remote sensing

MDPI

Article

SIM-MultiDepth: Self-Supervised Indoor Monocular Multi-Frame Depth Estimation Based on Texture-Aware Masking

Xiaotong Guo [1,2], Huijie Zhao [3,4,*], Shuwei Shao [5], Xudong Li [1,2], Baochang Zhang [3,6,7] and Na Li [3,4]

1 School of Instrumentation and Optoelectronic Engineering, Key Laboratory of Precision Opto-Mechatronics Technology, Ministry of Education, Beihang University, Beijing 100191, China; guoxiaotong0420@buaa.edu.cn (X.G.); xdli@buaa.edu.cn (X.L.)
2 Qingdao Research Institute, Beihang University, Qingdao 266104, China
3 Institute of Artificial Intelligence, Beihang University, Beijing 100191, China; bczhang@buaa.edu.cn (B.Z.); lina_17@buaa.edu.cn (N.L.)
4 Aerospace Optical-Microwave Integrated Precision Intelligent Sensing, Key Laboratory of Ministry of Industry and Information Technology, Beihang University, Beijing 100191, China
5 School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China; swshao@buaa.edu.cn
6 Hangzhou Research Institute, Beihang University, Hangzhou 310051, China
7 Nanchang Institute of Technology, Nanchang 330044, China
* Correspondence: hjzhao@buaa.edu.cn

Abstract: Self-supervised monocular depth estimation methods have become the focus of research since ground truth data are not required. Current single-image-based works only leverage appearance-based features, thus achieving a limited performance. Deep learning based multiview stereo works facilitate the research on multi-frame depth estimation methods. Some multi-frame methods build cost volumes and take multiple frames as inputs at the time of test to fully utilize geometric cues between adjacent frames. Nevertheless, low-textured regions, which are dominant in indoor scenes, tend to cause unreliable depth hypotheses in the cost volume. Few self-supervised multi-frame methods have been used to conduct research on the issue of low-texture areas in indoor scenes. To handle this issue, we propose SIM-MultiDepth, a self-supervised indoor monocular multi-frame depth estimation framework. A self-supervised single-frame depth estimation network is introduced to learn the relative poses and supervise the multi-frame depth learning. A texture-aware depth consistency loss is designed considering the calculation of the patch-based photometric loss. Only the areas where multi-frame depth prediction is considered unreliable in low-texture regions are supervised by the single-frame network. This approach helps improve the depth estimation accuracy. The experimental results on the NYU Depth V2 dataset validate the effectiveness of SIM-MultiDepth. The zero-shot generalization studies on the 7-Scenes and Campus Indoor datasets aid in the analysis of the application characteristics of SIM-MultiDepth.

Keywords: self-supervised learning; indoor monocular depth estimation; multi-frame depth estimation

check for updates

Citation: Guo, X.; Zhao, H.; Shao, S.; Li, X.; Zhang, B.; Li, N. SIM-MultiDepth: Self-Supervised Indoor Monocular Multi-Frame Depth Estimation Based on Texture-Aware Masking. *Remote Sens.* **2024**, *16*, 2221. https://doi.org/10.3390/rs16122221

Academic Editor: Henrique Lorenzo

Received: 8 April 2024
Revised: 10 June 2024
Accepted: 14 June 2024
Published: 19 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Depth estimation has been widely applied in many downstream computer vision tasks, such as automatic driving, robot navigation, and augmented reality. Compared with obtaining depth values from active sensors, monocular depth estimation using a single camera is much more flexible. In terms of whether labeled data are required during training, monocular depth estimation methods are divided into supervised and self-supervised ones. Because of the benefit of not requiring ground truth, self-supervised methods have become the focus of research.

Many self-supervised monocular depth estimation methods [1–3] adopt a single image as the test input. Single-frame methods rely on appearance-based cues to predict scene depth. Although multiple frames are used in training, the geometric matching

information between adjacent frames is still ignored while testing. To fully utilize geometric cues, some studies [4–6] have used multiple frames to estimate the depth at the time of the test. Taking inspiration from deep learning based multiview stereo (MVS) [7], these methods build 3D cost volumes based on depth hypotheses and differentiable warping. However, dynamic and low-textured areas often cause wrong matchings to occur, which results in mistakes in the cost volume. Further, the whole training process is subject to overfitting. With photometric stereo methods [8,9], in which the 3D surface normal is recovered from multiple images with directional lighting, such low-textured areas, caused by non-Lambertian reflection, can be mitigated by the use of max-pooling layers, which fuse visual cues from multiple images [10]. However, it is difficult for such methods to adapt to complex lighting conditions in the estimation of depth. To solve these problems, researchers [4,11–13] have leveraged a single-frame depth estimation network to predict camera pose and supervise the training of the multi-frame network. However, current multi-frame methods mainly focus on predicting depth on outdoor datasets such as KITTI [14] and Cityscapes [15]. Few self-supervised multi-frame works perform depth estimation on indoor datasets such as NYU Depth V2 [16] and 7-Scenes [17].

We propose a self-supervised indoor monocular multi-frame depth estimation framework called SIM-MultiDepth. Following the ManyDepth [4] architecture, we build a cost volume with feature map warping and depth hypotheses. The cost volume and target frame feature are then processed by the depth decoder to predict depth.

In indoor depth estimation, a great challenge concerns the existence of many low-textured areas, such as tables, walls, and floors. The multi-frame network suffers severely from such regions and often falls into cost volume overfitting. Inspired by previous methods, we introduce a single-frame network to calculate camera poses for the multi-frame network to solve the overfitting issue. Additionally, the single-frame network supervises the multi-frame depth learning in selected areas to improve the overall performance. Figure 1 provides a comparison of the performances of the methods.

Different from ManyDepth, we designed a new single-frame supervision method. We selected the intersection of two parts of areas as the supervised region. The first part includes the pixels not involved in calculating the patch-based photometric loss. These areas can be considered as lacking unique textures, i.e., low-textured areas. The second part includes those areas where multi-frame depth prediction results are considered unreliable. In this way, the low-textured areas for which multi-frame prediction may be inaccurate will be supervised by single-frame depth. The proposed texture-aware masking manner has proven effective in the experiments on the NYU Depth V2 dataset [16]. The experiments, regarding some limitations, were performed to illustrate the capabilities and constraints of the framework. Zero-shot generalization experiments were also conducted on the 7-Scenes [17] and our collected Campus Indoor datasets [18]. The scenarios applicable for multi-frame and single-frame networks were further investigated.

In brief, the following three contributions were made:

1. We propose a novel self-supervised indoor monocular multi-frame depth estimation framework called SIM-MultiDepth. To solve the overfitting problem and achieve a better performance in low-textured areas, a single-frame depth estimation network is introduced to compute camera poses and serve as supervision.
2. Considering the patch-based photometric loss, a texture-aware masking supervision strategy is well-designed. The corresponding depth consistency loss guarantees that the points with discriminative features are only involved in geometric reasoning, instead of being forced to be consistent with the single-frame depth.
3. The experiments on the NYU Depth V2 dataset illustrate the effectiveness of our proposed SIM-MultiDepth and the texture-aware masking strategy. All evaluation metrics improved compared with the single-frame method. The generalization experiments on the 7-Scenes and Campus Indoor datasets also reveal the characteristics of the multi-frame and single-frame methods.

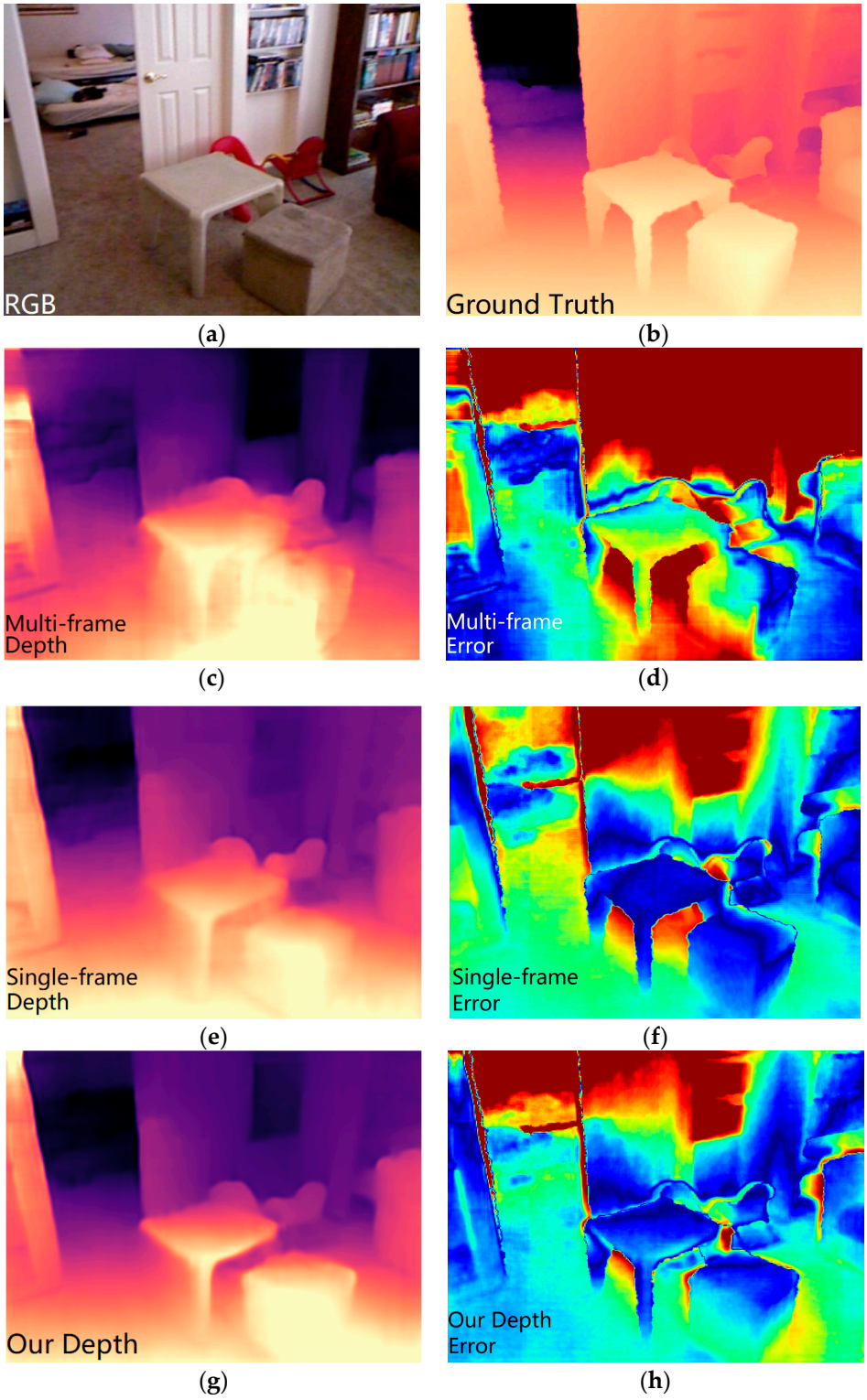


Figure 1. Depth estimation in indoor scenes. The original multi-frame method produces unsatisfactory results in low-textured areas such as doors and walls. The single-frame method handles these regions better. Our SIM-MultiDepth combines the advantages of both methods and performs the best. The absolute relative error is visualized for each method, ranging from blue (Abs Rel = 0) to red (Abs Rel = 0.3). (**a**) RGB image; (**b**) depth ground truth; (**c**) depth predicted by the multi-frame network; (**d**) error map of the multi-frame depth; (**e**) depth predicted by the single-frame network; (**f**) error map of the single-frame depth; (**g**) depth predicted by our SIM-MultiDepth; (**h**) error map of our SIM-MultiDepth.

