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

Xuepeng Shi[2] Georgi Dikov[1] Gerhard Reitmayr[1] Tae-Kyun Kim[2,3] Mohsen Ghafoorian[1]
[1]Qualcomm [2]Imperial College London [3]KAIST

Abstract

Self-supervised monocular depth estimation (SSMDE) aims at predicting the dense depth maps of monocular images, by learning to minimize a photometric loss using spatially neighboring image pairs during training. While SSMDE offers a significant scalability advantage over supervised approaches, it performs poorly on reflective surfaces as the photometric constancy assumption of the photometric loss is violated. We note that the appearance of reflective surfaces is view-dependent and often there are views of such surfaces in the training data that are not contaminated by strong specular reflections. Thus, reflective surfaces can be accurately reconstructed by aggregating the predicted depth of these views. Motivated by this observation, we propose 3D distillation: a novel training framework that utilizes the projected depth of reconstructed reflective surfaces to generate reasonably accurate depth pseudo-labels. To identify those surfaces automatically, we employ an uncertainty-guided depth fusion method, combining the smoother and more accurate projected depth on reflective surfaces and the detailed predicted depth elsewhere. In our experiments using the ScanNet and 7-Scenes datasets, we show that 3D distillation not only significantly improves the prediction accuracy, especially on the problematic surfaces, but also that it generalizes well over various underlying network architectures and to new datasets.

1. Introduction

Monocular depth estimation [37, 7] is the task of predicting the dense depth map of a monocular image. It is a fundamental and challenging problem in computer vision as it bridges the gap between 2D images and the 3D world. Supervised monocular depth estimation requires a large number of images from diverse scenes with ground truth depth. However, creating depth annotations involves expensive hardware and is time-consuming [12, 4, 39]. In contrast, self-supervised monocular depth estimation (SSMDE) [11, 55, 13, 14] only requires posed images as training data, such as stereo pairs and video sequences, and is therefore important for domains such as autonomous driving and virtual/augmented reality where the scalability of the data acquisition for various environments and camera setups matters. As a consequence, SSMDE has drawn much attention in recent years [43, 42].

Fundamentally, training an SSMDE model is based on the photometric loss [14]: given (i) the relative pose between two frames (source and target), (ii) the camera intrinsic parameters and (iii) the predicted depth map of the target frame, one can transform the source image into

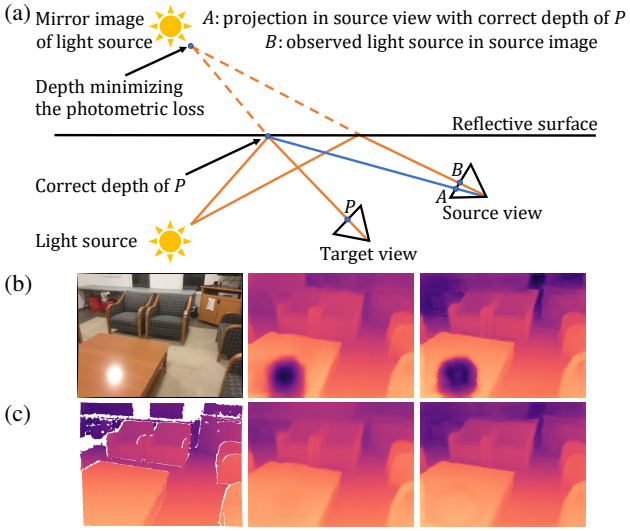
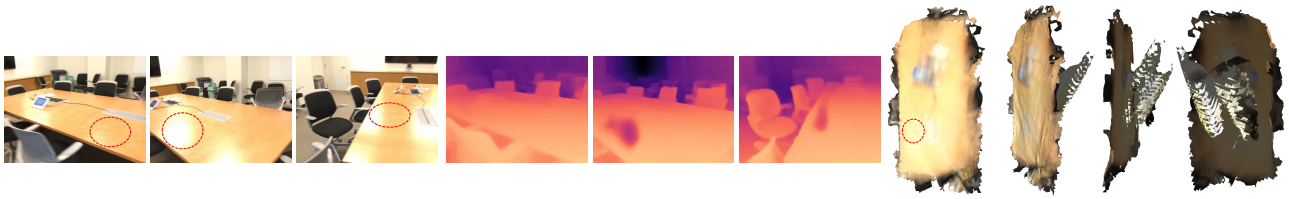


Figure 1: (a) On a reflective surface, predicting the correct surface depth does not minimize the photometric loss [14], due to the disparity between the projection with correct depth A and the observed location B. (b) L→R: Image from the ScanNet test set (scene0781_00) [4], predicted depth of Monodepth2 [14] and MonoViT [53] which overestimates the depth of the highlight. (c) L→R: Ground truth, predicted depth of [14, 53] with our 3D distillation.

Email: XuepengShi@outlook.com. X. Shi did the work while interning at Qualcomm.



(a) (b) (c)

Figure 2: (a) Images from the ScanNet train set [4] with annotated highlights. (b) Depth predictions of Monodepth2 [14]. (c) Mesh [32] of the table showing that the reflective surface can still be reconstructed correctly by aggregating the predicted depth from different view directions. The artifacts from the overestimated depths are occluded by the correct mesh surface.

the target view direction with interpolation-based warping. The photometric loss can then be used to guide the training of the underlying depth estimation model. The effectiveness of SSMDE in outdoor applications such as autonomous driving [12] has been demonstrated in many prior works [11, 55, 13, 34, 14, 15, 16, 22, 45].

On the other hand, applying SSMDE to indoor scenes [4, 39] is challenging due to commonly observed reflective surfaces such as shiny floors, tables, screens, *etc*. As illustrated in Fig. 1a, predicting the correct depth of a reflective surface does not minimize the photometric loss due to view-dependent effects violating the photometric constancy assumption. Specifically, perceiving depth at the mirror image of a light source on the reflective surface, appearing as a virtual faraway object, minimizes the said loss. Consequently, the network learns to predict overestimated depth for the specular reflection, see Fig. 1b. However, this issue has not been studied enough.

