remote sensing

MDPI

Technical Note

Geometric Prior-Guided Self-Supervised Learning for Multi-View Stereo

Liman Liu [1], Fenghao Zhang [1], Wanjuan Su [2,*], Yuhang Qi [2] and Wenbing Tao [2]

[1] School of Biomedical Engineering, South-Central Minzu University, Wuhan 430074, China
[2] National Key Laboratory of Science and Technology on Multi-Spectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
* Correspondence: suwanjuan@hust.edu.cn

Abstract: Recently, self-supervised multi-view stereo (MVS) methods, which are dependent primarily on optimizing networks using photometric consistency, have made clear progress. However, the difference in lighting between different views and reflective objects in the scene can make photometric consistency unreliable. To address this issue, a geometric prior-guided multi-view stereo (GP-MVS) for self-supervised learning is proposed, which exploits the geometric prior from the input data to obtain high-quality depth pseudo-labels. Specifically, two types of pseudo-labels for self-supervised MVS are proposed, based on the structure-from-motion (SfM) and traditional MVS methods. One converts the sparse points of SfM into sparse depth maps and combines the depth maps with spatial smoothness constraints to obtain a sparse prior loss. The other generates initial depth maps for semi-dense depth pseudo-labels using the traditional MVS, and applies a geometric consistency check to filter the wrong depth in the initial depth maps. We conducted extensive experiments on the DTU and Tanks and Temples datasets, which demonstrate that our method achieves state-of-the-art performance compared to existing unsupervised/self-supervised approaches, and even performs on par with traditional and supervised approaches.

Keywords: multi-view stereo; depth estimation; self-supervised learning

check for updates

Citation: Liu, L.; Zhang, F.; Su, W.; Qi, Y.; Tao, W. Geometric Prior-Guided Self-Supervised Learning for Multi-View Stereo. *Remote Sens.* **2023**, *15*, 2109. https://doi.org/10.3390/rs15082109

Academic Editor: John Trinder

Received: 20 March 2023
Revised: 10 April 2023
Accepted: 14 April 2023
Published: 17 April 2023

1. Introduction

Multi-view stereo (MVS) aims to generate a 3D model of a scene using a set of images with known poses, which has various applications in augmented reality, virtual reality, robotics, remote sensing, and more [1,2]. In the past years, the traditional MVS methods, such as MVE [3], OpenMVS [4], and COLMAP [5,6] have developed rapidly. Recently, the introduction of deep learning has allowed supervised MVS methods to outperform these traditional methods. Benefiting from the powerful feature representation ability, the learning-based MVS methods can efficiently reconstruct more complete 3D scenes [7]. Learning-based methods, however, require a significant quantity of large-scale 3D labeled data for training. This is difficult to obtain due to the challenges associated with creating 3D annotations [8,9], which generally involve capturing multiple synchronized images and depth sensors.

To solve the dependence on 3D annotated data, several unsupervised/self-supervised methods have been proposed [10–14]. These methods generally use the photometric consistency loss as the main loss, which measures the color consistency of original images and reconstructed images, based on estimated depth maps. In essence, it is assumed that the objects satisfy photometric consistency in different perspectives, that is, the projection of identical 3D scene points on different views conforms to color consistency. However, in the real world, due to the different lighting conditions from different perspectives and the reflection and occlusion problems in some areas, the assumption of photometric consistency sometimes does not hold. As a result, the photometric consistency loss is not always reliable. To this end, some self-supervised MVS methods, which [14,15] leverage pseudo-labels to solve the ambiguity of photometric consistency loss, have been proposed. These methods [14,15] usually adopt a two-stage training strategy. First, the network is primarily trained based on photometric consistency loss, resulting in the generation of an initial depth map; second, the pseudo-labels are generated by refining the initial depth map.

Although self-supervised methods based on pseudo-labels have achieved comparable performance to supervised methods, the process of generating pseudo-labels is very cumbersome and time-consuming, which is not conducive to practical application. To address this issue, we propose the geometric prior-guided multi-view stereo (GP-MVS) approach for self-supervised learning. The GP-MVS method uses geometry priors to efficiently generate high-quality depth pseudo-labels for self-supervised MVS. Specifically, we propose two types of depth pseudo-labels, sparse and semi-dense, based on the geometry information of the 3D scene. For the sparse labels, we use structure-from-motion (SfM) [5] to obtain sparse points, and convert them into depth maps as pseudo-labels. We add spatial smoothness constraints as supervision with the sparse labels to improve performance. For the semi-dense labels, we employ the traditional MVS method COLMAP [6] to produce the initial depth maps. We then apply geometric consistency constraints to remove outliers from these maps. As a result, we obtain high-quality pseudo-labels by combining the geometric priors, which can effectively avoid mis-estimation due to unreliable photometric consistency. From the experimental results, we can see that our method demonstrates exceptional performance when compared to other self-supervised methods, and is even comparable with some of the top supervised methods.

The key contributions of this work are as follows:

(1) An efficient geometric prior-guided self-supervised learning framework for MVS is proposed.

(2) A sparse prior loss, that combines sparse depth pseudo-labels from the SfM and the spatial smoothness constraint, is introduced, to better deal with depth discontinuities under sparse supervision.

(3) A semi-dense depth pseudo-label from the initial depth map estimated by COLMAP and geometric consistency is applied, to remove outliers caused by unreliable photometric consistency.

2. Related Work

2.1. Traditional MVS

Traditional MVS methods are able to be categorized into three groups according to how they represent the 3D scene: point cloud-based [16,17], volumetric-based [18], and depth map-based methods [6,19–25]. The first class of methods usually adopts the propagation strategy for matched keypoints, to gradually densify the reconstruction. However, due to the sequential propagation strategy, these methods are difficult to parallelize. The second class of methods represents the 3D space as regular voxels and determines the proximity of each voxel to the surface. These methods usually have high memory consumption, due to the voxel representation. The third class of methods separates the problem into a depth map estimation and depth map fusion, which are easy to parallelize and convert to a point cloud representation.

Depth map-based MVS methods can be implemented using various software packages, such as Multi-View Environment (MVE) [3], which offers end-to-end reconstruction capabilities including SfM, MVS, surface reconstruction, and texturing. Gipuma [21], COLMAP [6], ACMM [23], DP-MVS [24], and PatchMatch MVS [25] are PatchMatch-based [26] MVS methods. COLMAP employs geometric priors and photometric consistency to estimate surface normals and depth maps. ACMH [23] introduces an adaptive checkerboard sampling strategy to improve the efficiency of the PatchMatch-based method, ACMM further uses multi-scale geometric consistency based on ACMH, to improve the robustness of the method. In this paper, we adopt the widely used COLMAP to generate pseudo-labels.