2. Related Work

2.1. Indoor Monocular Single-Frame Depth Estimation

Monocular single-frame depth estimation aims to predict the depths of pixels in a single image. Many supervised methods [19–25] use pixel-wise regression to estimate depth. IEBins [26] treat depth estimation as classification–regression and leverages iterative elastic bins to calculate accurate depths. Recent advances in the attention mechanism have led many studies [27–35] to replace previously used Convolutional Neural Network (CNN) modules with Transformer [36]. Diffusion models have also been applied in supervised monocular depth estimation tasks [37,38]. Since there often exists high geometric regularity in indoor scenes, many supervised works design appropriate geometry-related loss terms to achieve better performance. Surface normal loss is applied in some works [39,40]. Plane-aware constraints are also introduced in many methods, such as PlaneNet [41], PlaneReg [42], P3Depth [43], and NDDepth [44].

Self-supervised monocular single-frame depth estimation methods usually adopt photometric consistency loss as the mainly used loss term. SfMLearner [1] first proposes self-supervised depth learning with view synthesis as supervision. Monodepth2 [2] is another classical method that reduces the impact of occlusion and moving object issues in auto-driving datasets such as KITTI. However, Monodepth2 fails to achieve a satisfactory performance in indoor scenarios because there exist many low-textured regions. Simple photometric consistency loss can lead to a large number of false matchings. To improve accuracy on indoor datasets, P^2Net [3] adopts patch-based photometric consistency loss and planar consistency loss based on Monodepth2 architecture. Similarly, StructDepth [45] uses Manhattan normal loss and coplanar loss as extra supervision. F^2Depth [18] leverages an improved optical flow network to provide accurate pixel motion in low-textured areas. Based on SC-Depth V1 [46], SC-Depth V2 [47] introduces an Auto-Rectify Network to handle the issue of excessive camera rotation in indoor scenes. For the same problem related to camera rotation, MonoIndoor [48] and MonoIndoor++ [49] use a residual pose estimation module after the initial pose estimation. From the expert network DPT [50,51], DistDepth [52] distills structure to the student network. GasMono [53] utilizes the structure-from-motion tool COLMAP [54] to perform coarse pose estimations and enhances feature processing in low-textured regions with Transformer.

For monocular depth estimation from a single image, the sequence information between adjacent frames is ignored. Thus, monocular multi-frame depth estimation that makes use of geometric cues has received more attention.

2.2. Monocular Multi-Frame Depth Estimation

Different from single-frame depth estimation, monocular multi-frame depth estimation exploits sequential multiple frames to estimate scene depth. Test-time refinement methods (e.g., [55]) iteratively refine the single-frame model for global temporal consistency during testing. These kinds of approaches will increase the computational cost. Another group of methods (e.g., [56]) are based on recurrent networks. The recently proposed MAMo [57] leverages memory and an attention mechanism to augment single-frame depth estimation networks into multi-frame depth estimation. Although these works use multiple frames to estimate depth at the time of test, geometric information is still not utilized explicitly.

To perform explicit geometry reasoning, some multiview stereo (MVS) studies [7,58,59] construct a 3D cost volume by image differentiable warping. However, these methods assume camera poses are known for training. For monocular depth estimation based on consecutive frames, the relative poses between adjacent frames are usually unknown.

Some supervised approaches [60,61] require relative pose as input to cost volumes at test time. Some methods [62,63] do not need known camera poses, but the number of frames used is unchangeable. Dynamic areas in outdoor scenarios tend to violate multiview consistency and cause false depth hypotheses in the cost volume. Recently, in [64] dynamic areas were managed by proposing a cross-cue fusion (CCF) module. The CCF module fuses single-frame and multi-frame cues to enhance both representations.

To eliminate the dependence on ground truth and known poses, ManyDepth [4] proposes a multi-frame depth estimation network based on single-frame Monodepth2 architecture. ManyDepth can process a variable number of frames during testing. For the overfitting problem in dynamic and low-textured areas, ManyDepth leverages the single-frame network to supervise the training of the multi-frame network. Following ManyDepth, DynamicDepth [5] proposes a Dynamic Object Cycle Consistent scheme to deal with the dynamic area problem. An Occlusion-aware Cost Volume is also designed to handle the occlusion issue. IterDepth [6] formulates depth estimation as residual learning and refines depth prediction results through iterative residual refinement.

Similar to single-frame depth estimation, Transformer architecture is also applied in multi-frame depth estimation. DepthFormer [65] enhances feature matching between adjacent frames during the generation of the cross-attention cost volume. Apart from attention-based modules, Dyna-DepthFormer [66] was proposed to estimate the motion field of moving objects to meet the challenge of dynamic areas.

Since single-frame depth estimation mainly leverages appearance-based features, while multi-frame methods rely on geometric constraints, some works tend to combine both to improve accuracy. MOVEDepth [12] adopts single-frame depth prediction results and velocity cues prior to the search for multi-frame depths. Furthermore, MOVEDepth fuses single-frame and multi-frame depths through uncertainty learning in the cost volume. In [13], a hybrid decoder consisting of single-frame and multi-frame pathways is proposed, and these two kinds of information are then fused in a multistage manner. In addition, [11] proposes a self-supervised framework in which single-frame and multi-frame depths supervise each other to fully leverage the mutual influence between them.

Almost all of the works mentioned above focus only on multi-frame depth estimation in outdoor scenes. Few researchers have proposed approaches to indoor monocular multi-frame depth estimation. Although MAMo [57] conducts experiments on the indoor NYU Depth V2 dataset, it is a supervised method and lacks explicit geometric reasoning.

Inspired by ManyDepth, we propose a self-supervised indoor monocular multi-frame depth estimation network called SIM-MultiDepth. In terms of the loss function, the manner in which the single-frame depth supervises the multi-frame depth is well-designed.

3. Methods

In this section, we introduce the proposed self-supervised indoor monocular multi-frame depth estimation framework, SIM-MultiDepth. The target frame and its source frame are taken as inputs to the PoseCNN to estimate the relative camera poses. Based on the feature map's differentiable warping and hypothesized depth values, a cost volume is constructed. Then, the features in the target image and the cost volume are processed by the encoder and depth decoder to generate the depth results. To address overfitting caused by the cost volume, we updated the PoseCNN parameters relying only on the single-frame network prediction results. In addition, the generated depths by the single-frame network were used to supervise the multi-frame depth learning in precisely selected areas. The contributions guarantee the performance of SIM-MultiDepth both in textured and low-textured areas.

In Section 3.1, the overall multi-frame depth estimation network architecture is explained. Section 3.2 introduces the well-designed texture-aware depth consistency loss, and the overall loss function is illustrated in Section 3.3.

3.1. Overview of SIM-MultiDepth

We aimed for it to perform geometric reasoning with adjacent frames, usually called source frames, to estimate the target frame depth. A pipeline diagram of SIM-MultiDepth is shown in Figure 2.

Multiple frames are processed by building a cost volume that fuses their features. The target frame, $I_t \in \mathbb{R}^{H\times W\times 3}$, and the source frame, $I_s \in \mathbb{R}^{H\times W\times 3}$, are input into the PoseCNN to compute the relative camera pose, $T_{t\to s}$. The target frame feature map,

$F_t \in \mathbb{R}^{H/4\times W/4\times C_f}$, and the source frame feature map, $F_s \in \mathbb{R}^{H/4\times W/4\times C_f}$, are extracted by the encoder. C_f is the number of feature map channels.

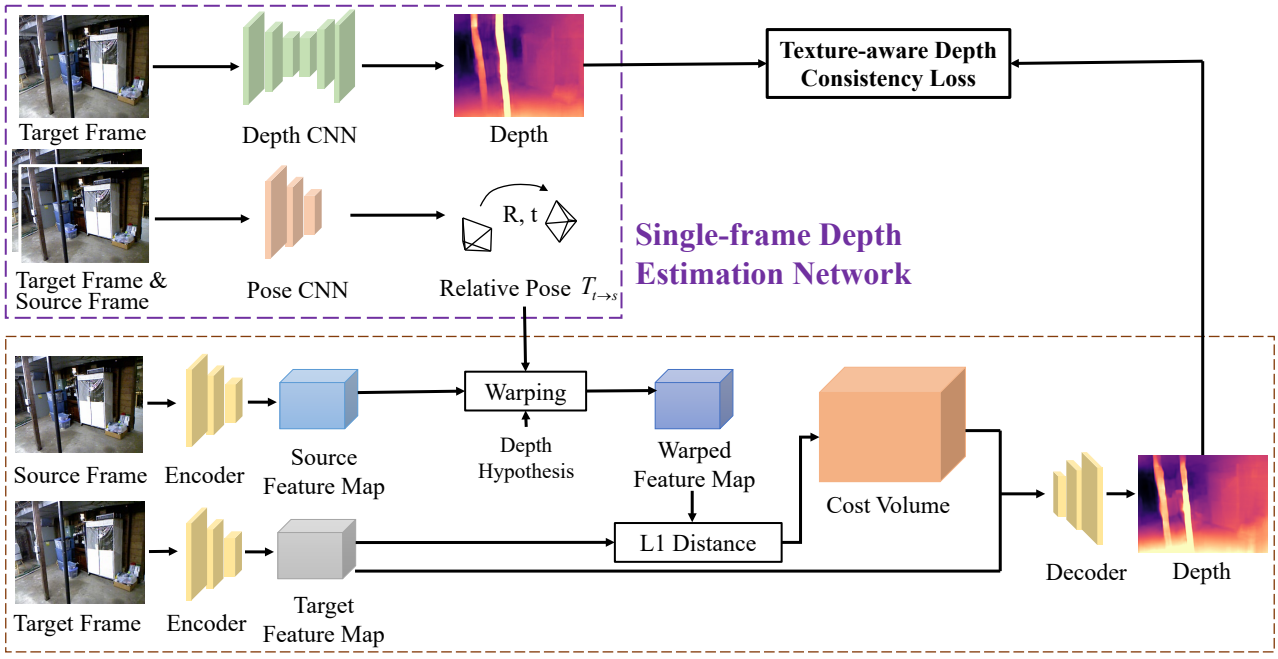


Figure 2. A pipeline diagram of SIM-MultiDepth. Multiple frames are input to encoders to generate feature maps. With the relative poses computed by the single-frame network, the source frame feature map is warped to build the cost volume. The cost volume and the target frame feature are then decoded into depths. The depths predicted by the single-frame network serve as supervision for the multi-frame depth by computing texture-aware depth consistency loss.