To address this issue, we propose 3D distillation: a general training framework to improve SSMDE on reflective surfaces. As shown in Fig. 2a and Fig. 2b, we observe that specular highlights are view-dependent, and that there are some view directions in which the surface appearance is not contaminated by them. Thus, reflective surfaces can be accurately reconstructed by aggregating the predicted depth of these views, as shown in Fig. 2c. Inspired by this observation, we utilize the projected depth of reconstructed scenes to generate accurate depth pseudo-labels for challenging reflective surfaces. However, while the projected depth is more accurate at reflective surfaces, it is lacking high-frequency details due to volumetric averaging over multiple views. To overcome this over-smoothing problem, we propose a fusion scheme in which the projected and predicted depth are combined under the guidance of an uncertainty map associated to the predicted depth. Our 3D distillation is agnostic to the underlying network architectures [14, 29, 53] and significantly improves the depth prediction accuracy on reflective surfaces, as shown in Fig. 1c.

We highlight the contributions of this paper as follows:

1. We propose 3D distillation: a novel training framework that utilizes multi-view 3D information to improve depth prediction accuracy on reflective surfaces without adding computational cost or model parameters during inference.

2. We originally fuse the predicted and projected depth for pseudo-label generation, and propose an uncertainty-based approach that accurately identifies specular highlights.

3. To validate the effectiveness, we select a subset of the ScanNet dataset [4] which is rich in specular reflections and glossy surfaces and thus provide a foundation for benchmarking future works tackling this issue.

4. Through extensive evaluations, we show that 3D distillation significantly improves the depth accuracy of reflective surfaces on ScanNet [4] and 7-Scenes [39], while being agnostic to the underlying networks.

2. Related Work

2.1. Self-Supervised Monocular Depth Estimation

Self-supervised monocular depth estimation (SSMDE) aims to learn the dense depth maps of monocular images, training with the photometric loss [14] using stereo pairs or monocular videos. Monodepth [13] learns depth from stereo pairs. Monodepth2 [14] further uses temporally neighboring frames to minimize the photometric loss, and introduces auto-masking and minimum reprojection loss to solve the problem of stationary pixels and occlusions. To deal with dynamic objects, semantic information is utilized in SGDepth [22] and motion maps are introduced in [26]. Feature space reconstruction losses are used in [52, 40] to improve the depth accuracy. DeFeatNet [41] introduces a cross-domain dense feature representation and a warped feature consistency to improve the depth accuracy. In [34], a complex architecture is deployed to supervise a more compact one. HR-Depth [29] introduces high-resolution feature representation and feature fusion squeeze-and-excitation block. MonoViT [53]

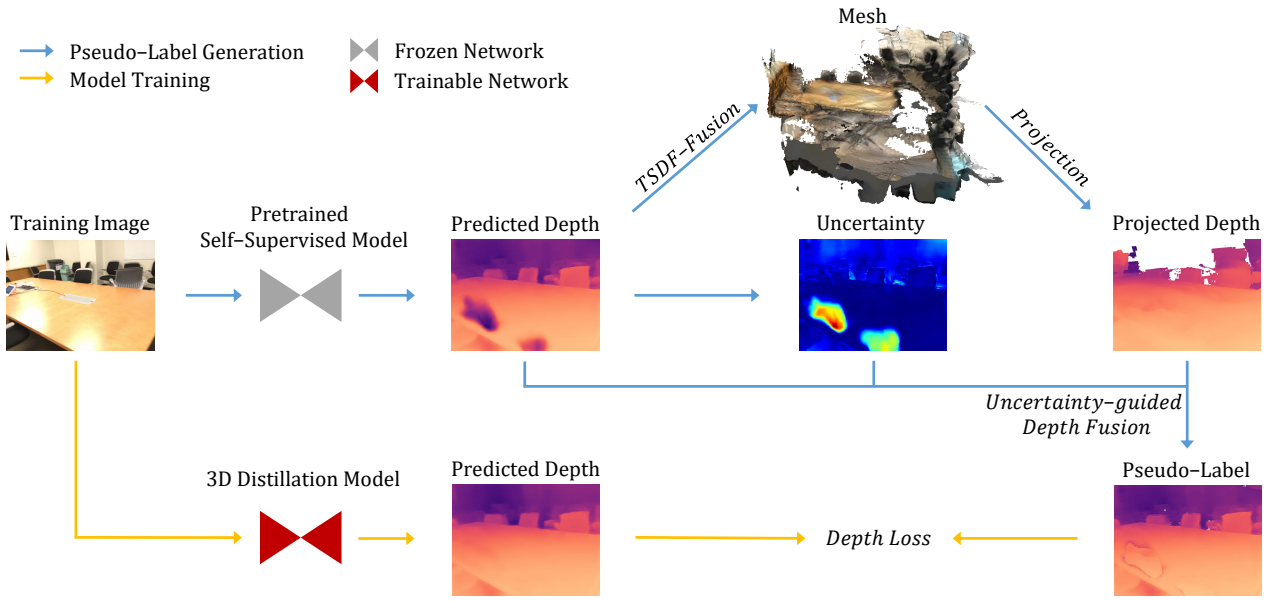


Figure 3: Pipeline of our 3D distillation training. Firstly, a pretrained self-supervised depth model is used to get the predicted depth of the training images. Then, the mesh of the training scene is reconstructed using the predicted depth, and the projected depth of the training images can be obtained. Finally, the predicted and projected depth are fused to generate pseudo-labels, under the guidance of the uncertainty of the predicted depth (low high). 3D distillation can generate accurate pseudo-labels by utilizing the multi-view 3D information aggregated from the predicted depth of multiple video frames.

uses MPViT [24] as the encoder of the depth model and achieves state-of-the-art SSMDE accuracy [42]. Planar assumptions are used in [51, 25] to improve the depth estimation accuracy for indoor scenes, which is achieved by utilizing external superpixel segmentation [9] or vanishing point detection [28] methods. MonoIndoor [19] is designed to handle indoor dynamic depth ranges and be more robust to rotational motions. DistDepth [46] uses an external depth model [36] as a teacher, which is trained in a supervised manner, to guide the training of SSMDE models.

In this paper, we improve the SSMDE accuracy on reflective surfaces in indoor scenes in a self-supervised manner, which has not been studied in these existing works. The proposed 3D distillation utilizes the multi-view 3D information aggregated from the predicted depth of multiple video frames, instead of utilizing external segmentation or depth models [51, 25, 46]. To demonstrate the generalizability, we experiment on three SSMDE architectures [14, 29, 53].