2.2. Learning-Based MVS

Learning-based MVS methods have recently begun to shown great potential. MVSNet [27] presented an end-to-end pipeline, which estimated depth by building a 3D cost volume and using 3D CNN to regularize and regress the initial depth map. Following this, most learning-based methods [28–30] have mainly followed the pipeline of MVSNet. Some methods [31–33] leverage the RNN to regularize the cost volume sequentially, reducing memory overhead while increasing inference time. To both reduce time and memory consumption, the authors in [1,34–36] adopt the coarse-to-fine strategy. They estimate coarse estimates first and then make accurate estimates based on the previous stage's results. These methods achieve high-accuracy and high-resolution estimates, with acceptable memory and time cost. Based on the coarse-to-fine architecture, MVSFormer [37] introduced the pretrained ViT enhanced multi-view feature extraction network, which can learn more reliable feature representations, benefiting from informative priors from ViT. In this paper, CasMVSNet [35] is used as the backbone.

Supervised learning methods rely on hard-to-obtain 3D annotations. Thus, researchers began to focus on unsupervised/self-supervised methods [10–14]. UnsupMVS [10] was the first learning-based network that solved the MVS problem without ground-truth training data, which relies on the photometric consistency between multiple views. MVS^2 [11] adopts the geometric consistency between multiple views. M^{3}VSNet [12] extracts features with more semantic information by using a pretrained VGG network, and optimizes the initial depth map with normal–depth consistency. JDACS [13] introduces a self-supervised MVS framework based on co-segmentation and data-augmentation. Self-sup CVP-MVSNet [14] uses a two-stage training strategy, where the initial depth maps are estimated based on photometric consistency, followed by depth map refinement from high-resolution images and neighboring views. U-MVS [15] uses the correspondence information provided by optical flows and uncertainty maps to handle wrong supervision in the foreground and background, respectively. These methods heavily depend on photometric consistency, that is ambiguous in real 3D scenes. To overcome this, we consider leveraging traditional methods to generate pseudo-labels for self-learning.

3. Method

Our objective is to produce accurate and reliable pseudo-labels to facilitate self-supervised learning of MVS. In this section, we first analyze the wrong supervision caused by photometric consistency loss, then we present the two proposed kinds of pseudo-labels. The first type is sparse pseudo-labels, which are obtained by generating sparse points from SfM and then converting them into sparse depth maps. The second type is semi-dense pseudo-labels, which are generated using the traditional MVS method to estimate initial depth maps, followed by filtering out the outliers using geometric consistency.

The overall geometric prior-guided self-supervised learning framework is depicted in Figure 1. At the top of the figure, are the two pseudo-labels proposed for network training, which will be given in Sections 3.2 and 3.3. The bottom part shows a sketch of a learning-based MVS network.

3.1. Photometric Consistency Loss Revisited

The photometric consistency loss, $\mathcal{L}_{photo}$, measures the resemblance between the source image, $\mathcal{I}_{ref}^{j}$, projected to the reference view, according to the estimated depth maps and the reference image, $\mathcal{I}_{ref}$:

$$\mathcal{L}_{photo} = \sum_{j=1}^{\mathcal{N}} ||(\mathcal{I}_{ref} - \mathcal{I}_{ref}^{j}) \bigodot \mathcal{M}_j||_2 + ||(\bigtriangledown\mathcal{I}_{ref} - \bigtriangledown\mathcal{I}_{ref}^{j}) \bigodot \mathcal{M}_j||_2 \quad (1)$$

where $\bigtriangledown$ represents the gradient, and $\mathcal{M}_j$ is the effective area of the image.

Although the photometric consistency loss between different views can serve as supervision for self-supervised learning, there are still two issues that remain unresolved. (1) As shown in Figure 2, in a real scene, there is often interference from reflective surfaces, object occlusion, or other factors, and the corresponding points in different perspectives do not always meet the conditions of photometric consistency; (2) the supervision of the background areas in the DTU dataset is invalid. Specifically, there are invisible areas between different views, thus, the reconstructed image, $\mathcal{I}^{j}_{ref}$, usually contains invalid areas, and using photometric consistency loss will introduce large errors.

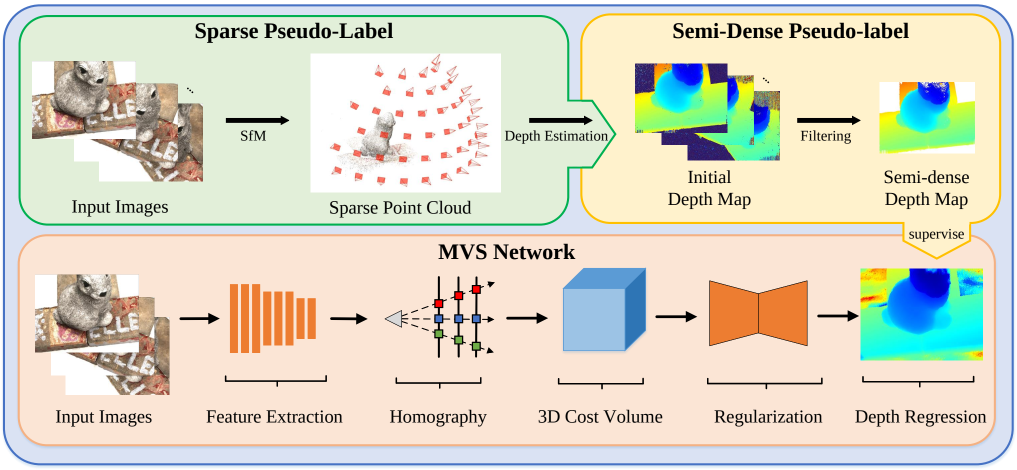


Figure 1. Geometric prior-guided self-supervised learning MVS framework. We generate the sparse and semi-dense pseudo-labels by using SfM and the traditional MVS method. These labels are then used to supervise the training of our MVS network.

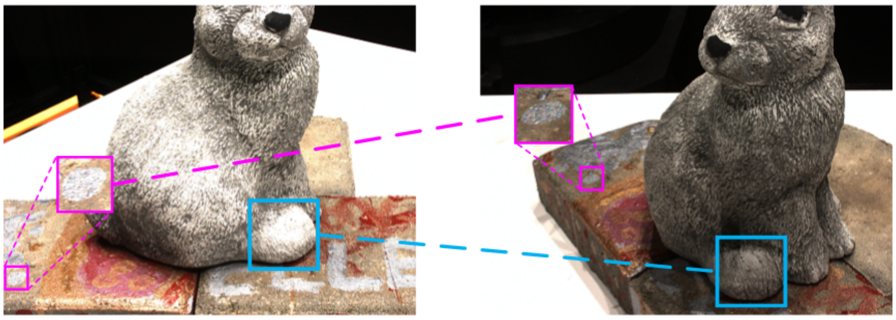


Figure 2. Ambiguity when adopting photometric consistency.