We suppose that there are N_d depth values, written as d_i. For each d_i, the synthesized feature map, $F^i_{s\to t}$, is the following:

$$F^i_{s\to t} = F_s\langle proj(d_i, T_{t\to s}, K)\rangle \tag{1}$$

where K denotes the camera intrinsic parameters matrix, $\langle\rangle$ denotes the sampling operation, $proj()$ describes the projection relationship between the predicted depth and the 2D homogeneous coordinates in the feature maps. In the cost volume, $CV \in \mathbb{R}^{H/4\times W/4\times N_d}$, every depth value, d_i, corresponds to $CV_i \in \mathbb{R}^{H/4\times W/4\times 1}$, which can be calculated by the following:

$$CV_i = \frac{1}{C_f}\sum_{c=1}^{C_f}(|F_t(c) - F^i_{s\to t}(c)|) \tag{2}$$

Briefly, the cost volume describes the probability that the pixel (u, v) corresponds to the depth value d_i. Then, the cost volume concatenated with the target frame feature is taken as input into the depth decoder to compute the depth.

We can observe that the training process falls into overfitting because there are many low-textured areas in indoor scenarios. In these areas, the information in the cost volume becomes untrustworthy. Mistakes in the cost volume affect the depth prediction results for both training and testing. Different from the multi-frame network's overreliance on geometric matching information, a single-frame network tends to leverage appearance cues to estimate depth. The single-frame network is more robust in low-textured areas.

To solve the overfitting problem of the multi-frame network, we introduce a trained single-frame depth estimation network to provide pose information. The multi-frame

network does not participate in updating the parameters of the PoseCNN. In order to further improve the accuracy, the depth results of the single-frame network serve as supervision for the multi-frame depth learning in some specific areas. In Section 3.2, we show how the supervised areas are selected.

3.2. Texture-Aware Depth Consistency Loss

To improve the overall depth estimation results, an effective single-frame supervision strategy was designed. We precisely select those pixels with depth predictions that are unreliable in low-textured areas to be supervised by the single-frame depth.

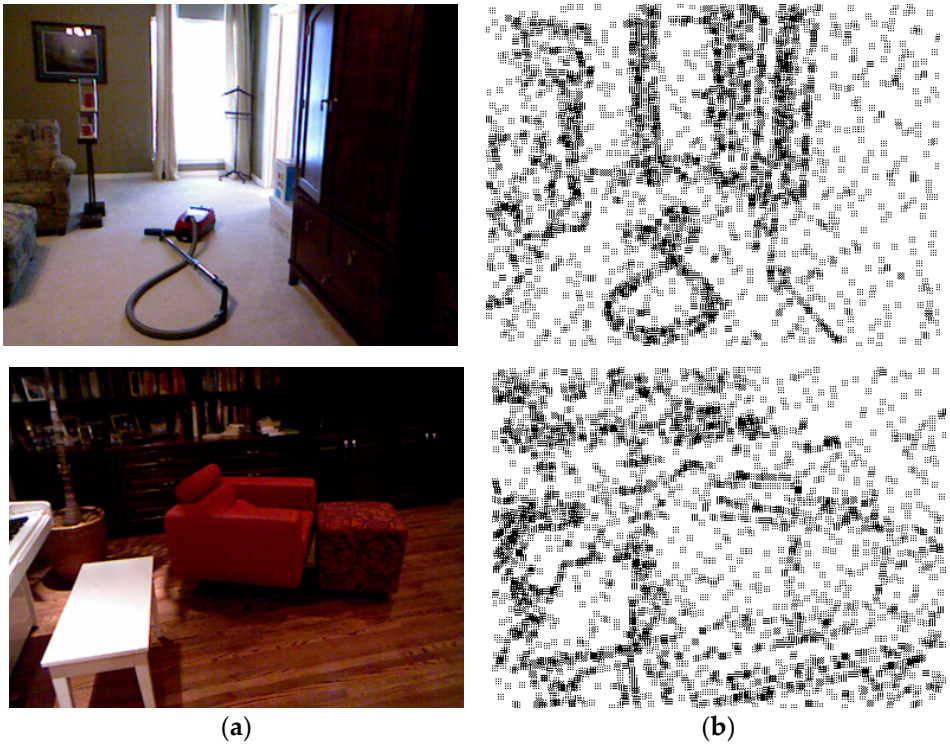
Firstly, we need to determine which pixels can be considered as lacking in discriminative textures. The loss function contains a patch-based photometric consistency loss. The photometric loss involves pixels around extracted key points with significant features. The pixels are distributed in 3×3 patches centered at each key point, which can be written as the following:

$$P = \{(x_i + x_k, y_i + y_k), x_k \in \{-N, 0, N\}, y_k \in \{-N, 0, N\}, 1 \leq i \leq n, i \in \mathbb{N}\} \quad (3)$$

where (x_i, y_i) are the 2D coordinates of the key point, n is the number of key points, and N is set to 2. We set the binary mask, M_1, to 1 in the following regions, except for the pixels in P:

$$M_1 = p \notin P \quad (4)$$

As shown in Figure 3, the black patches in the images have distinctive features while the white areas are considered to be nondiscriminative.



(**a**) (**b**)

Figure 3. Examples of images and extracted patches: (**a**) RGB image; (**b**) extracted patches.

Since the overall loss function has been elaborately designed for indoor scenes, not all pixels in low-textured areas will produce wrong depth results. The supervision from the

single-frame depth should be more targeted. There are some unreliable prediction results that can be formulated, as follows:

$$M_2 = \max\left(\frac{D_{cv} - \hat{D}_t}{\hat{D}_t}, \frac{\hat{D}_t - D_{cv}}{D_{cv}}\right) > 1 \tag{5}$$

where M_2 is a binary mask representing which depth values are predicted to potentially be wrong; $\hat{D}_t$ is the depth generated by the single-frame network. D_{cv} is calculated by the following:

$$D_{cv} = \text{argmin}(CV) \tag{6}$$

The final texture-aware supervision mask, M, is computed as follows:

$$M = M_1 \circ M_2 \tag{7}$$

Intuitively, M is set to 1 at the intersection of pixels producing unreliable depths and those not involved in the patch-based photometric loss.

The texture-aware depth consistency loss term is written as follows:

$$L_{consistency} = \left|D_t(p_t) - \hat{D}_t(p_t)\right| \tag{8}$$

where p_t represents the pixels, M is set to 1, and D_t is the predicted depth of the multi-frame network.

ManyDepth [4] adopts a similar single-frame supervisory approach. The difference is that the multi-frame network is supervised to approach single-frame depths in all M_2 areas. The disadvantage of this strategy is that some points with significant features are forced to approach wrong single-frame prediction results. These points could have achieved better accuracy through geometric inferencing of the multi-frame network. The advantage of selecting the intersection of M_1 and M_2 is that it not only ensures that almost all strong feature points participate in the geometric reasoning of the multi-frame network but also filters out points that are identified as being unreliable. The benefits of the multi-frame depth estimation network and the single-frame network combine well to achieve a better performance.

3.3. Overall Loss Functions

For the multi-frame depth estimation network SIM-MultiDepth, the final loss function is defined as follows:

$$L = L_{ph} + \lambda_1 L_{sm} + \lambda_2 L_{spp} + L_{rigid} + \lambda_3 L_{feature} + L_{consistency} \tag{9}$$

where L_{ph} denotes the patch-based photometric consistency loss, L_{sm} is the smoothness loss, and L_{spp} represents the planar consistency loss. These three loss terms have the same definitions as in P^{2}Net [3]. Moreover, λ_1 is set to 0.001, λ_2 is set to 0.05, and λ_3 is set to 3 in the experiments.

The patch-based photometric consistency loss, L_{ph}, is formulated as follows:

$$L_{L1} = \left\| I_t\left[P_i^t\right] - I_s\left[P_i^{t\to s}\right] \right\|_1 \tag{10}$$

$$L_{SSIM} = SSIM(I_t[P_i^t], I_s[P_i^{t\to s}]) \tag{11}$$

$$L_{ph} = \alpha L_{SSIM} + (1-\alpha) L_{L1} \tag{12}$$

where P is as defined in Equation (3), and α is set to 0.85.

The smoothness loss, L_{sm}, is formulated as follows:

$$L_{sm} = |\partial_x d_t^*| e^{-|\partial_x I_t|} + |\partial_y d_t^*| e^{-|\partial_y I_t|} \tag{13}$$

where $d_t^* = d_t / \overline{d}_t$ denotes the normalized depth prediction.

The planar consistency loss, L_{spp}, is formulated as follows:

$$L_{spp} = \sum_{m=1}^{M} \sum_{n=1}^{N} |D(p_n) - D'(p_n)| \tag{14}$$

where L_{spp} is as proposed in P^{2}Net [3], M denotes the number of segmented plane super-pixels, and N denotes the number of pixels in every plane super-pixel. L_{spp} helps improve the accuracy in the planar regions.