2.2. Self-Supervised Multi-View Stereo

Self-supervised multi-view stereo predicts depth from multi-view images, without using ground truth depth labels during training. Generating pseudo-labels is prevalent in this topic. U-MVS [47] uses uncertainty [10] to filter out unreliable pseudo-labels. In [48], the projected depth from reconstructed meshes is used as pseudo-labels and low-resolution training is introduced to improve the accuracy. In contrast, our 3D distillation originally fuses the predicted depth and projected depth under the guidance of uncertainty [23] to generate reliable pseudo-labels. RC-MVSNet [3] uses NeRF [31] as a teacher to improve the accuracy, and training models on an object-level dataset [1]. However, designing a general NeRF model [31] for scene-level datasets [4] is challenging [17]. In contrast, our 3D distillation can work on scene-level datasets [4] and does not rely on external models like NeRF [31].

2.3. Uncertainty Estimation

Uncertainty estimation [10, 23, 20] aims to quantify the uncertainty of predictions. Regression uncertainty [20] and MC-dropout [10] are used to select reliable pseudo-labels in semi-supervised object detection [27] and self-supervised multi-view stereo [47], respectively. In SSMDE, different strategies are explored in [35] to model uncertainty. In this paper, we work on SSMDE and use an ensemble-based uncertainty [23] to guide the fusion of depth training labels from different sources.

3. Method

In this section, we first discuss the self-supervised pretraining, then detail our 3D distillation training which aggregates multi-view 3D information to improve the depth accuracy on reflective surfaces. An overview of our 3D distillation training pipeline is shown in Fig. 3.

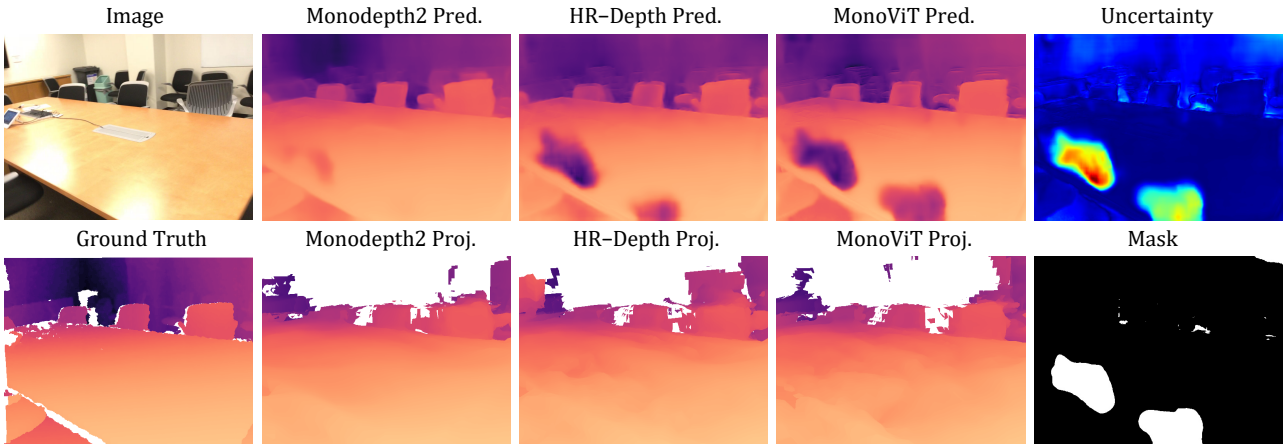


Figure 4: First row (L→R): Image from the ScanNet train set (scene0066_00) [4]; predicted depth of Monodepth2 [14], HR-Depth [29], and MonoViT [53], respectively; uncertainty map [23] of the predicted depth. Second row (L→R): Ground truth depth; projected depth of [14, 29, 53], respectively; binary mask of the uncertainty map. The predicted depth can keep more high-frequency details and the projected depth is more accurate on reflective surfaces. In our 3D distillation, the predicted depth and projected depth are fused under the guidance of uncertainty, which combines the best of the two worlds.

3.1. Self-Supervised Pretraining

In this stage, a self-supervised depth model is obtained by training with the photometric loss [13, 14]. Let $S_I = \{I_t\}_{t=1}^N$ be a sequence of video frames for training. Following common notation, we denote with $T_{t\to s}$ the relative pose for a source image I_s, with respect to a target image I_t, and with K the world-to-pixel coordinate camera projection matrix. The goal is to predict a dense depth map D_t of a target image I_t which minimizes the photometric loss $\mathcal{L}_{\text{recons}}$ as follows:

$$\mathcal{L}_{\text{recons}} = \ell(I_t, \text{warp}(I_s, T_{t\to s}, K, D_t)), \tag{1}$$

where $\text{warp}(I, T, K, D) = I\langle KTDK^{-1}x\rangle_{x\in\text{coords}(I)}$ denotes an image warping transformation with bilinear interpolation sampling. Following [13, 14], we use:

$$\ell(I_a, I_b) = \frac{\alpha}{2}(1-\text{SSIM}(I_a, I_b)) + (1-\alpha)\|I_a - I_b\|_1, \tag{2}$$

a combination of pixel-wise L_1 and SSIM [44] losses, where $\alpha = 0.85$.

In practice, we follow [14] to extend the photometric loss from Eq. (1) to account for multiple source frames using the minimum reprojection loss and add a smoothness regularization. During training, the ground truth camera poses are used to calculate the relative pose $T_{t\to s}$, and thus the predicted depth is metric. Any existing SSMDE network architecture [14, 29, 53] fits into this framework of self-supervised training.

3.2. 3D Distillation Training

In this stage, the self-supervised model is first used to generate the predicted depth of training images. Then, the meshes of training scenes are reconstructed using the predicted depth, and the projected depth of training images can be obtained. Finally, the predicted and projected depth are fused to generate pseudo-labels, and a 3D distillation model is trained using the pseudo-labels. Note that the self-supervised model is frozen in this stage.