To solve these issues, we propose two pseudo-labels as supervision. In Figure 1, the sparse pseudo-label is located in the top left, while the semi-dense pseudo-label can be found in the top right. The pseudo-label-based self-supervised MVS framework can learn 3D information well and efficiently, even under reflective surfaces, object occlusion, and illumination changes.

3.2. Sparse Pseudo-Label

This section covers the generation of sparse depth map pseudo-labels. SfM [5], as a pre-step for MVS, aims to predict camera parameters of input images and 3D sparse point clouds of the scene. By triangulating feature points that match across multiple images, a set of sparse 3D points is obtained. These points are then optimized through bundle adjustment and outlier filtering, ensuring that the remaining sparse points are sufficiently reliable. Figure 3 shows the generation of sparse pseudo-labels, where only the white points in the depth map contain sparse prior information. Specifically, we generate sparse pseudo-labels based on the sparse point clouds P_{world}, and the camera parameters $\{K, R, t\}$

from SfM. For visible sparse point P^i_{world} in view i, we transform the world coordinates into camera coordinates P^i_{cam}, with the extrinsic $\{R_i, t_i\}$:

$$P^i_{cam} = R_i \cdot P^i_{world} + t_i \tag{2}$$

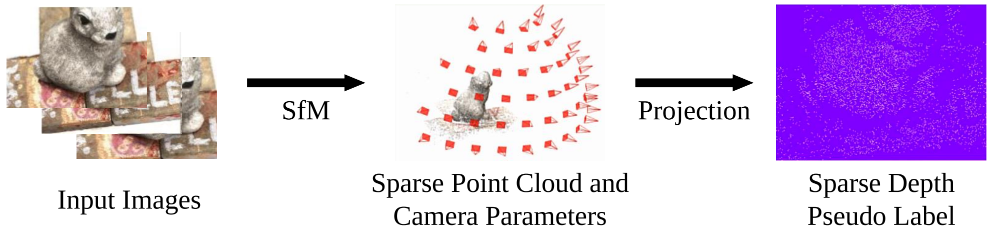


Figure 3. Generation of sparse pseudo-label.

Then, we project the sparse 3D points (camera coordinates) to the 2D image. For point (x, y, z) in a sparse point cloud, we get the projected point (u, v), with the intrinsics:

$$\begin{cases} u &= f_x \frac{x}{z} + c_x \\ v &= f_y \frac{y}{z} + c_y \\ d &= z \end{cases} \tag{3}$$

where (f_x, f_y) and (c_x, c_y) from K_i are the pixel focal length and the principal point, respectively. z denotes the depth d of (u, v). For those points without prior depth values, we set their depth as 0.

The sparse depth map only provides supervision for some pixels in the estimate. Therefore, we add the depth smoothing loss. The aims of the depth smoothing loss are to make the gradient of the estimate change smoothly and allow discontinuities in depth with large color changes. The depth smoothing loss considers variations in gradients of the input image:

$$\mathcal{L}_{smooth} = \sum_p |\bigtriangledown \mathcal{D}(p)|^{\mathcal{T}} \cdot e^{-|\bigtriangledown \mathcal{I}(P)|} \tag{4}$$

where p is the pixel in the depth map $\mathcal{D}$ and the image $\mathcal{I}$, and $\bigtriangledown\mathcal{D}$ is the gradient of the estimate. Thus, the sparse prior loss $\mathcal{L}_{sparse}$ that we adopt when training the network is:

$$\mathcal{L}_{sparse} = \sum_{s=1}^{S} \lambda_s (\mathcal{L}_1 + \mu \cdot \mathcal{L}_{smooth}) \tag{5}$$

where $\mathcal{L}_1$ represents the loss between the sparse pseudo-label and estimate. μ is the weight parameters of the two losses in our training with a sparse pseudo-label, μ is empirically set to 0.1 [13]. λ_s is the weight coefficient of the $\mathcal{L}_1$ in different stages.

3.3. Semi-Dense Pseudo-Label

The sparse pseudo-label can only provide supervision for a few points in the estimated depth maps, which can constrain the network's learning capability. Therefore, we consider generating a pseudo-label based on the traditional MVS method, to obtain more dense pseudo-labels. COLMAP [6] is a widely used method for 3D reconstruction, which performs pixel-wise normal and depth estimation based on geometric and photometric consistency. However, for weak textures and background areas, the reconstruction results of COLMAP are generally not reliable, this is due to the ambiguity of photometric consistency. To ensure the production of dependable pseudo-labels for self-supervised learning, we initially employ COLMAP to generate a preliminary depth map and then utilize multi-view geometric consistency to eliminate any outliers.

The initial depth map produced by COLMAP undergoes a filtering process that involves checking for geometric consistency through depth reprojection error, as shown in

Figure 4. To be specific, the image i and j are related by their relative position represented by the matrix $[\boldsymbol{R}ij|\boldsymbol{t}ij]$. The estimated depth of point p in the reference view is denoted as d_p. Back-projecting the point p into 3D space based on d_p, X_p is obtained. Projecting X_p to the source image, gives the projected pixel q of the source view. Back-projecting the q in the source image based on its depth estimate d_q to 3D space, gives the point X_q. Projecting X_q to the reference image gives the projected pixel coordinates $q^{'}$. The coordinate reprojection error is expressed as $||p - p^{'}||_2$. Similarly, the relative depth reprojection error is expressed as $\frac{||d^{'}_p - d_p||_1}{d_p}$.

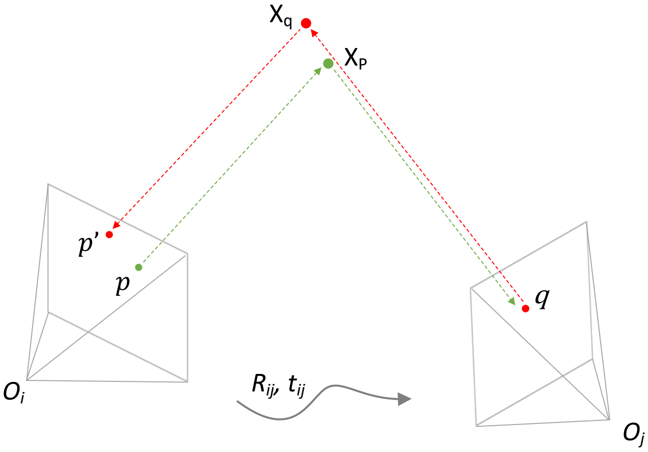


Figure 4. Cross-view geometric consistency.

We define a criterion $c(\cdot)$ to determine whether the estimated depth d_p of pixel p satisfies the cross-view geometric consistency, which comprehensively considers the coordinate reprojection error and relative depth reprojection error of the depth map. We consider the d_p to be consistent between the two views if the following equation is satisfied:

$$c(p) = \begin{cases} 1, \text{ if } ||p - p^{'}||_2 < \alpha \text{ and } \frac{||d^{'}_p - d_p||_1}{d_p} < \beta \\ 0, \text{ otherwise} \end{cases} \tag{6}$$

where α and β are empirically set to 1 and 0.01 based on the geometric consistency used in the previous method [27].