L_{rigid} is our previously proposed optical flow consistency loss for the single-frame depth estimation network [18]. It guarantees that the optical flow generated from the depth estimation network is more accurate. The loss term is as follows:

$$L_{rigid} = \left| f_{rigid}(p_t) - f_{flow}(p_t) \right| \tag{15}$$

where $f_{flow}(p_t)$ stands for the optical flow generated by the introduced flow estimation network.

The multiscale feature map synthesis loss, $L_{feature}$, is another previously designed constraint for the single-frame network [18], which is calculated as follows:

$$L_{feature} = \left| F_{src}(p_t) - \hat{F}(p_t) \right| \tag{16}$$

where $F_{src}(p_t)$ denotes feature maps of the source frame computed by the flow estimation network, and $\hat{F}(p_t)$ denotes warped feature maps by the optical flow of the depth estimation network.

4. Experiments

4.1. Implementation Details

Datasets: The multi-frame depth estimation network SIM-MultiDepth was trained and evaluated on the publicly available NYU Depth V2 dataset [16]. NYU Depth V2 is made up of 582 indoor scenes, which were captured by Microsoft Kinect. We adopted the same train split as in previous works [3,67]. The training dataset contains 21,483 images recorded in 283 scenes. All training images were undistorted and sampled at intervals of every 10 frames. The images $\{I_{t-1}, I_t, I_{t+1}\}$ were used for training. Our SIM-MultiDepth was evaluated on the official test split consisting of 654 labeled target images, I_t, and their source images, I_{t-1}. We selected the 10th frame before the target image as the source image. The images $\{I_t, I_{t-1}\}$ participated in constructing the cost volume during both the training and evaluation.

The 7-Scenes dataset [17] is composed of 7 indoor scenes. The official test video sequences are provided. For the zero-shot generalization experiments, we sampled the test split at intervals of 10 frames. The final test set contains 1700 images.

In addition to 7-Scenes, we also conducted zero-shot generalization studies on the Campus Indoor dataset collected by us [18]. The Campus Indoor dataset contains 99 images of 18 indoor scenes on a campus. Monocular images were captured using a FUJIFILM X-T30 camera. For calculating depth, 13–16 points were selected from each image, forming a total of approximately 1500 points. The selected points are distributed evenly over a large depth range. As shown in Figure 4, the red stars represent the points. Their ground truth values were measured using a laser rangefinder.

Experimental setup: The SIM-MultiDepth architecture is based on ManyDepth [4]. The adopted single-frame network is our previously proposed monocular depth estimation network, F^{2}Depth [18]. The parameters of the single-frame network were frozen. Training images were color augmented and flipped randomly. The images are of size 288×384 for the training. The SIM-MultiDepth was trained on a single NVIDIA GeForce RTX 3090 GPU. We trained for 23 epochs with a batch size of 12. The learning rate was 10^{-4} without adjustment.

Evaluation metrics: Similar to previous research [3,18,67], we used a median-scaling strategy, since there exists scale ambiguity. The evaluation metrics include mean absolute

relative error (*Abs Rel*), root mean squared error (*RMS*), mean log10 error, and three accuracies under different thresholds (1.25, 1.25^2, and 1.25^3). The metrics were calculated according to the following:

$$Abs\ Rel = \frac{1}{N}\sum_{i \in N}\frac{\left|D_i - \hat{D}_i\right|}{\hat{D}_i} \tag{17}$$

$$RMS = \sqrt{\frac{1}{N}\sum_{i \in N}\left|\left|D_i - \hat{D}_i\right|\right|^2} \tag{18}$$

$$Mean\ \log 10 = \frac{1}{N}\sum_{i \in N}\left|\log_{10} D_i - \log_{10} \hat{D}_i\right| \tag{19}$$

$$Accuracies = \%\ of\ D_i\ s.t.\ \max\left(\frac{D_i}{\hat{D}_i}, \frac{\hat{D}_i}{D_i}\right) = \delta\ <\ thr \tag{20}$$

where $\hat{D}$ denotes the ground truth, D stands for the predicted depth results, N represents the total number of pixels, and thr is the threshold for calculating the accuracies.

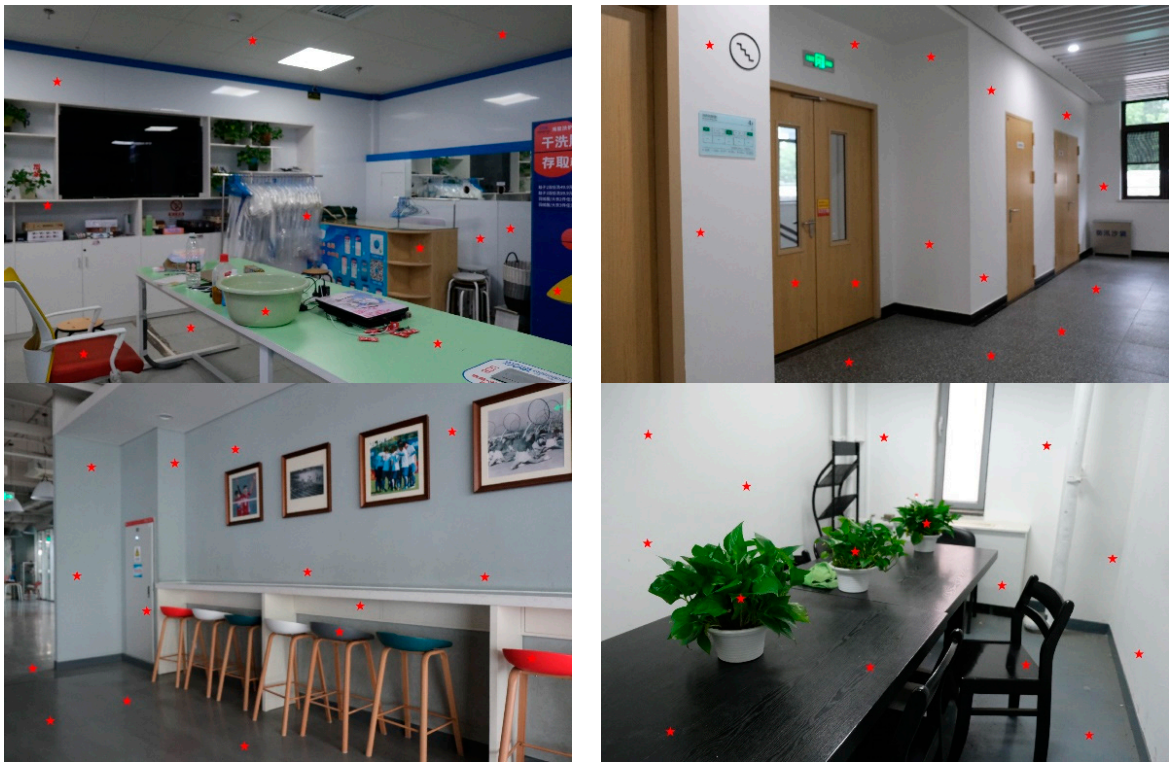


Figure 4. Examples of points in the Campus Indoor dataset. The red stars represent the selected points.

4.2. Results

4.2.1. Evaluation Results on NYU Depth V2

An evaluation was conducted using the official test set of NYU Depth V2 to compare with other supervised and self-supervised methods. To our knowledge, no multi-frame methods have been evaluated using the same test set, so only single-frame methods were selected for comparison. The results are listed in Table 1. Compared with our single-frame network, F^2Depth, the proposed multi-frame method, SIM-MultiDepth, achieved significant improvements in all metrics. The results of the evaluation indicate the effectiveness of SIM-MultiDepth in indoor scenarios.

Table 1. Quantitative results of SIM-MultiDepth and other existing methods on the NYU Depth V2 dataset. The best results are in bold. ↓ indicates that the lower the value, the better; ↑ indicates that the higher the value, the better.

Method	Supervision	REL ↓	RMS ↓	Log10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Liu [68]	✓	0.335	1.060	0.127	-	-	-
Li [69]	✓	0.232	0.821	0.094	0.621	0.886	0.968
Liu [25]	✓	0.213	0.759	0.087	0.650	0.906	0.976
Eigen [21]	✓	0.158	0.641	-	0.769	0.950	0.988
Li [22]	✓	0.143	0.635	0.063	0.788	0.958	0.991
PlaneNet [41]	✓	0.142	0.514	0.060	0.827	0.963	0.990
PlaneReg [42]	✓	0.134	0.503	0.057	0.827	0.963	0.990
Laina [20]	✓	0.127	0.573	0.055	0.811	0.953	0.988
DORN [19]	✓	0.115	0.509	0.051	0.828	0.965	0.992
VNL [40]	✓	0.108	0.416	0.048	0.875	0.976	0.994
P3Depth [43]	✓	0.104	0.356	0.043	0.898	0.981	0.996
Jun [31]	✓	0.100	0.362	0.043	0.907	0.986	0.997
DDP [38]	✓	**0.094**	**0.329**	**0.040**	**0.921**	**0.990**	**0.998**
Moving Indoor [67]	×	0.208	0.712	0.086	0.674	0.900	0.968
TrianFlow [70]	×	0.189	0.686	0.079	0.701	0.912	0.978
Zhang [71]	×	0.177	0.634	-	0.733	0.936	-
Monodepth2 [2]	×	0.170	0.617	0.072	0.748	0.942	0.986
ADPDepth [72]	×	0.165	0.592	0.071	0.753	0.934	0.981
SC-Depth [46]	×	0.159	0.608	0.068	0.772	0.939	0.982
P^2Net [3]	×	0.159	0.599	0.068	0.772	0.942	0.984
F^2Depth [18]	×	0.158	0.583	0.067	0.779	0.947	0.987
Ours	×	**0.152**	**0.567**	**0.065**	**0.792**	**0.950**	**0.988**

Figure 5 shows the predicted depth maps of the single-frame network P^2Net, F^2Depth, and our multi-frame network SIM-MultiDepth. From the visualization of the results, it can be observed that SIM-MultiDepth produced more accurate and detailed depth estimation results than F^2Depth and P^2Net. Not only arethe prediction results for significantly textured areas, such as chairs and sofas, satisfactory, but the depth results for low-textured areas, like tables, arealso good.

For some of the limitations that may be present in indoor scenes, we provide a comprehensive analysis to illustrate the capabilities and constraints of SIM-MultiDepth.