3.2.1 Predicted Depth Generation

For the training video sequence $S_I = \{I_t\}_{t=1}^N$, the self-supervised depth model is used to obtain the predicted depth, *i.e.*, $S_D = \{D_t\}_{t=1}^N$. The predicted depth is accurate on high-frequency details, such as the boundary of an object. However, the predicted depth can perform poorly on reflective surfaces, as the photometric constancy assumption of the photometric loss in Eq. (1) is violated.

3.2.2 Projected Depth Generation

To get better training supervision for reflective surfaces, we aggregate the multi-view 3D information from the predicted depth of multiple video frames. Specifically, with the predicted depth S_D and ground truth camera poses of the training images, we use TSDF-fusion [32] to reconstruct the 3D mesh of the scene; then we project the 3D mesh according to the camera poses of the images and obtain the corresponding projected depth, *i.e.*, $S_P = \{P_t\}_{t=1}^N$. The projected depth is more accurate than the predicted depth on reflective surfaces, because reflective surfaces are view-dependent and often there are views of such surfaces in the training data that are not contaminated by strong specular reflections. Fig. 4 illustrates that the predicted and projected

	[48]	RC-MVSNet [3]	Ours
Pseudo-Label	proj. depth	Depth from NeRF [31]	pred. depth+proj. depth
Technique	LR Training	Depth-guided Sampling	Uncer-guided Fusion
Training Data	Object [1]	Object [1]	Scene [4]

Table 1: Different strategies to aggregate multi-view 3D information to generate pseudo-labels. 'Technique' means the proposed technique to improve the quality of pseudo-labels. 'LR Training' denotes low-resolution training. 'Uncer' denotes uncertainty. 'Object' and 'Scene' denote object-level datasets and scene-level datasets, respectively.

depth are complementary.

In this step, mesh reconstruction is necessary because: (i) mesh creation improves the completeness of the projected depth S_P; (ii) meshes can model occlusions.

3.2.3 Uncertainty-guided Depth Fusion

We fuse the predicted depth D_t and projected depth P_t under the guidance of the uncertainty of the predicted depth. With three self-supervised models with different network architectures [14, 29, 53], we can use an ensemble-based uncertainty [23] to obtain the uncertainty maps $S_U = \{U_t\}_{t=1}^{N}$. Specifically, the standard deviation of the three depth predictions for a pixel is the uncertainty of this pixel. As shown in Fig. 4 (top row), the ability of these networks to capture high-frequency information is different, so the depth predictions at specular highlights are varying as well, which increases the uncertainty there. We do not use MC-dropout [10] here, as MC-dropout [10] may not work well in SSMDE with known scale, as discussed in [35]. We set a threshold $\alpha_{\text{uncer}} = 0.4$ and fuse the predicted and projected depth to get the pseudo-labels $S_L = \{L_t\}_{t=1}^{N}$, formulated as:

$$L_t(x) = \begin{cases} P_t(x), & \text{if } U_t(x) \geq \alpha_{\text{uncer}} \\ D_t(x), & \text{otherwise} \end{cases} \tag{3}$$

where x is a pixel on an image frame I_t.

We compare different strategies to aggregate multi-view 3D information to generate pseudo-labels in Tab. 1. In [48], the projected depth from reconstructed meshes is used and low-resolution training/high-resolution testing is introduced to improve the accuracy. However, cross-resolution testing is challenging for monocular depth estimation [8, 18]. In [3], NeRF [31] is used as a teacher to improve the accuracy. However, designing a general NeRF model [31] for scene-level datasets [4] is not trivial [17]. Our 3D distillation originally fuses the predicted depth and projected depth to generate pseudo-labels, which works well for scene-level monocular depth estimation.

3.2.4 Model Training

We use the pseudo-labels S_L to train the 3D distillation model. Following [50, 38], the training loss is:

$$\mathcal{L}_{\text{depth}} = |\log F_t - \log L_t|, \tag{4}$$

where F_t is the prediction of image I_t and L_t is the pseudo-label of image I_t. To demonstrate the benefit of aggregating multi-view 3D information, the 3D distillation model uses the same network architecture as the self-supervised model, and is trained from scratch on the pseudo-labels instead of fine-tuning the self-supervised model. Our 3D distillation framework only modifies the training stage, without introducing additional computational cost or model parameters during inference.

3.3. Implementation Details

We experiment using three network architectures [14, 29, 53]. For numerical stability during training, the depth models predict disparity and the output is activated by a sigmoid function. The input/output resolution of the depth models is 384×288. We implement our method with PyTorch [33]. The training batch size for Monodepth2 [14] and HR-Depth [29] is 12 and for MonoViT [53] is 8. All the models are trained for 41 epochs with the Adam optimizer [21]. The initial learning rate is 10^{-4} and reduced by a factor of 10 after 26 and 36 epochs. Flipping and color augmentations are used during training, following [14]. For the scene reconstruction, we use TSDF-fusion [32] and mesh extraction in Open3D [54]. The voxel size is 0.05m and the truncation distance is 1.0m. To speed up the reconstruction, we only integrate every 10^{th} frame during TSDF-fusion. To obtain the projected depth of meshes, we use Pyrender [30].

4. Experiments

In this section, we first introduce the datasets we use, then present the main results, and finally discuss the ablation experiments. We also show the effectiveness of our 3D distillation qualitatively in Fig. 5.

4.1. Datasets

ScanNet (v2) dataset [4] is a large-scale indoor RGB-D dataset that includes both 2D and 3D data. It contains 1613 indoor scenes with ground truth camera poses and depth maps. We use the official train set (1201 scenes) for our model training. During training, we only use images and ground truth camera poses, without using ground truth depth data. We consider every 10^{th} frame as a target frame to reduce redundancy and for each, we find a source frame both backwards and forwards in time with a relative translation of 5-10 cm and a relative rotation of at most 3 degrees, forming 45 539 training triples. We evaluate using