The initial depth map, estimated based on the traditional geometric method, is denoted as $\{\boldsymbol{\mathcal{D}}_{pm}|\boldsymbol{\mathcal{D}}_{pm} \in R^{h \times w}\}_{i=0}^{N}$. For the p in the reference image, there are $N - 1$ source images for the multi-view geometric consistency check, and we can obtain $N - 1$ pixels reprojected to the reference image. If the reprojected depth values are consistent for at least n_{min} views, i.e., $\sum_{i=1}^{N-1} c(p) \geq n_{min}$, then the estimate is considered dependable; n_{min} represents the minimum number of views necessary to achieve depth consistency. The retained high-confidence depth map is denoted as $\{\boldsymbol{\mathcal{D}}^{'}_{pm}|\boldsymbol{\mathcal{D}}^{'}_{pm} \in R^{h \times w}\}_{i=0}^{N}$, which is the semi-dense pseudo-label used for model training.

As shown in Figure 5, after the cross-view geometric consistency check, the erroneous background area in the depth map is filtered basically, while the depth estimation in the foreground part is retained. The multi-view geometric consistency check avoids invalid supervision of the background area, which is more conducive to the training of the network model.

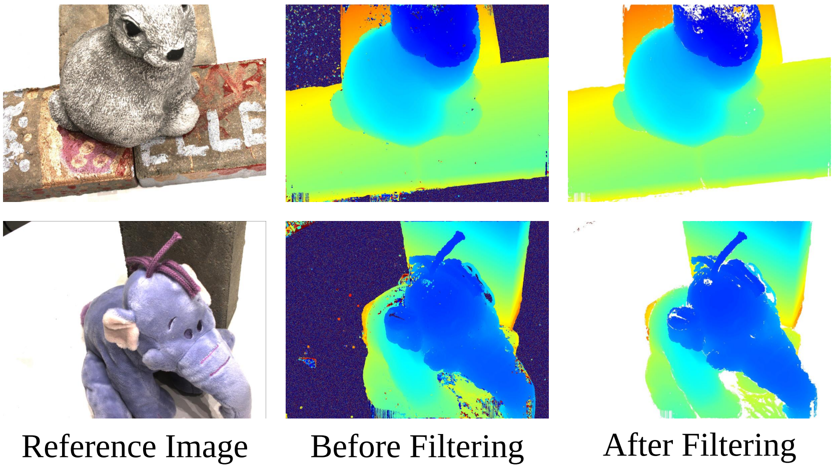


Figure 5. Semi-dense pseudo-label with cross-view geometric consistency.

We adopt the semi-dense loss $\mathcal{L}_{semi-dense}$ when training the network with semi-dense pseudo-labels:

$$\mathcal{L}_{semi-dense} = \sum_{l=0}^{\mathcal{L}} \sum_{p \in \Omega_{valid}} \lambda_l \cdot ||\boldsymbol{\mathcal{D}}^l(p) - \boldsymbol{\mathcal{D}}^l_{semi-dense}(p)||_1 \tag{7}$$

we follow the multi-stage training strategy of the backbone network [35], where $\boldsymbol{\mathcal{D}}^l(p)$ is the estimate of stage l, and $\boldsymbol{\mathcal{D}}^l_{semi-dense}(p)$ is the pseudo-label. Ω_{valid} denotes valid pixels in the pseudo-label. λ_l is the weight of the loss items in different stages.

3.4. Geometric Prior-Guided Multi-View Stereo Network

CasMVSNet [35] is used as our baseline model, and we apply the proposed sparse prior loss or semi-dense loss to supervise the network during training. The coarse-to-fine strategy is utilized by CasMVSNet [35] for estimating high-resolution depth maps. It first uses a weight-sharing feature pyramid network [38] to extract multi-scale features $\{F_i^s\}_{i=0}^{N-1}$ $(s = 1, 2, 3)$ from all input images, with resolution $H \times W$. For each scale, the features are then warped into fronto-parallel planes of the reference view, using differentiable homography [27], to obtain $N - 1$ feature volumes. By calculating the variance-based similarity, the feature volumes are combined to construct the 3D cost volume. Subsequently, the raw cost volume is regularized using a 3D UNet, resulting in a pixel-wise depth probability distribution. From this distribution, the D_0^1 is obtained by taking the expectation value. Finally, the depth map D_0^3, with resolution $H \times W$, can be obtained by gradually decreasing the depth sampling range and the depth sampling number of cost volumes, according to the predictions of previous stages.

4. Experiments

In this section, the performance of the GP-MVS framework is evaluated on the DTU [39] and Tanks and Temples benchmark [40]. We begin by describing these datasets and providing implementation details. Subsequently, we present the benchmarking process carried out on the aforementioned datasets. Finally, an ablation study is presented, to showcase the benefits of utilizing the proposed pseudo-labels.

4.1. Datasets and Implementation Details

4.1.1. Datasets

DTU is a dataset that comprises over 100 indoor scenes captured in a laboratory environment and featuring 7 distinct lighting conditions. Each scene consists of 39 or 64 images. We adopt the sparse prior generation process proposed in Section 3.2, and use the camera projection transformation to obtain the sparse depth map pseudo-labels. For semi-dense depth map pseudo-labels, as presented in Section 3.3, the initial depth maps are generated using COLMAP [5], and then we employ geometric consistency to remove any outliers.

We use metrics of mean accuracy, mean completeness, overall score, and the 0.5 mm F score for this dataset. These metrics are given by Equations (8)–(11), respectively.

$$\text{Acc.} = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} |e_{\mathbf{r} \to \mathcal{G}}| \tag{8}$$

where $e_{\mathbf{r} \to \mathcal{G}} = \min_{\mathbf{g} \in \mathcal{G}} ||\mathbf{r} - \mathbf{g}||$, $e_{\mathbf{r} \to \mathcal{G}}$ measures the distance from each point $\mathbf{r}$ in the reconstructed point cloud $\mathcal{R}$ to the ground-truth point cloud $\mathcal{G}$.

$$\text{Comp.} = \frac{1}{|\mathcal{G}|} \sum_{\mathbf{g} \in \mathcal{G}} |e_{\mathbf{g} \to \mathcal{R}}| \tag{9}$$

where $e_{\mathbf{g} \to \mathcal{R}} = \min_{\mathbf{r} \in \mathcal{R}} ||\mathbf{g} - \mathbf{r}||$, $e_{\mathbf{g} \to \mathcal{R}}$ measures the distance from each point $\mathbf{r}$ in the ground truth point cloud $\mathcal{R}$ to the reconstructed point cloud $\mathcal{R}$

$$\text{Overall} = \frac{\text{Acc.} + \text{Comp.}}{2} \tag{10}$$

$$F_1(\tau) = \frac{2P(\tau)R(\tau)}{P(\tau) + R(\tau)} \tag{11}$$

where the precision $P(\tau)$, measures the percentage of the number of reconstructed point clouds that fall within a given distance threshold τ, to the total number of reconstructed point clouds, the recall $R(\tau)$, measures the percentage of the number of ground-truth point clouds to the total number of ground-truth point clouds at a given distance threshold τ. The $P(\tau)$ and $R(\tau)$ are given by Equations (12) and (13), respectively.

