Watson, Victor Prisacariu, Michael Firman, and Clément Godard. Simplerecon: 3d reconstruction without 3d convolutions. In *ECCV*, 2022. 1

[8] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *3DV*, 2022. 1, 3, 4