Architecture	Model	ScanNet Val Set						
		Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Monodepth2 [14]	Self-Supervised [14]	0.167	0.100	0.385	0.203	0.764	0.935	0.981
	Self-Teaching [35]	0.160	0.090	0.365	0.193	0.780	0.941	0.983
	3D Distillation (ours)	**0.157**	**0.083**	**0.357**	**0.190**	**0.782**	**0.943**	**0.985**
HR-Depth [29]	Self-Supervised [14]	0.166	0.100	0.381	0.200	0.771	0.937	0.982
	Self-Teaching [35]	0.159	0.090	0.360	0.190	0.785	0.943	0.984
	3D Distillation (ours)	**0.154**	**0.080**	**0.349**	**0.186**	**0.788**	**0.945**	**0.986**
MonoViT [53]	Self-Supervised [14]	0.138	0.077	0.331	0.171	0.831	0.955	0.986
	Self-Teaching [35]	0.133	0.071	0.314	0.163	0.844	0.959	0.988
	3D Distillation (ours)	**0.128**	**0.060**	**0.296**	**0.157**	**0.846**	**0.962**	**0.990**
Architecture	Model	ScanNet Test Set						
		Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Monodepth2 [14]	Self-Supervised [14]	0.189	0.116	0.407	0.217	0.731	0.921	0.974
	Self-Teaching [35]	0.184	0.109	0.392	0.210	0.742	0.925	**0.976**
	3D Distillation (ours)	**0.181**	**0.105**	**0.388**	**0.208**	**0.746**	**0.927**	**0.976**
HR-Depth [29]	Self-Supervised [14]	0.184	0.111	0.399	0.212	0.739	0.925	0.976
	Self-Teaching [35]	0.178	0.102	0.381	0.204	0.752	0.931	**0.979**
	3D Distillation (ours)	**0.176**	**0.098**	**0.378**	**0.202**	**0.754**	**0.932**	**0.979**
MonoViT [53]	Self-Supervised [14]	0.154	0.082	0.343	0.182	0.801	0.948	0.984
	Self-Teaching [35]	0.152	0.081	0.329	0.177	0.811	0.948	0.983
	3D Distillation (ours)	**0.149**	**0.075**	**0.324**	**0.174**	**0.812**	**0.949**	**0.985**

Table 2: Main results on the ScanNet val and test sets [4]. 'Self-Supervised' indicates that the model is trained with the photometric loss [14]. 'Self-Teaching' indicates that the model is supervised by the predicted depth from self-supervised models and trained with the depth loss in Eq. (4). '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 Eq. (4). **Bold** indicates the best result of an architecture.

the complete official val set (312 scenes) and the complete official test set (100 scenes). To better evaluate the accuracy on reflective surfaces, we create ScanNet-Reflection, a subset in which reflective surfaces can be observed in every image. ScanNet-Reflection val and test sets consist of 439 and 121 images from the official val and test sets, respectively. To evaluate the accuracy on non-reflective surfaces, we also create a ScanNet-NoReflection val set, which consists of 1012 images without reflective surfaces from the official val set. We evaluate the absolute depth and use the standard depth metrics [7]. In the supplementary material, we provide the list of the training triples, the lists of the ScanNet-Reflection and ScanNet-NoReflection subsets, and the definitions of the evaluation metrics.

7-Scenes dataset [39] is a challenging RGB-D dataset captured in indoor scenes. To show the cross-dataset generalizability, we use models trained on ScanNet [4] to test on 7-Scenes [39], following [5, 38]. We use the test set in [5, 38], which consists of 13 sequences, and evaluate using the ground truth depth from [2]. We evaluate the relative depth as the camera intrinsics of different datasets [4, 39] are different, and use the standard depth metrics [7].

4.2. Main Results

ScanNet [4] results with and without our 3D distillation are shown in Tab. 2. We can see: (i) 3D distillation models achieve the best accuracy under all seven metrics, for three different backbones [14, 29, 53] and on both val and test sets. For example, using Monodepth2 architecture [14], 3D distillation can decrease the Sq Rel of the self-teaching model by 7.78% and 3.67% on the val and test sets, respectively; using HR-Depth architecture [29], 3D distillation can decrease the Sq Rel of the self-teaching model by 11.11% and 3.92% on the val and test sets, respectively; the corresponding improvements for MonoViT [53] are 15.49% and 7.41% on the val and test sets, respectively. (ii) The observed improvements of 3D distillation for a stronger model are larger. Specifically, on the val set, 3D distillation can decrease the Sq Rel of the self-teaching models by 7.78% / 11.11% / 15.49% for Monodepth2 [14], HR-Depth [29], and MonoViT [53] architectures, respectively; on the test set, 3D distillation can decrease the Sq Rel of the self-teaching models by 3.67% / 3.92% / 7.41% for Monodepth2 [14], HR-Depth [29], and MonoViT [53] architectures, respectively. We assume stronger models can better capture high-frequency information, thus their depth pre-

Architecture	Model	ScanNet-Reflection Val Set						
		Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Monodepth2 [14]	Self-Supervised [14]	0.206	0.227	0.584	0.246	0.750	0.912	0.961
	Self-Teaching [35]	0.192	0.188	0.548	0.233	0.764	0.920	0.967
	3D Distillation (ours)	**0.156**	**0.093**	**0.442**	**0.191**	**0.786**	**0.943**	**0.987**
HR-Depth [29]	Self-Supervised [14]	0.213	0.244	0.605	0.255	0.741	0.906	0.961
	Self-Teaching [35]	0.202	0.208	0.565	0.243	0.756	0.914	0.964
	3D Distillation (ours)	**0.153**	**0.090**	**0.430**	**0.188**	**0.789**	**0.948**	**0.989**
MonoViT [53]	Self-Supervised [14]	0.179	0.206	0.557	0.227	0.819	0.930	0.963
	Self-Teaching [35]	0.176	0.195	0.537	0.224	0.823	0.930	0.963
	3D Distillation (ours)	**0.126**	**0.068**	**0.367**	**0.159**	**0.851**	**0.965**	**0.991**
Architecture	Model	ScanNet-Reflection Test Set						
		Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Monodepth2 [14]	Self-Supervised [14]	0.181	0.160	0.521	0.221	0.758	0.932	0.976
	Self-Teaching [35]	0.179	0.146	0.502	0.218	0.750	0.938	0.980
	3D Distillation (ours)	**0.156**	**0.096**	**0.459**	**0.195**	**0.766**	**0.945**	**0.988**
HR-Depth [29]	Self-Supervised [14]	0.182	0.168	0.530	0.225	0.749	0.937	0.979
	Self-Teaching [35]	0.175	0.145	0.492	0.215	0.757	0.936	0.982
	3D Distillation (ours)	**0.152**	**0.089**	**0.451**	**0.190**	**0.771**	**0.956**	**0.990**
MonoViT [53]	Self-Supervised [14]	0.154	0.129	0.458	0.197	0.822	0.955	0.979
	Self-Teaching [35]	0.151	0.130	0.439	0.191	0.837	0.950	0.978
	3D Distillation (ours)	**0.127**	**0.069**	**0.379**	**0.162**	**0.846**	**0.961**	**0.992**

Table 3: Main results on the ScanNet-Reflection val and test sets [4]. ScanNet-Reflection is a subset in which specular reflections or glossy surfaces can be observed in every image.