$$P(\tau) = \frac{100}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} \left[e_{\mathbf{r} \to \mathcal{G}} < \tau \right] \tag{12}$$

$$R(\tau) = \frac{100}{|\mathcal{G}|} \sum_{\mathbf{g} \in \mathcal{G}} \left[e_{\mathbf{g} \to \mathcal{R}} < \tau \right] \tag{13}$$

Tanks and Temples is a dataset consisting of indoor and outdoor scenes captured in realistic environments, and it includes the *intermediate set* and the *advanced set*. Our method is evaluated for its generalization performance using this dataset, with the F score serving as the primary metric.

4.1.2. Implementation Details

Training. We used generated pseudo-labels to supervise the backbone network [35,37] on the training set of DTU. Similar to CasMVSNet [35], the high-resolution input images and pseudo-label depth maps, with resolution 1600 × 1200, were down-sampled and center-cropped to obtain image and depth maps with a resolution of 640 × 512 when training. PyTorch was used to implement the network, and a total of 16 epochs were used to train the network with the Adam optimizer. The initial learning rate of 0.001 was halved at the 10th, 12th, and 14th epochs, to prevent the network training from falling into a local optimum. Following CasMVSNet [35], we employed 48, 32, and 8 hypothesis planes at each stage, and the λ_l / λ_s for each stage was set to 0.5, 1.0, and 2.0.

Depth Fusion. After generating depth maps for all reference views, we fused them to create a dense 3D point cloud model, using a similar approach to previous work [35]. We started by filtering out unreliable depth values with low confidence, using the probability map generated by the network. We then applied the geometric consistency check in Equation (6) to verify the depth maps, further filtering out unreliable depths. The final depth estimation for each pixel was obtained by taking the average over all reprojected depths. Finally, we directly reprojected the filtered depth maps into space to generate the 3D point cloud.

4.2. Benchmark Performance

Results on DTU. Our method's performance is assessed on the DTU test set using the network trained on the DTU training set. As for the supervised backbone CasMVSNet [35], the resolution of the input is resized to 1152 × 864 and five images are used for depth map prediction (one reference image and four source images).

We evaluate the point clouds reconstructed by our method using the overall score. As summarized in Figure 6, our approach that utilizes semi-dense depth pseudo-labels delivers performance that is comparable to self-supervised learning approaches and even outperforms the supervised MVSNet [27], R-MVSNet [31], and Point-MVSNet [34], and the result is roughly on par with those of CasMVSNet [35] and CVP-MVSNet [36].

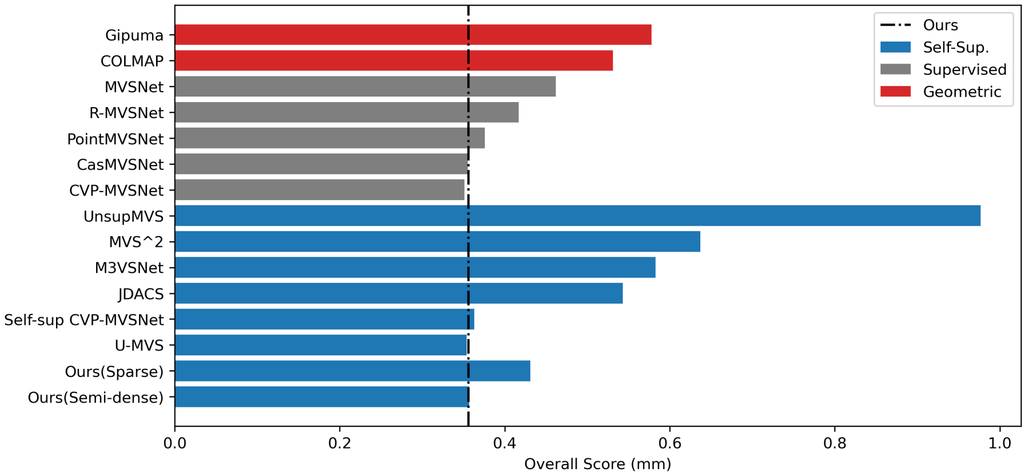


Figure 6. Comparison between SOTA MVS methods on DTU dataset (lower is better).

The quantitative results of various self-supervised MVS methods, including the proposed pseudo-label based method, are presented in Table 1. Our methods (trained with sparse pseudo-labels or semi-dense pseudo-labels) perform better than UnsupMVS [10], MVS^2 [11], M^{3}VSNet [12], and JDACS [13]. The model trained with our semi-dense depth map pseudo-labels (semi-dense) achieved comparable performance compared with Self-sup CVP-MVSNet [14] and U-MVS [15]. Note that the pseudo-labels generation process of Self-sup CVP-MVSNet and U-MVS is much more complicated compared with that of our method. For Self-sup CVP-MVSNet, after obtaining the initial depth map from the unsupervised model, an iterative refinement process is performed to obtain the pseudo-labels, which involves several steps, such as initial depth estimation from a high-resolution image, consistency check-based filtering for estimates, and fusion of the depth from multiple views, to obtain final pseudo-labels. For U-MVS, it uses the pretrained unsupervised model based on the uncertainty to generate pseudo-labels, which requires sampling up to 20 times to obtain reliable uncertainty maps for depth filtering.

Table 1. Quantitative results of our method against self-supervised MVS methods on the DTU dataset (lower is better). The best results are in bold, while the second ones are underlined.

Method	Acc. ↓	Comp. ↓	Overall ↓
UnsupMVS [10]	0.881	1.073	0.977
MVS^2 [11]	0.760	0.515	0.637
M^{3}VSNet [12]	0.636	0.531	0.583
JDACS [13]	0.571	0.515	0.543
Self-sup CVP-MVSNet [14]	**0.308**	0.418	0.363
U-MVS [15]	0.354	0.354	**0.354**
Ours (sparse)	0.419	0.443	0.431
Ours (semi-dense)	0.399	**0.316**	0.357

Table 2 showcases a comparison between the proposed methods and traditional/supervised MVS methods. Our approach surpasses the traditional approaches Gipuma [21] and COLMAP [6]. MVSFormer [37] has been improved on the basis of CasMVSNet [35], achieving the best performance of the supervised methods on the DTU dataset. Our self-supervised method is comparable to the supervised multi-scale MVS network CVP-MVSNet [14], and the point cloud reconstructed by our method has better completeness. We also compare the self-supervised approach proposed with the backbone network CasMVSNet. Table 2 presents a numerical evaluation of our approach compared to the CasMVSNet on the DTU dataset. Our approach shows slightly lower quantitative results but the qualitative results, as shown in Figure 7, suggest that our approach can reconstruct 3D point clouds with high accuracy, especially in capturing local details.