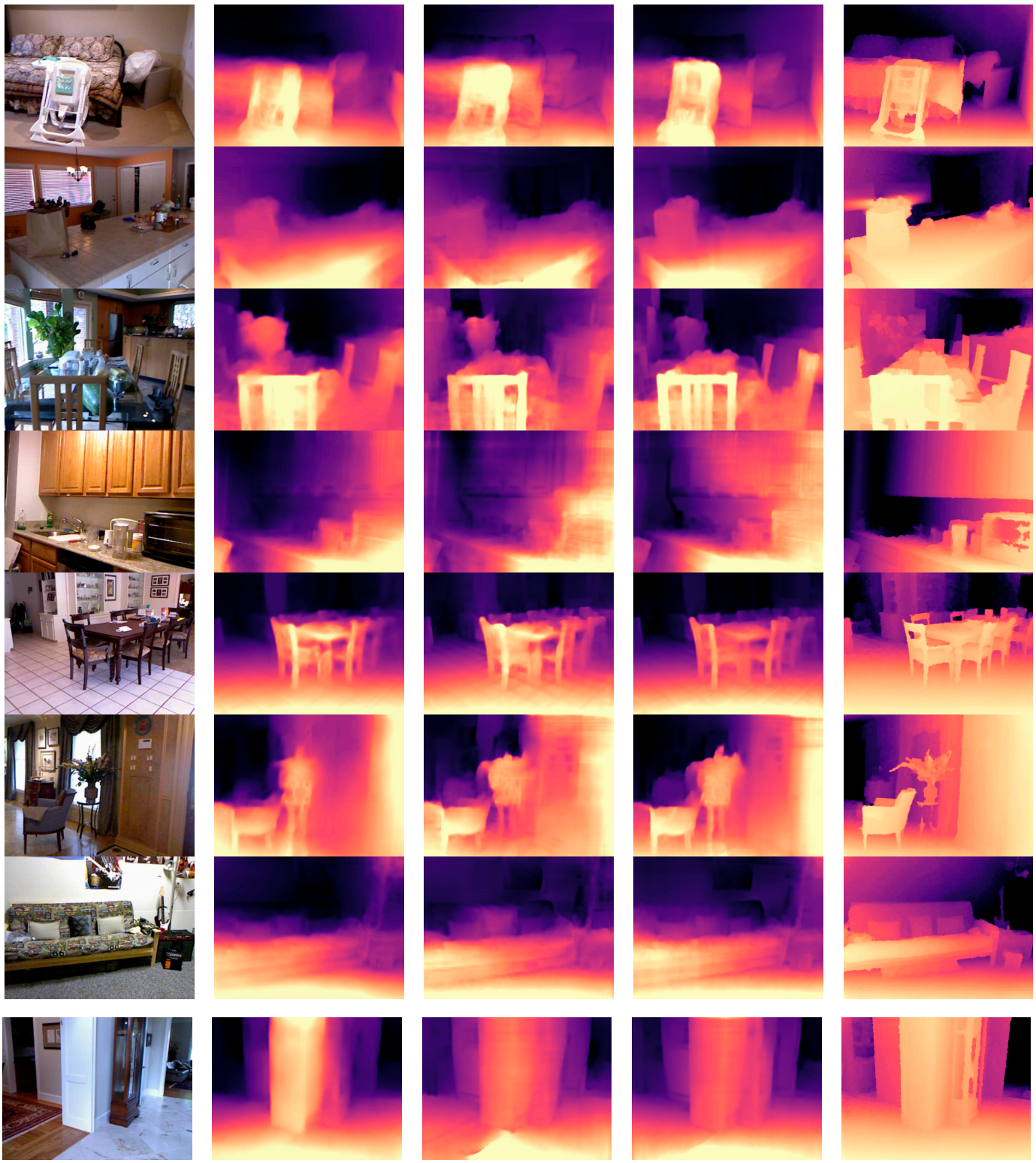
For transparent objects such as glass, we selected a bathroom with a glass door to evaluate SIM-MultiDepth's performance. The results of the evaluation are shown in Figure 6. Theoretically, there should be no large changes in depth on either side of the green line in the RGB image. The predictions by SIM-MultiDepth aremuch better than the ground truth. The ground truth were captured with a Microsoft Kinect device. The infrared light emitted to the glass was not reflected back to the camera correctly.

For reflective surfaces, we selected a reflective television screen as an example. The prediction results are shown in Figure 7. There is no significant change in depth in the reflective areas of the screen. These results illustrate that SIM-MultiDepth is robust to interference from reflections.

Changes in illumination are another challenge that cannot be ignored. We multiplied the two frames, $\{I_t, I_{t-1}\}$, by the same brightness factor, b, to simulate common changes in illumination in indoor scenes; b was set to 0.5, 0.8, 1, 1.2, 1.5, and 2. The corresponding illumination changes are shown in Figure 8.

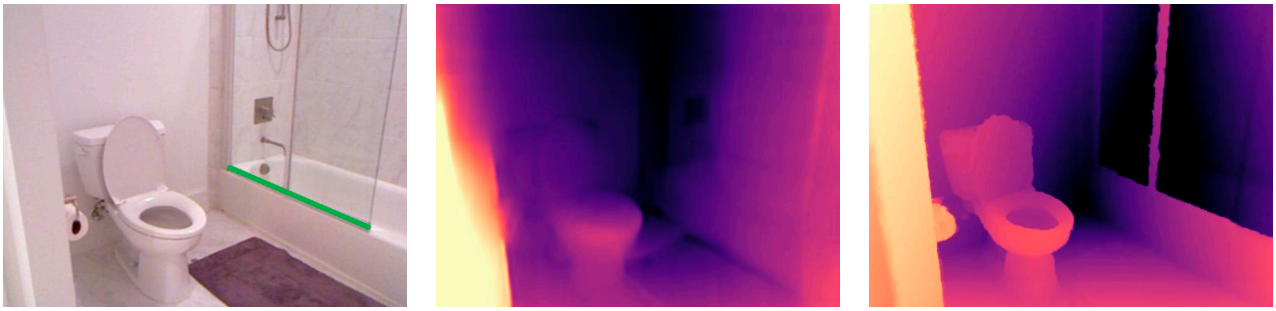
We evaluated SIM-MultiDepth on the test set with the six brightness factors. The results of the evaluation are listed in Table 2. Compared with the original test set, the results are almost unchanged when b is set to 0.8 and 1.2. This is largely because the training data augmentation randomly selects b from [0.8, 1.2]. If b falls outside of [0.8, 1.2], the prediction results will be significantly worse. To handle this issue, the value range for b can be further

expanded in the augmentation of the training data. A semantic consistency loss can also be designed to meet the challenge of changes in illumination.



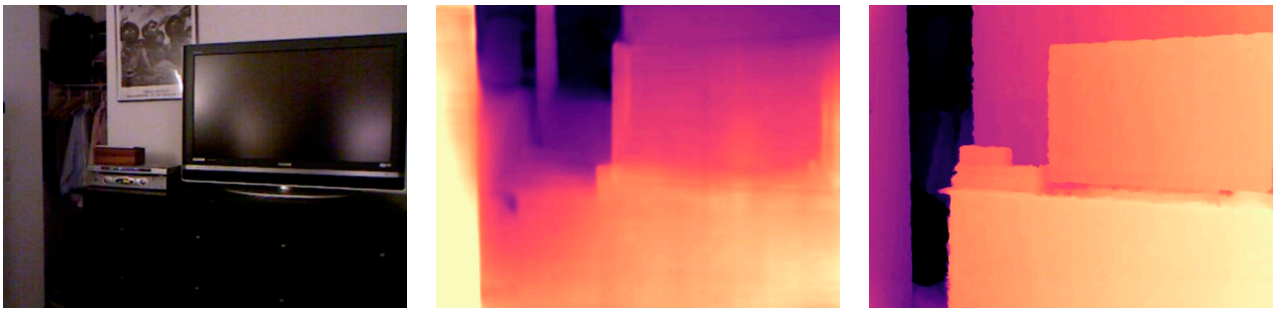
(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 5. Qualitive depth prediction results for the NYU Depth V2 dataset: (**a**) RGB image; (**b**) P^2Net [3]; (**c**) F^2Depth [18]; (**d**) SIM-MultiDepth; (**e**) ground truth.



(**a**) (**b**) (**c**)

Figure 6. Qualitive depth prediction results in a scenewith a transparent glass door: (**a**) RGB image; (**b**) SIM-MultiDepth; (**c**) ground truth.



(**a**) (**b**) (**c**)

Figure 7. Qualitive depth prediction results with reflective surfaces: (**a**) RGB image; (**b**) SIM-MultiDepth; (**c**) ground truth.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

Figure 8. The changes in illumination corresponding to different brightness factor, b, values: (**a**) $b = 0.5$; (**b**) $b = 0.8$; (**c**) $b = 1$; (**d**) $b = 1.2$; (**e**) $b = 1.5$; (**f**) $b = 2$.

Table 2. Evaluation results with different brightness factors. The underlined results were predicted by the original test set. ↓ indicates that the lower the value, the better; ↑ indicates that the higher the value, the better.

b	REL ↓	RMS ↓	Log10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
0.5	0.156	0.578	0.066	0.783	0.946	0.987
0.8	0.152	0.567	0.065	0.792	0.950	0.988
1	0.152	0.567	0.065	0.792	0.950	0.988
1.2	0.152	0.568	0.065	0.791	0.949	0.988
1.5	0.153	0.571	0.065	0.788	0.948	0.988
2	0.157	0.581	0.067	0.780	0.946	0.987

Additionally, experiments focusing on baseline changes of I_t and I_{t-1} were performed. I_t and I_{t-1} in the original test set were selected at intervals of 10 frames. Here, we selected them at intervals of 5, 20, 40, 80, and 160 frames. Correspondingly, five experiments were conducted, in all of which exactly the same results as the original test set were achieved. The results verify that SIM-MultiDepth is highly adaptable to baseline changes.