Architecture	Model	ScanNet-NoReflection Val Set						
		Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Monodepth2 [14]	Self-Supervised [14]	0.169	0.100	0.395	0.206	0.759	0.932	0.979
	Self-Teaching [35]	0.161	0.090	0.375	0.196	0.777	0.939	0.981
	3D Distillation (ours)	**0.159**	**0.087**	**0.373**	**0.195**	**0.779**	**0.941**	**0.983**
HR-Depth [29]	Self-Supervised [14]	0.169	0.102	0.388	0.202	0.766	0.933	0.980
	Self-Teaching [35]	0.160	0.089	0.367	0.192	0.784	0.941	0.982
	3D Distillation (ours)	**0.158**	**0.086**	**0.365**	**0.190**	**0.786**	**0.942**	**0.983**
MonoViT [53]	Self-Supervised [14]	0.140	0.074	0.333	0.171	0.829	0.952	0.984
	Self-Teaching [35]	0.134	0.068	0.317	0.164	**0.840**	**0.956**	**0.987**
	3D Distillation (ours)	**0.133**	**0.065**	**0.311**	**0.162**	0.838	**0.956**	**0.987**

Table 4: Main results on the ScanNet-NoReflection val set [4]. ScanNet-NoReflection is a subset without reflective surfaces.

diction is more influenced by reflective surfaces. (iii) Self-teaching models are better than self-supervised models. We assume the depth loss, *i.e.*, Eq. (4), can decrease the contribution of challenging and faraway depth during training, thus improving overall accuracy. Nevertheless, our 3D distillation models are much better than self-teaching models.

ScanNet-Reflection [4] results with and without our 3D distillation are shown in Tab. 3. Our 3D distillation can significantly improve the depth accuracy, which supports the effectiveness for reflective surfaces. For example, on the Sq Rel metric of the val set, 3D distillation can improve the self-teaching models by 50.53% / 56.73% / 65.13% for Monodepth2 [14], HR-Depth [29], and MonoViT [53] architectures, respectively; on the Sq Rel metric of the test set, 3D distillation can improve the self-teaching models by 34.25% / 38.62% / 46.92% for Monodepth2 [14], HR-Depth [29], and MonoViT [53] architectures, respectively.

ScanNet-NoReflection [4] results are shown in Tab. 4. We can observe that 3D distillation improvements extend beyond reflective patches.

Architecture	Model	7-Scenes						
		Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Monodepth2 [14]	Self-Supervised [14]	0.153	0.071	0.323	0.190	0.793	0.959	0.989
	Self-Teaching [35]	0.152	0.069	0.321	0.188	0.796	0.961	0.989
	3D Distillation (ours)	**0.149**	**0.065**	**0.308**	**0.185**	**0.800**	**0.963**	**0.990**
HR-Depth [29]	Self-Supervised [14]	0.157	0.078	0.334	0.193	0.790	0.957	0.988
	Self-Teaching [35]	0.149	0.067	0.315	0.185	0.802	0.963	**0.990**
	3D Distillation (ours)	**0.147**	**0.064**	**0.304**	**0.183**	**0.804**	**0.965**	**0.990**
MonoViT [53]	Self-Supervised [14]	0.140	0.059	0.297	0.176	0.821	0.967	0.992
	Self-Teaching [35]	0.137	0.057	0.293	0.174	0.827	0.968	0.992
	3D Distillation (ours)	**0.134**	**0.053**	**0.284**	**0.170**	**0.831**	**0.972**	**0.993**

Table 5: Main results on 7-Scenes [39]. All the models are trained using the ScanNet train set [4].

Training Label	ScanNet Val Set						
	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
pred. depth	0.160	0.090	0.365	0.193	0.780	0.941	0.983
proj. depth	0.176	0.102	0.421	0.224	0.710	0.919	0.980
proj. depth with low-resolution training	0.344	0.341	0.833	0.484	0.326	0.602	0.804
pred. depth + proj. depth with diff	0.166	0.094	0.397	0.212	0.740	0.926	0.981
pred. depth + proj. depth with mv	0.158	0.085	0.365	0.195	0.773	0.940	0.984
pred. depth + proj. depth with uncer (ours)	**0.157**	**0.083**	**0.357**	**0.190**	**0.782**	**0.943**	**0.985**
Training Label	ScanNet-Reflection Val Set						
	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
pred. depth	0.192	0.188	0.548	0.233	0.764	0.920	0.967
proj. depth	0.189	0.130	0.565	0.249	0.664	0.901	0.979
proj. depth with low-resolution training	0.377	0.469	1.128	0.561	0.262	0.508	0.730
pred. depth + proj. depth with diff	0.172	0.115	0.521	0.229	0.709	0.914	0.983
pred. depth + proj. depth with mv	0.163	0.109	0.469	0.204	0.772	0.935	0.984
pred. depth + proj. depth with uncer (ours)	**0.156**	**0.093**	**0.442**	**0.191**	**0.786**	**0.943**	**0.987**

Table 6: Ablation experiments using different training labels. The network is Monodepth2 [14] architecture. 'pred. depth' and 'proj. depth' indicate only using the predicted depth or projected depth as training supervision, respectively. 'low-resolution training' denotes training the model with low-resolution [48]. 'diff' denotes fusing using the difference strategy. 'mv' denotes fusing using the multi-view consistency check [49]. 'uncer' denotes fusing using the uncertainty [23].