Table 2. Quantitative results of our approach against traditional and supervised MVS methods on the DTU dataset (lower is better). The best results are in bold, while the second ones are underlined.

Method	Classification	Acc. ↓	Comp. ↓	Overall ↓
Gipuma [21]	Traditional	**0.283**	0.873	0.578
COLMAP [6]	Traditional	0.401	0.661	0.531
MVSNet [27]	Supervised	0.396	0.527	0.462
R-MVSNet [31]	Supervised	0.383	0.452	0.417
Point-MVSNet [34]	Supervised	0.342	0.411	0.376
CVP-MVSNet [14]	Supervised	0.296	0.406	0.351
CasMVSNet [35]	Supervised	0.325	0.385	0.355
MVSFormer [37]	Supervised	0.327	**0.251**	**0.289**
Ours (semi-dense)	Self-supervised	0.399	0.316	0.357

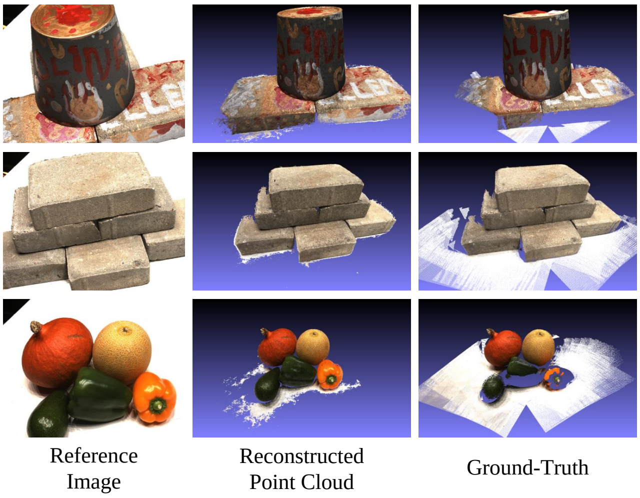


Figure 7. Qualitative results of our approach on the DTU dataset in terms of reconstructed point clouds.

Results on Tanks and Temples. To assess the generalization capability of the proposed methods, the models were trained on the DTU dataset and performed an evaluation on the Tanks and Temples dataset, without any fine-tuning. Specifically, five input images were used as an input, with a resolution of 1920 × 1056. As displayed in Table 3, our approach surpasses the traditional methods and supervised methods by a significant margin, which proves that the MVS network supervised with our proposed pseudo-label is effective. Additionally, Figure 8 illustrates the qualitative results of both subsets. The proposed method can reconstruct denser point clouds with more details, making them more visually appealing.

Table 3. The performance of our approach on the Tanks and Temples benchmark (intermediate set) with F score (%) (higher is better). The best results are in bold, while the second ones are underlined.

Method	Sup.	Mean	Family	Francis	Horse	Lighthouse	M60	Panther	Playground	Train
COLMAP [6]	-	42.14	50.41	22.25	25.63	56.43	44.83	46.97	48.53	42.04
MVSNet [27]	√	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69
MVSCRF [28]	√	45.73	59.83	30.60	29.93	51.15	50.61	51.45	52.60	39.68
CIDER [30]	√	46.76	56.79	32.39	29.89	54.67	53.46	53.51	50.48	42.85
R-MVSNet [31]	√	48.40	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38
Point-MVSNet [34]	√	48.27	61.79	41.15	34.20	50.79	51.97	50.85	52.38	43.06
CasMVSNet [35]	√	56.42	76.36	58.45	46.20	55.53	56.11	54.02	58.17	49.56
CVPMVSNet [36]	√	54.03	76.50	47.74	36.34	55.12	57.28	54.28	57.43	47.54
UCSNet [1]	√	54.83	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89
MVS^2 [11]	×	37.21	47.74	21.55	19.50	44.54	44.86	46.32	43.38	29.72
M^{3}VSNet [12]	×	37.67	47.74	24.38	18.74	44.42	43.45	44.95	47.39	30.31
JDACS [13]	×	45.48	66.62	38.25	36.11	46.12	46.66	45.25	47.69	37.16
Self-sup CVP-MVSNet [14]	×	56.54	76.35	49.06	43.04	**57.35**	**60.64**	**57.35**	**58.47**	50.06
U-MVS [15]	×	**57.15**	76.49	**60.04**	**49.20**	55.52	55.33	51.22	56.77	**52.63**
Ours	×	56.65	**77.32**	59.88	48.96	56.17	54.78	50.82	55.52	49.76

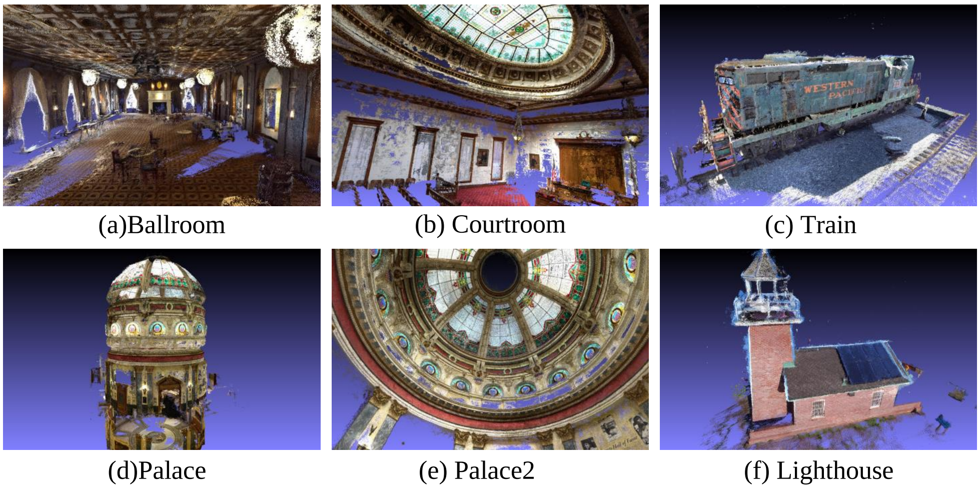
The advanced set of Tanks and Temples contains challenging scenes. Our approach demonstrates superior performance compared to other approaches in most evaluation metrics, as presented in Table 4. This proves that the proposed depth map pseudo-labels based on the geometry prior, can effectively capture the geometric information in the 3D scene. Due to overfitting on the DTU dataset, supervised methods, such as the backbone CasMVSNet, exhibit limited generalization performance. Thus, even though our method achieved slightly lower reconstruction performance on the DTU compared to the backbone network, the use of our proposed pseudo-labels has the potential to enhance the network's generalization ability. This proves that the MVS network supervised with our proposed pseudo-labels is effective.