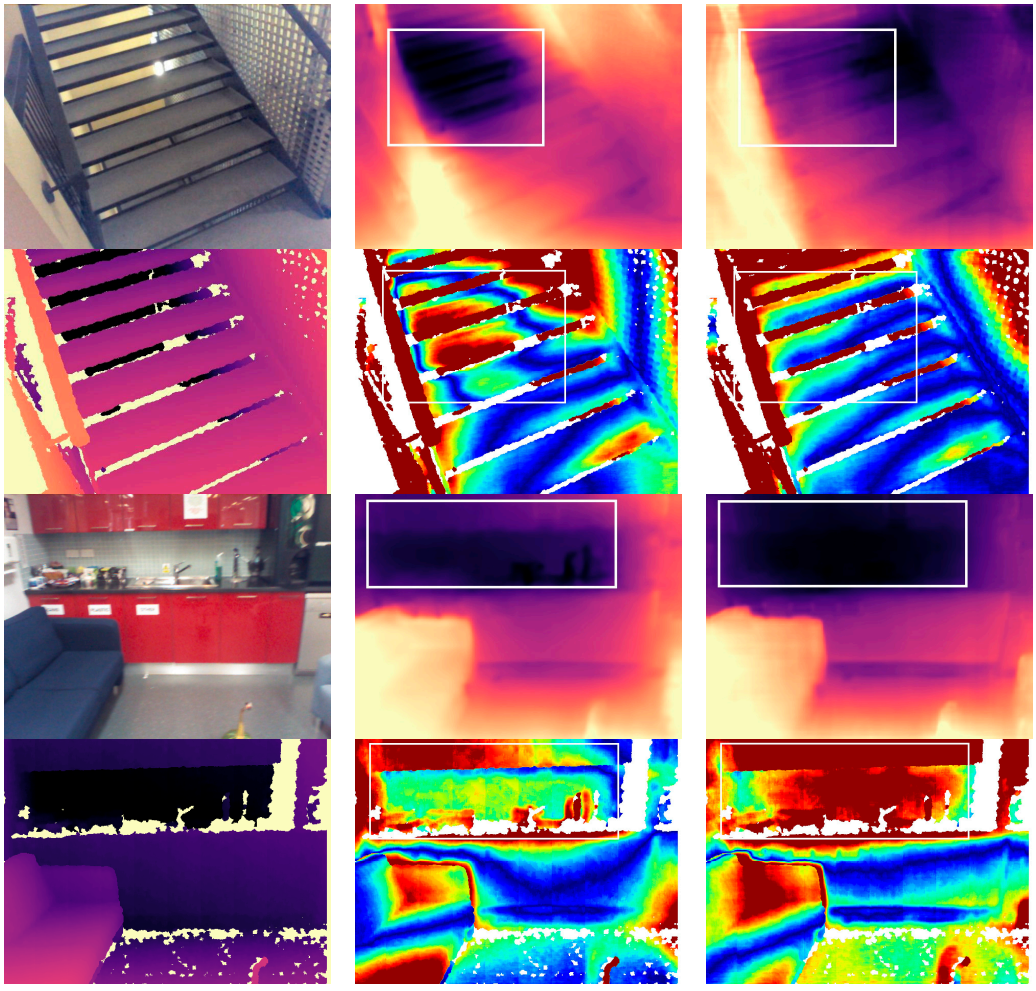
4.2.2. Zero-Shot Generalization Results on 7-Scenes

To evaluate the generalization abilities of SIM-MultiDepth, we conducted zero-shot generalization studies on the 7-Scenes dataset. A comparison of the results for SIM-MultiDepth and F^{2}Depth is provided in Table 3.

Table 3. Comparison results of our SIM-MultiDepth and F^{2}Depth [18] on 7-Scenes. The best results are in bold. ↓ represents the lower the better, ↑ represents the higher the better.

Methods	Our SIM-MultiDepth						F^{2}Depth [18]					
Scene	REL ↓	RMS ↓	Log10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑	REL ↓	RMS ↓	Log10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Chess	**0.186**	0.411	**0.081**	0.668	**0.940**	0.992	**0.186**	**0.409**	**0.081**	**0.671**	0.936	**0.993**
Fire	0.180	0.332	0.078	0.690	0.950	**0.991**	**0.176**	**0.322**	**0.076**	**0.701**	**0.951**	**0.991**
Heads	0.196	0.204	0.082	0.674	0.924	**0.985**	**0.185**	**0.195**	**0.078**	**0.718**	**0.930**	0.983
Office	**0.159**	**0.363**	**0.066**	**0.766**	**0.970**	**0.996**	0.162	0.370	0.067	0.762	0.963	**0.996**
Pumpkin	0.136	0.372	0.059	0.813	0.978	**0.996**	**0.127**	**0.350**	**0.056**	**0.846**	**0.980**	0.995
RedKitchen	**0.171**	**0.414**	**0.073**	**0.724**	**0.950**	**0.994**	0.173	0.416	0.074	0.722	0.946	0.992
Stairs	**0.147**	**0.437**	**0.064**	**0.784**	**0.922**	**0.974**	0.159	0.455	0.068	0.766	0.912	0.972
Average	**0.167**	0.376	**0.071**	0.734	**0.954**	**0.992**	**0.167**	**0.375**	**0.071**	**0.740**	0.950	**0.992**

The SIM-MultiDepth achieved a similar generalization performance as F^{2}Depth. In some scenarios, such as Stairs and Office, SIM-MultiDepth outperformed F^{2}Depth. In other scenes, such as with Pumpkin and Heads, F^{2}Depth generalized better than SIM-MultiDepth. To analyze the respective applicable scenarios of F^{2}Depth and SIM-MultiDepth, we visualized their predicted depths in different scenes, as shown in Figure 9.



(**a**) (**b**) (**c**)

Figure 9. Depth visualization results for the 7-Scenes dataset. The first two rows depict the scene Stairs, for which SIM-MultiDepth outperformed F^{2}Depth [18]. The last two rows display thescene

Pumpkin, for which F^{2}Depth generalized better than SIM-MultiDepth. (**a**) RGB images and ground truth; (**b**) F^{2}Depth generalization results and error maps; (**c**) SIM-MultiDepth generalization results and error maps. The absolute relative error ranges, in color, from blue (*Abs Rel* = 0) to red (*Abs Rel* = 0.2).

From Figure 9, it can be found that the multi-frame network SIM-MultiDepth generalized better in regions with significant features or sharp color changes. A typical example is that SIM-MultiDepth could better distinguish differences in the depths between the steps on the set of stairs than F^{2}Depth. But the single-frame network F^{2}Depth made better predictions in low-textured areas, such as cupboard doors, walls, and floors.

To investigate the cause of the variance in the performances on the 7-Scenes dataset, we analyzed the data characteristics for the Pumpkin and Heads for which F^{2}Depth outperformed SIM-MultiDepth. The Pumpkin scene is mainly composed of low-textured planes, as shown in Figure 9. The Heads also has a low-textured desktop as its main component, as shown in Figure 10.

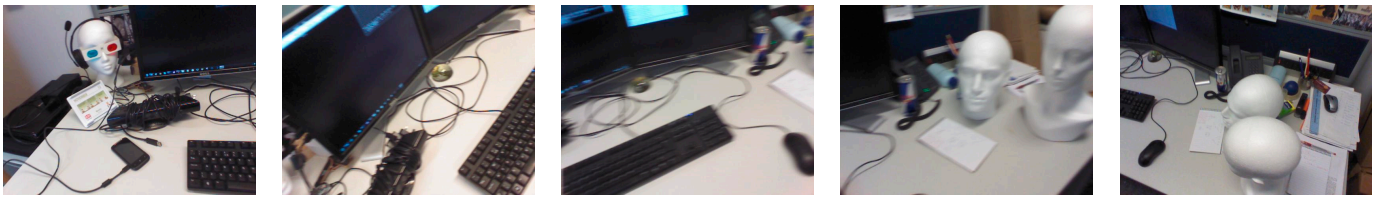


Figure 10. Typical examples of the scene Heads in the 7-Scenes dataset.

4.2.3. Zero-Shot Generalization Results on Campus Indoor

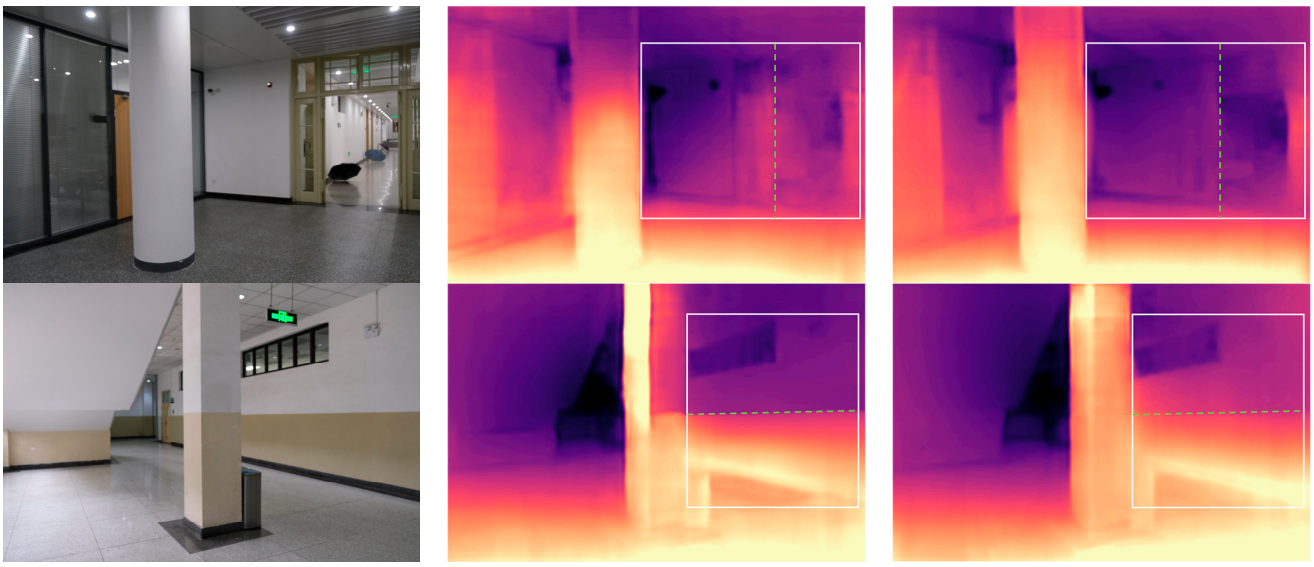
In addition to the 7-Scenes dataset, we also conducted generalization studies on the Campus Indoor dataset to provide a more comprehensive analysis. The experiment results are listed in Table 4.