7-Scenes [39] results with and without our 3D distillation are shown in Tab. 5. 3D distillation models are still the best, which demonstrates the cross-dataset generalizability of 3D distillation. For example, on the Sq Rel metric, 3D distillation can improve the self-teaching models by 5.80% / 4.48% / 7.02% for Monodepth2 [14], HR-Depth [29], and MonoViT [53] architectures, respectively.

4.3. Ablation Experiments

We train models using different training labels and evaluate these models in Tab. 6. We use Monodepth2 architecture [14], as its training time is the shortest. We can make the following observations: (i) 'proj. depth' is much worse than 'pred. depth', as the projected depth is over-smoothing. This supports that it is important to fuse the predicted depth and projected depth. (ii) Among the strategies to fuse the predicted depth and projected depth, 'uncer' is better than 'diff' and 'mv'. 'diff' strategy means that, for a pixel x, if $D_t(x) - P_t(x) > 0.25 P_t(x)$, this pixel will be regarded as being on a reflective surface. 'mv' strategy means pixels which fail in multi-view consistency check [49] will be regarded as being on reflective surfaces. Specifically, in a training triple we check the pixels in the target frame with the aid of two source frames, *i.e.*, the three view consistency in [49]. (iii) Low-resolution training/high-resolution testing [48] is the worst, because cross-resolution testing for monocular depth estimation is challenging [8, 18]. Specifically, we train the model with the resolution of 128×96 and test the model with the resolution of 384×288.

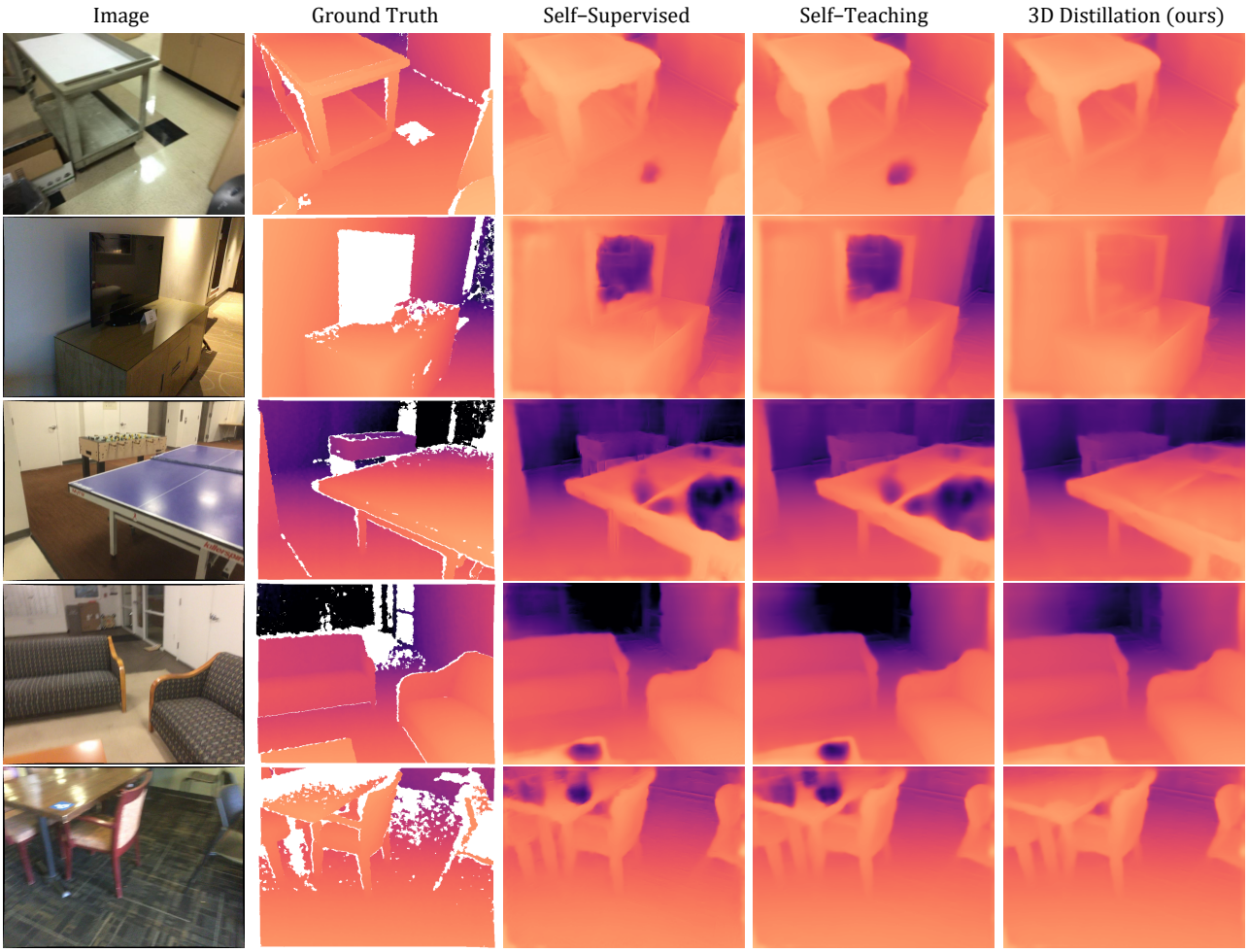


Figure 5: Our 3D distillation can significantly improve the depth prediction accuracy on reflective surfaces. The network is MonoViT architecture [53]. The image of the first row is from scene0704_00 of the ScanNet val set [4]. The images of the other rows are from scene0721_00, scene0776_00, scene0781_00, and scene0796_00 of the ScanNet test set [4], respectively.

5. Conclusion

We have proposed 3D distillation: a novel training framework for improving SSMDE on reflective surfaces. Motivated by the view-dependent property of reflective surfaces, 3D distillation utilizes the multi-view 3D information aggregated from the predicted depth of multiple frames to generate accurate pseudo-labels for reflective surfaces. 3D distillation significantly improves the depth estimation accuracy for various architectures [14, 29, 53] and on multiple datasets [4, 39], without adding computational cost or model parameters during inference.