Table 4. The performance of our approach on the Tanks and Temples benchmark (advanced set) with F score (%) (higher is better). The best results are in bold.

Method	Sup.	Mean	Auditorium	Ballroom	Courtroom	Museum	Palace	Temple
COLMAP [6]	-	27.24	16.02	25.23	**34.70**	41.51	18.05	27.94
R-MVSNet [31]	√	24.91	12.55	29.09	25.06	38.68	19.14	24.96
CIDER [30]	√	23.12	12.77	24.94	25.01	33.64	19.18	23.15
CasMVSNet [35]	√	31.12	19.81	38.46	29.10	43.87	27.36	28.11
U-MVS [15]	×	30.97	**22.79**	35.39	28.90	36.70	**28.77**	**33.25**
Ours	×	**32.68**	21.62	**40.41**	29.52	**46.79**	28.16	29.61

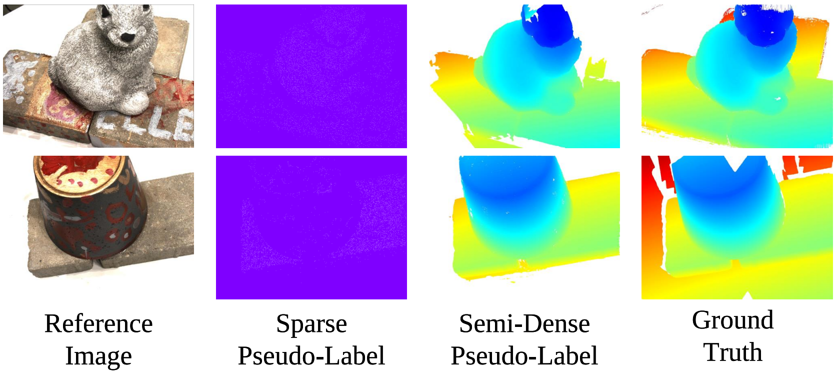
4.3. Ablation Study

Accuracy of pseudo-labels. Figure 9 shows the visualization of different pseudo-labels. The white dots in the sparse depth map are the pixel positions with sparse prior information. The sparse depth map can only describe the basic geometric structure of the 3D scene, focusing more on the rich texture part. The supervision of the semi-dense pseudo-label depth map in the foreground area is more complete. Upon comparison with the ground truth, it can be inferred that the foreground, using semi-dense pseudo-labels, is more complete, while removing false background estimates. We assess the accuracy of the network using various pseudo-labels as supervision on the DTU dataset, with depth prediction accuracy serving as the evaluation metric. In addition, we provide the density (means of percentage of labeled pixels in each image) of different pseudo-labels. Note that the density of the initial depth map without filtering is 100%.


(a)Ballroom (b) Courtroom (c) Train

(d)Palace (e) Palace2 (f) Lighthouse

Figure 8. Visualization of the reconstructed point clouds on the Tanks and Temples dataset.


Reference Image

Sparse Pseudo-Label

Semi-Dense Pseudo-Label

Ground Truth

Figure 9. Visualization of different pseudo-labels.

From Table 5, it can be concluded that the accuracy of the sparse depth map has already achieved a high accuracy (86% pixels of the sparse depth map are accurate within 2 mm). However, due to the few labeled points, the sparse pseudo-labels have certain limitations as supervision. The semi-dense pseudo-labels, after removing the wrong points, has the highest accuracy.

Table 5. Evaluation of different pseudo-labels. The best results are in bold.

Pseudo-Label	Acc_2 mm ↑	Acc_4 mm ↑	Acc_8 mm ↑	Density
Sparse depth map	86.74%	90.75%	93.23%	0.65%
Initial depth map	74.75%	79.17%	81.91%	–
Semi-dense depth map	**90.72%**	**93.86%**	**95.20%**	**64.78%**

Analysis of Different Supervisions. Table 6 reflects the accuracy of depth maps estimated by models trained under different supervision. The results show that the network trained with semi-dense depth map pseudo-labels achieves the second best accuracy, which is comparable to that of the supervised CasMVSNet, while outperforming the network based on photometric consistency loss and the sparse prior loss.

Table 6. Qualitative results of depth estimation on the DTU dataset (lower is better). The best results are in bold, while the second ones are underlined.

Method	Acc_2 mm ↑	Acc_4 mm ↑	Acc_8 mm ↑
CasMVSNet [35]	**69.90%**	**75.35%**	**78.80%**
Photometric consistency loss	65.31%	72.23%	76.35%
Sparse pseudo-labels	60.99%	67.70%	71.99%
Semi-dense pseudo-labels	<u>69.82%</u>	<u>74.69%</u>	<u>77.64%</u>

As shown in Figure 10, using photometric consistency loss as supervision, leads to noticeable errors at the boundaries. In contrast, using semi-dense pseudo-labels as the network's supervision, allows for more precise depth map predictions, especially at the border between foreground and background.

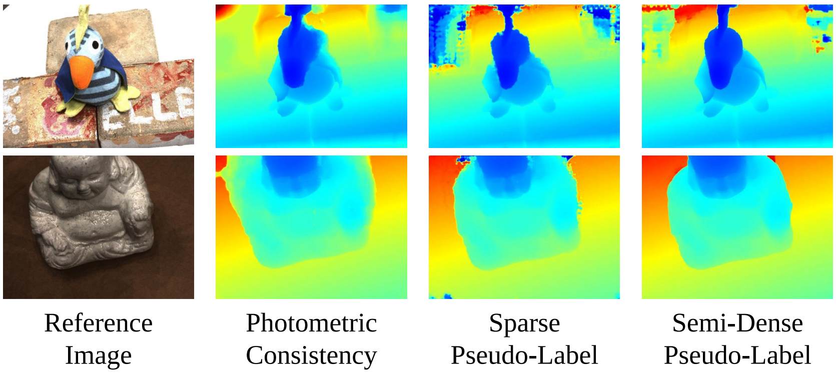


Figure 10. Quantitative results of depth estimation on the DTU dataset.

Table 7 shows the results of point clouds reconstructed by models with different supervisions. We use the overall and the F score under the 1 mm threshold as the evaluation metrics. Methods based on semi-dense pseudo-labels have the best quality.

Table 7. Qualitative results of point cloud reconstruction on the DTU dataset (lower is better). The best results are in bold.