Table 4. Comparison of the results of SIM-MultiDepth and F^{2}Depth [18] on the Campus Indoor dataset. The best results are in bold. ↓ indicates that the lower the value, the better; ↑ indicates that the higher the value, the better.

Methods	Our SIM-MultiDepth						F^{2}Depth [18]					
Scene No.	REL ↓	RMS ↓	Log10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑	REL ↓	RMS ↓	Log10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
1	**0.156**	**0.590**	**0.063**	0.750	**0.977**	**1**	0.164	0.592	0.064	**0.766**	0.947	**1**
2	0.144	1.083	0.066	0.752	**1**	**1**	**0.135**	**0.949**	**0.060**	**0.798**	0.981	**1**
3	0.218	**0.887**	**0.083**	**0.685**	0.946	**1**	**0.215**	0.889	0.084	0.644	**0.973**	**1**
4	0.142	**0.559**	0.060	**0.818**	**1**	**1**	**0.134**	0.588	**0.058**	0.777	**1**	**1**
5	0.183	1.224	0.078	0.667	**0.973**	**1**	**0.159**	**1.091**	**0.067**	**0.746**	0.946	**1**
6	**0.136**	**1.487**	**0.066**	**0.774**	**0.933**	**0.972**	0.154	1.625	0.073	0.757	0.919	0.932
7	**0.244**	**2.224**	**0.109**	**0.514**	**0.830**	**0.973**	0.266	2.321	0.122	0.436	0.802	0.960
8	0.172	0.791	0.071	0.745	0.915	**1**	**0.149**	**0.699**	**0.063**	**0.772**	**0.929**	**1**
9	0.099	0.418	0.041	0.907	**1**	**1**	**0.088**	**0.372**	**0.037**	**0.946**	**1**	**1**
10	0.232	0.761	0.091	0.583	**0.933**	**1**	**0.193**	**0.686**	**0.076**	**0.717**	0.917	**1**
11	0.183	0.674	0.076	0.733	0.973	0.987	**0.168**	**0.667**	**0.072**	**0.787**	**0.973**	**1**
12	**0.124**	**0.512**	**0.058**	**0.763**	**1**	**1**	0.143	0.536	0.066	0.736	**1**	**1**
13	**0.144**	**0.594**	**0.067**	0.790	**0.974**	**1**	0.147	**0.594**	**0.067**	**0.803**	**0.974**	**1**
14	0.141	0.443	0.062	0.827	0.987	**1**	**0.135**	**0.422**	**0.060**	**0.880**	**1**	**1**
15	0.256	0.433	0.087	**0.747**	0.813	**0.947**	**0.236**	**0.409**	**0.084**	0.720	**0.867**	0.946
16	**0.163**	**0.527**	**0.067**	**0.787**	**0.960**	**1**	0.202	0.647	0.079	0.707	0.920	**1**
17	0.127	0.270	0.050	**0.907**	0.947	0.987	**0.114**	**0.244**	**0.047**	0.893	**0.960**	**1**
18	**0.144**	**0.227**	0.063	**0.800**	**0.987**	**1**	**0.144**	0.230	**0.061**	0.787	**0.987**	**1**
Average	0.165	0.769	**0.069**	0.754	**0.956**	**0.993**	**0.164**	**0.753**	**0.069**	**0.760**	0.950	0.991

SIM-MultiDepth and F^{2}Depth exhibited similar performances on the Campus Indoor dataset, similar to the 7-Scenes dataset. SIM-MultiDepth surpassed F^{2}Depth on nine scenes, while F^{2}Depth performed better on the remaining nine scenes. In scene Nos. 6 and 7, SIM-MultiDepth displayed a particularly superior performance, as visualized in Figure 11. The depths of the pillars were predicted much more reliably by SIM-MultiDepth. Moreover, SIM-MultiDepth predicted the changes in depth for the entire scene more accurately, as framed in Figure 11. Specifically, in No. 6, the depth to the right of the green dotted line should be larger than that of the left; SIM-MultiDepth correctly predicted this. Also, for No. 7, the depths above and below the green line should almost be the same. Although

these two methods both failed to make perfect predictions, SIM-MultiDepth reduced the difference in depths between the two parts.



(**a**) (**b**) (**c**)

Figure 11. Depth visualization results for scene Nos. 6 and 7 in the Campus Indoor dataset. The first row shows scene No. 6. The second row shows scene No. 7. (**a**) RGB images; (**b**) F^{2}Depth [18] generalization results; (**c**) SIM-MultiDepth generalization results.

For scene No. 5, F^{2}Depth was the most superior to SIM-MultiDepth. It can be seen that the overall performance of SIM-MultiDepth was good. We determined the points that caused the main gap in the performances, which are represented by the white stars in Figure 12. Compared with scene Nos. 6 and 7, the scene structure of No. 5 is relatively simple. The main parts of No. 5 are low-textured walls and floors. We noticed that F^{2}Depth generalized better for the entire wall than SIM-MultiDepth. White stars are distributed where SIM-MultiDepth generalized unsatisfactorily, which is the upper part of the wall. Since the Campus Indoor dataset is mainly made up of images of public spaces instead of private rooms, large areas with white walls exist. SIM-MultiDepth did not perform that satisfactorily in these regions.



(**a**) (**b**) (**c**)

Figure 12. Depth visualization results of scene No. 5 in the Campus Indoor dataset: (**a**) RGB image; (**b**) F^{2}Depth [18] generalization results; (**c**) SIM-MultiDepth generalization results.

4.2.4. Ablation Studies on NYU Depth V2

To validate the effectiveness of the proposed supervision mask, M, in texture-aware depth consistency loss, we conducted ablation studies on the NYU Depth V2. In Table 5, we present the prediction results without the texture-aware depth consistency loss in the first row. The evaluation results corresponding with M_1 and M_2 are reported in the second

and third rows. The fourth row presents the results of our well-designed mask. It can be observed that the proposed texture-aware masking had positive effects on the multi-frame depth estimations.

Table 5. Ablation studies of our masking manner on the NYU Depth V2 dataset. The best results are in bold. ↓ indicates that the lower the value, the better; ↑ indicates that the higher the value, the better.

Final Mask M	REL ↓	RMS ↓	Log10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
No consistency loss	0.158	0.591	0.067	0.779	0.943	0.985
M_1	0.153	0.570	**0.065**	0.791	0.949	0.987
M_2	**0.152**	0.570	**0.065**	0.790	0.949	0.987
$M_1 \circ M_2$	**0.152**	**0.567**	**0.065**	**0.792**	**0.950**	**0.988**

5. Conclusions and Discussions

In this paper, we presented a self-supervised monocular multi-frame depth estimation framework for indoor scenes, called SIM-MultiDepth. We built a cost volume to perform geometric reasoning for depth estimation. For the overfitting problem caused by the cost volume, we introduced a trained single-frame network to compute the relative poses. To improve the prediction accuracy, a single-frame network was also leveraged to supervise the multi-frame network. We design a supervision strategy to precisely supervise necessary areas. The intersections of low-textured areas and areas producing unreliable depths were selected to be supervised by the single-frame network. The texture-aware masking approach allowed for pixels with unique textures to only participate in multi-frame geometric reasoning and without being consistent with the single-frame depth. The useful cues of both the single-frame and multi-frame depth estimation were fused in our proposed SIM-MultiDepth. The contributions were validated as effective in the experiments on the NYU Depth V2 dataset. In the generalization studies, SIM-MultiDepth outperformed the single-frame network in most areas, except low-textured areas. It can be concluded that the single-frame network relying on appearance cues still has better generalization abilities in low-textured areas. Future work will focus on building more robust cost volumes to further handle correspondence ambiguity. The attention-based Transformer module will be applied to enhance feature learning. More reliable loss terms will also be exploited to address problems related to indoor multi-frame depth estimation.

Author Contributions: Conceptualization, H.Z., X.G. and B.Z.; methodology, S.S.; software, X.G.; formal analysis, X.G. and S.S.; investigation, X.G.; resources, H.Z. and X.L.; data curation, X.G.; writing—original draft preparation, X.G.; writing—review and editing, S.S., N.L. and X.L.; visualization, X.G.; supervision, H.Z. and N.L.; project administration, N.L., X.L. and B.Z.; funding acquisition, H.Z. and B.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Application Innovation Project of CASC (grant number: 6230109004); National Key Research and Development Program of China (grant number: 2023YFC3300029); Zhejiang Provincial Natural Science Foundation of China (grant number: LD24F020007); "One Thousand Plan" projects in Jiangxi Province (grant number: Jxsg2023102268).

Data Availability Statement: The Campus Indoor dataset is not publicly available because of the campus confidentiality policy and privacy issues.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhou, T.; Brown, M.; Snavely, N.; Lowe, D.G. Unsupervised learning of depth and ego-motion from video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1851–1858.
2. Godard, C.; Mac Aodha, O.; Firman, M.; Brostow, G.J. Digging into self-supervised monocular depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3828–3838.