Limitations and Future Work In the 3D distillation training framework, (i) perfect reflections such as mirrors are not handled; (ii) ensemble-based uncertainty requires multiple models during training; (iii) camera poses are assumed known during training. In future work, all these limitations could be tackled, *e.g.*, by using dedicated networks to predict reflection masks and camera poses. Besides, since depth and normal estimation are synergistic tasks [6], it could be a promising future direction to combine 3D distillation training with normal estimation and use predicted depth and surface normal to refine each other. Moreover, applying 3D distillation recursively could lead to more improvements. For the sake of simplicity, in this paper, we opt for a single iteration that already proves to be effective.

Acknowledgements T-K. Kim is supported by NST grant (CRC 21011, MSIT) and KOCCA grant (R2022020028, MCST).

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *IJCV*, 2016. 3, 5

[2] Eric Brachmann and Carsten Rother. Learning less is more - 6d camera localization via 3d surface regression. In *CVPR*, 2018. 6

[3] Di Chang, Aljaz Bozic, Tong Zhang, Qingsong Yan, Yingcong Chen, Sabine Süsstrunk, and Matthias Nießner. Rc-mvsnet: Unsupervised multi-view stereo with neural rendering. In *ECCV*, 2022. 3, 5

[4] 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, 2, 3, 4, 5, 6, 7, 8, 9

[5] Arda Düzçeker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. Deepvideomvs: Multi-view stereo on video with recurrent spatio-temporal fusion. In *CVPR*, 2021. 6

[6] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *ICCV*, 2021. 9

[7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014. 1, 6

[8] José M. Fácil, Benjamin Ummenhofer, Huizhong Zhou, Luis Montesano, Thomas Brox, and Javier Civera. Camconvs: Camera-aware multi-scale convolutions for single-view depth. In *CVPR*, 2019. 5, 8

[9] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004. 3

[10] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 3, 5

[11] Ravi Garg, B. G. Vijay Kumar, Gustavo Carneiro, and Ian D. Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016. 1, 2

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012. 1, 2

[13] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 1, 2, 4

[14] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 1, 2, 3, 4, 5, 6, 7, 8, 9

[15] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, 2020. 2

[16] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *ICLR*, 2020. 2

[17] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3d scene reconstruction with the manhattan-world assumption. In *CVPR*, 2022. 3, 5

[18] Mu He, Le Hui, Yikai Bian, Jian Ren, Jin Xie, and Jian Yang. Ra-depth: Resolution adaptive self-supervised monocular depth estimation. In *ECCV*, 2022. 5, 8

[19] 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

[20] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017. 3

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[22] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *ECCV*, 2020. 2

[23] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017. 3, 4, 5, 8

[24] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *CVPR*, 2022. 3

[25] Boying Li, Yuan Huang, Zeyu Liu, Danping Zou, and Wenxian Yu. Structdepth: Leveraging the structural regularities for self-supervised indoor depth estimation. In *ICCV*, 2021. 3

[26] Hanhan Li, Ariel Gordon, Hang Zhao, Vincent Casser, and Anelia Angelova. Unsupervised monocular depth learning in dynamic scenes. In *CoRL*, 2020. 2

[27] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *CVPR*, 2022. 3

[28] Xiaohu Lu, Jian Yao, Haoang Li, and Yahui Liu. 2-line exhaustive searching for real-time vanishing point estimation in manhattan world. In *WACV*, 2017. 3

[29] 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. 2, 3, 4, 5, 6, 7, 8, 9

[30] Matthew Matl. Pyrender. https://github.com/mmatl/pyrender, 2019. 5

[31] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3, 5

[32] 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. 2, 4, 5

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5

[34] Andrea Pilzer, Stéphane Lathuilière, Nicu Sebe, and Elisa Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *CVPR*, 2019. 2

[35] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *CVPR*, 2020. 3, 5, 6, 7, 8

[36] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 3

[37] Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng. Learning depth from single monocular images. In *NeurIPS*, 2005. 1

[38] Mohamed Sayed, John Gibson, Jamie Watson, Victor Prisacariu, Michael Firman, and Clément Godard. Simplerecon: 3d reconstruction without 3d convolutions. In *ECCV*, 2022. 5, 6

[39] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew W. Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *CVPR*, 2013. 1, 2, 6, 8, 9

[40] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *ECCV*, 2020. 2

[41] Jaime Spencer, Richard Bowden, and Simon Hadfield. Defeat-net: General monocular depth via simultaneous unsupervised representation learning. In *CVPR*, 2020. 2

[42] Jaime Spencer, C. Stella Qian, Chris Russell, Simon Hadfield, Erich W. Graf, Wendy J. Adams, Andrew J. Schofield, James H. Elder, Richard Bowden, Heng Cong, Stefano Mattoccia, Matteo Poggi, Zeeshan Khan Suri, Yang Tang, Fabio Tosi, Hao Wang, Youmin Zhang, Yusheng Zhang, and Chaoqiang Zhao. The monocular depth estimation challenge. In *WACVW*, 2023. 1, 3

[43] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Deconstructing self-supervised monocular reconstruction: The design decisions that matter. *TMLR*, 2022. 1

[44] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 4

[45] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel J. Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *CVPR*, 2021. 2

[46] Cho-Ying Wu, Jialiang Wang, Michael Hall, Ulrich Neumann, and Shuochen Su. Toward practical monocular indoor depth estimation. In *CVPR*, 2022. 3

[47] Hongbin Xu, Zhipeng Zhou, Yali Wang, Wenxiong Kang, Baigui Sun, Hao Li, and Yu Qiao. Digging into uncertainty in self-supervised multi-view stereo. In *ICCV*, 2021. 3

[48] Jiayu Yang, Jose M. Alvarez, and Miaomiao Liu. Self-supervised learning of depth inference for multi-view stereo. In *CVPR*, 2021. 3, 5, 8

[49] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018. 8

[50] Wang Yifan, Carl Doersch, Relja Arandjelovic, João Carreira, and Andrew Zisserman. Input-level inductive biases for 3d reconstruction. In *CVPR*, 2022. 5

[51] Zehao Yu, Lei Jin, and Shenghua Gao. P^2net: Patch-match and plane-regularization for unsupervised indoor depth estimation. In *ECCV*, 2020. 3

[52] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian D. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *CVPR*, 2018. 2

[53] 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, 2, 3, 4, 5, 6, 7, 8, 9

[54] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *CoRR*, 2018. 5

[55] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 1, 2