3. Yu, Z.; Jin, L.; Gao, S. P^2Net: Patch-match and plane-regularization for unsupervised indoor depth estimation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXIV. pp. 206–222.
4. Watson, J.; Mac Aodha, O.; Prisacariu, V.; Brostow, G.; Firman, M. The temporal opportunist: Self-supervised multi-frame monocular depth. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 1164–1174.
5. Feng, Z.; Yang, L.; Jing, L.; Wang, H.; Tian, Y.; Li, B. Disentangling object motion and occlusion for unsupervised multi-frame monocular depth. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 228–244.
6. Feng, C.; Chen, Z.; Zhang, C.; Hu, W.; Li, B.; Lu, F. IterDepth: Iterative residual refinement for outdoor self-supervised multi-frame monocular depth estimation. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *34*, 329–341. [CrossRef]
7. Xu, H.; Zhou, Z.; Qiao, Y.; Kang, W.; Wu, Q. Self-supervised multi-view stereo via effective co-segmentation and data-augmentation. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 3030–3038. [CrossRef]
8. Shi, B.; Wu, Z.; Mo, Z.; Duan, D.; Yeung, S.-K.; Tan, P. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3707–3716.
9. Ju, Y.; Lam, K.-M.; Xie, W.; Zhou, H.; Dong, J.; Shi, B. Deep Learning Methods for Calibrated Photometric Stereo and Beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* 2024; *early access*. [CrossRef]
10. Chen, G.; Han, K.; Shi, B.; Matsushita, Y.; Wong, K.-Y.K. Deep photometric stereo for non-lambertian surfaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 129–142. [CrossRef]
11. Xiang, J.; Wang, Y.; An, L.; Liu, H.; Liu, J. Exploring the mutual influence between self-supervised single-frame and multi-frame depth estimation. *IEEE Robot. Autom. Lett.* **2023**, *8*, 6547–6554. [CrossRef]
12. Wang, X.; Zhu, Z.; Huang, G.; Chi, X.; Ye, Y.; Chen, Z.; Wang, X. Crafting monocular cues and velocity guidance for self-supervised multi-frame depth learning. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 2689–2697. [CrossRef]
13. Long, Y.; Yu, H.; Liu, B. Two-stream based multi-stage hybrid decoder for self-supervised multi-frame monocular depth. *IEEE Robot. Autom. Lett.* **2022**, *7*, 12291–12298. [CrossRef]
14. Uhrig, J.; Schneider, N.; Schneider, L.; Franke, U.; Brox, T.; Geiger, A. Sparsity invariant cnns. In Proceedings of the 2017 International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; pp. 11–20.
15. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
16. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from rgbd images. In *Proceedings of Computer Vision–ECCV 2012: 12th European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 746–760.
17. Shotton, J.; Glocker, B.; Zach, C.; Izadi, S.; Criminisi, A.; Fitzgibbon, A. Scene coordinate regression forests for camera relocalization in RGB-D images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2930–2937.
18. Guo, X.; Zhao, H.; Shao, S.; Li, X.; Zhang, B. F^2Depth: Self-supervised indoor monocular depth estimation via optical flow consistency and feature map synthesis. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108391. [CrossRef]
19. Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Tao, D. Deep ordinal regression network for monocular depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2002–2011.
20. Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; Navab, N. Deeper depth prediction with fully convolutional residual networks. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 239–248.
21. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2650–2658.
22. Li, J.; Klein, R.; Yao, A. A two-streamed network for estimating fine-scaled depth maps from single rgb images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3372–3380.
23. Zhang, S.; Yang, L.; Mi, M.B.; Zheng, X.; Yao, A. Improving deep regression with ordinal entropy. *arXiv* **2023**, arXiv:2301.08915.
24. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2366–2374.
25. Liu, F.; Shen, C.; Lin, G.; Reid, I. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 2024–2039. [CrossRef]
26. Shao, S.; Pei, Z.; Wu, X.; Liu, Z.; Chen, W.; Li, Z. IEBins: Iterative elastic bins for monocular depth estimation. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 53025–53037.
27. Agarwal, A.; Arora, C. Attention attention everywhere: Monocular depth prediction with skip attention. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 5861–5870.

28. Agarwal, A.; Arora, C. Depthformer: Multiscale vision transformer for monocular depth estimation with global local information fusion. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 3873–3877.
29. Bhat, S.F.; Alhashim, I.; Wonka, P. LocalBins: Improving depth estimation by learning local distributions. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part I. pp. 480–496.
30. Bhat, S.F.; Alhashim, I.; Wonka, P. Adabins: Depth estimation using adaptive bins. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4009–4018.
31. Jun, J.; Lee, J.-H.; Lee, C.; Kim, C.-S. Depth map decomposition for monocular depth estimation. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part II. pp. 18–34.
32. Ning, J.; Li, C.; Zhang, Z.; Wang, C.; Geng, Z.; Dai, Q.; He, K.; Hu, H. All in tokens: Unifying output space of visual tasks via soft token. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 19900–19910.
33. Shao, S.; Pei, Z.; Chen, W.; Li, R.; Liu, Z.; Li, Z. URCDC-Depth: Uncertainty rectified cross-distillation with cutflip for monocular depth estimation. *IEEE Trans. Multimed.* **2023**, *26*, 3341–3353. [CrossRef]
34. Piccinelli, L.; Sakaridis, C.; Yu, F. iDisc: Internal discretization for monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 21477–21487.
35. Yuan, W.; Gu, X.; Dai, Z.; Zhu, S.; Tan, P. Neural window fully-connected CRFs for monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3916–3925.
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
37. Zhao, W.; Rao, Y.; Liu, Z.; Liu, B.; Zhou, J.; Lu, J. Unleashing text-to-image diffusion models for visual perception. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 5729–5739.
38. Ji, Y.; Chen, Z.; Xie, E.; Hong, L.; Liu, X.; Liu, Z.; Lu, T.; Li, Z.; Luo, P. DDP: Diffusion model for dense visual prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 21741–21752.
39. Hu, J.; Ozay, M.; Zhang, Y.; Okatani, T. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 7–11 January 2019; pp. 1043–1051.
40. Yin, W.; Liu, Y.; Shen, C.; Yan, Y. Enforcing geometric constraints of virtual normal for depth prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5684–5693.
41. Liu, C.; Yang, J.; Ceylan, D.; Yumer, E.; Furukawa, Y. PlaneNet: Piece-wise planar reconstruction from a single rgb image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2579–2588.
42. Yu, Z.; Zheng, J.; Lian, D.; Zhou, Z.; Gao, S. Single-image piece-wise planar 3d reconstruction via associative embedding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1029–1037.
43. Patil, V.; Sakaridis, C.; Liniger, A.; Van Gool, L. P3Depth: Monocular depth estimation with a piecewise planarity prior. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1610–1621.
44. Shao, S.; Pei, Z.; Chen, W.; Wu, X.; Li, Z. NDDepth: Normal-distance assisted monocular depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 7931–7940.
45. Li, B.; Huang, Y.; Liu, Z.; Zou, D.; Yu, W. StructDepth: Leveraging the structural regularities for self-supervised indoor depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 12663–12673.
46. Bian, J.; Li, Z.; Wang, N.; Zhan, H.; Shen, C.; Cheng, M.-M.; Reid, I. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–11.
47. Bian, J.-W.; Zhan, H.; Wang, N.; Chin, T.-J.; Shen, C.; Reid, I. Auto-rectify network for unsupervised indoor depth estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 9802–9813. [CrossRef]
48. Ji, P.; Li, R.; Bhanu, B.; Xu, Y. MonoIndoor: Towards good practice of self-supervised monocular depth estimation for indoor environments. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 12787–12796.
49. Li, R.; Ji, P.; Xu, Y.; Bhanu, B. MonoIndoor++: Towards better practice of self-supervised monocular depth estimation for indoor environments. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 830–846. [CrossRef]
50. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision transformers for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 12179–12188.
51. Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; Koltun, V. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1623–1637. [CrossRef]
52. Wu, C.-Y.; Wang, J.; Hall, M.; Neumann, U.; Su, S. Toward practical monocular indoor depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3814–3824.

53. Zhao, C.; Poggi, M.; Tosi, F.; Zhou, L.; Sun, Q.; Tang, Y.; Mattoccia, S. GasMono: Geometry-aided self-supervised monocular depth estimation for indoor scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 16209–16220.
54. Schonberger, J.L.; Frahm, J.-M. Structure-from-motion revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4104–4113.
55. Luo, X.; Huang, J.-B.; Szeliski, R.; Matzen, K.; Kopf, J. Consistent video depth estimation. *ACM Trans. Graph. (ToG)* **2020**, *39*, 71. [CrossRef]
56. Patil, V.; Van Gansbeke, W.; Dai, D.; Van Gool, L. Don't forget the past: Recurrent depth estimation from monocular video. *IEEE Robot. Autom. Lett.* **2020**, *5*, 6813–6820. [CrossRef]
57. Yasarla, R.; Cai, H.; Jeong, J.; Shi, Y.; Garrepalli, R.; Porikli, F. MAMo: Leveraging memory and attention for monocular video depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 8754–8764.
58. Yang, J.; Alvarez, J.M.; Liu, M. Self-supervised learning of depth inference for multi-view stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7526–7534.
59. Ding, Y.; Zhu, Q.; Liu, X.; Yuan, W.; Zhang, H.; Zhang, C. KD-MVS: Knowledge distillation based self-supervised learning for multi-view stereo. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 630–646.
60. Liu, C.; Gu, J.; Kim, K.; Narasimhan, S.G.; Kautz, J. Neural rgb→d sensing: Depth and uncertainty from a video camera. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10986–10995.
61. Hou, Y.; Kannala, J.; Solin, A. Multi-view stereo by temporal nonparametric fusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2651–2660.
62. Wu, Z.; Wu, X.; Zhang, X.; Wang, S.; Ju, L. Spatial correspondence with generative adversarial network: Learning depth from monocular videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7494–7504.
63. Wimbauer, F.; Yang, N.; Von Stumberg, L.; Zeller, N.; Cremers, D. MonoRec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6112–6122.
64. Li, R.; Gong, D.; Yin, W.; Chen, H.; Zhu, Y.; Wang, K.; Chen, X.; Sun, J.; Zhang, Y. Learning to fuse monocular and multi-view cues for multi-frame depth estimation in dynamic scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 21539–21548.
65. Guizilini, V.; Ambruș, R.; Chen, D.; Zakharov, S.; Gaidon, A. Multi-frame self-supervised depth with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 160–170.
66. Zhang, S.; Zhao, C. Dyna-DepthFormer: Multi-frame transformer for self-supervised depth estimation in dynamic scenes. *arXiv* **2023**, arXiv:2301.05871.
67. Zhou, J.; Wang, Y.; Qin, K.; Zeng, W. Moving Indoor: Unsupervised video depth learning in challenging environments. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8618–8627.
68. Liu, M.; Salzmann, M.; He, X. Discrete-continuous depth estimation from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 716–723.
69. Li, B.; Shen, C.; Dai, Y.; Van Den Hengel, A.; He, M. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1119–1127.
70. Zhao, W.; Liu, S.; Shu, Y.; Liu, Y.-J. Towards better generalization: Joint depth-pose learning without posenet. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9151–9161.
71. Zhang, Y.; Gong, M.; Li, J.; Zhang, M.; Jiang, F.; Zhao, H. Self-supervised monocular depth estimation with multiscale perception. *IEEE Trans. Image Process.* **2022**, *31*, 3251–3266. [CrossRef]
72. Song, X.; Hu, H.; Liang, L.; Shi, W.; Xie, G.; Lu, X.; Hei, X. Unsupervised monocular estimation of depth and visual odometry using attention and depth-pose consistency loss. *IEEE Trans. Multimed.* **2023**, *26*, 3517–3529. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.