Supervisory Signal	Acc. ↓	Comp. ↓	Overall ↓	F Score@1 mm ↑
Photometric consistency loss	0.441	0.335	0.388	83.04%
Sparse prior	0.419	0.443	0.431	81.62%
Semi-dense depth map	**0.399**	**0.316**	**0.357**	**85.98%**

By comparing the performance of different methods, we aimed to provide further evidence of the effectiveness of our self-supervised approach utilizing pseudo-labels. Figure 11 displays the reconstructed results of scan9, scan33, and scan49 in the DTU dataset. UnsupMVS [10] is an unsupervised MVS method based on photometric consistency loss. The self-supervised MVS method based on pseudo-labels produces denser 3D point clouds with more complete local details compared to other methods, as shown in Figure 11.

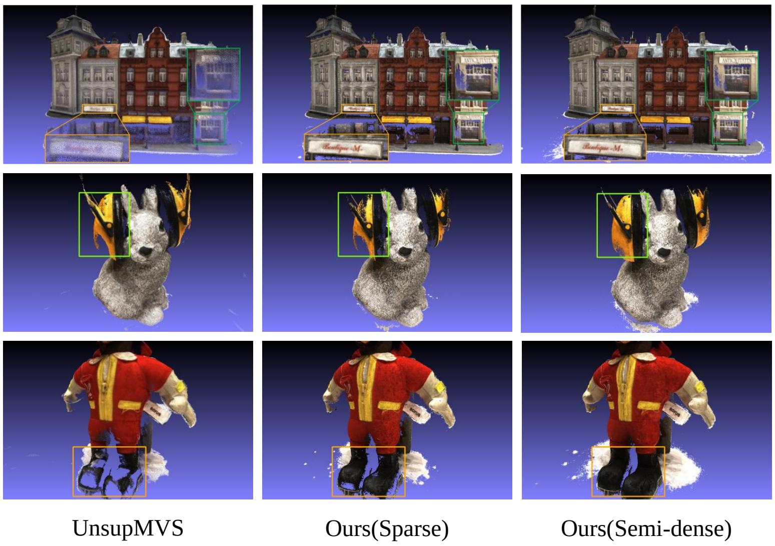


Figure 11. Quantitative results of point cloud reconstruction on the DTU dataset.

In addition, we conducted a comparison between the point clouds generated by our method and those obtained using traditional methods. The sparse point clouds were generated from the SfM described in Section 3.2. Table 8 demonstrates that although the sparse point cloud has an acceptable accuracy, its completeness is compromised, due to the sparse distribution of points. It should be noted, that our self-supervised methods using semi-dense pseudo-labels outperformed the traditional method COLMAP. In addition, using the dense depth map reconstructed by COLMAP as a pseudo-label for training, the accuracy is not only better than COLMAP itself, but also close to the best supervised learning method, and even stronger in generalization ability. These results highlight the strengths of our proposed pseudo-label approach.

Table 8. Comparison between traditional methods on the DTU dataset (lower is better). The best results are in bold.

Class	Acc. ↓	Comp. ↓	Overall ↓	F Score@1 mm ↑
Sparse Point Cloud [5]	0.452	3.450	1.951	11.75%
COLMAP [6]	0.401	0.661	0.531	76.61%
Ours(Sparse)	0.419	0.443	0.431	81.62%
Ours(Semi-dense)	**0.399**	**0.316**	**0.357**	**85.98%**

Statistical Analysis. To further show the effectiveness of the proposed semi-dense pseudo-labels, a statistical analysis based on the paired t-test is conducted for CasMVSNet [35] and CasMVSNet [35] combined with semi-dense pseudo-labels. The statistic t of the paired t-test is calculated as:

$$t = \frac{\bar{d} - d_0}{s_d / \sqrt{n}} \tag{14}$$

where $\bar{d}$ denotes the sample mean of differences, d_0 denotes the hypothesized population mean difference, s_d denotes the standard deviation of differences, and n denotes the sample size. The degrees of freedom $df = n - 1$. The p-value is determined by checking the corresponding threshold table based on the t statistic. Table 9 shows the results of the paired t-tests for CasMVSNet [35] and CasMVSNet with our semi-dense pseudo-labels on the DTU dataset and the Tanks and Temples dataset, the significance level α is set to 0.05. The p values for the DTU and intermediate subsets are 0.8260 and 0.2794, respectively, indicating no significant difference between the experimental results of CasMVSNet and CasMVSNet with our semi-dense pseudo-labels on these datasets. This suggests that

CasMVSNet with our semi-dense pseudo-labels is competitive with CasMVSNet on these datasets. On the advanced subset, the p value is 0.0076, indicating a significant difference between the experimental results of CasMVSNet and CasMVSNet with our semi-dense pseudo-labels on this dataset. Therefore, our method outperforms CasMVSNet significantly on this dataset.

Table 9. The paired t-test results for CasMVSNet [35] with our semi-dense pseudo-labels on the DTU dataset and the Tanks and Temples dataset (significance level α = 0.05).

	DTU	Intermediate	Advanced
t	0.2226	1.1724	4.3204
df	21	7	5
p	0.8260	0.2794	0.0076

5. Discussion

Our network's success can be mainly attributed to the utilization of self-supervised multi-view stereo learning, guided by pseudo-labels. Our pseudo-label-guided method effectively avoids the ambiguity of the breadth of the image reconstruction loss monitoring signal, resulting in a trained network model with stronger generalization performance. However, our work also has some limitations. For instance, the sparse depth supervised network model can only describe the basic structure of the scene, due to insufficient monitoring signals. Additionally, the depth map output from the network model based on sparse prior depth map supervision may not accurately estimate finer details. While using semi-dense pseudo-labels as a supervisory signal can achieve better performance than using sparse pseudo-labels, it is limited by the inherent difficulties of traditional MVS methods in estimating reliable depth in some areas such as occlusion, textureless, and non-Lambertian surfaces, where it cannot provide a supervisory signal for the network.

6. Conclusions

In this paper, a geometric prior-guided MVS framework for self-supervised learning is proposed. Unlike other methods that use photometric consistency loss as supervision, we propose two pseudo-labels: sparse depth map and semi-dense depth map. This can effectively address issues arising from illumination changes across images and inadequate supervision in the background area. Specifically, we employed SfM to obtain a sparse 3D point cloud, and produced depth maps using the traditional MVS method. After post-processing, we obtain two high-quality pseudo-labels, namely sparse and semi-dense. By using these pseudo-labels, our approach outperforms self-supervised methods and performs similarly to supervised learning frameworks. The sparse point cloud mentioned in this paper is low-level information, which only contains geometric information of the 3D scene, while the input image contains more semantic information. Our future work will consider how to combine an image's semantic information to assist MVS.

Author Contributions: Methodology, F.Z., Y.Q. and L.L.; software, F.Z. and Y.Q.; validation, F.Z., W.S. and Y.Q.; writing—original draft preparation, F.Z. and Y.Q.; writing—review and editing, W.S. and L.L.; visualization, F.Z. and Y.Q.; supervision, L.L. and W.T.; funding acquisition, L.L. and W.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (grants 61976227 and 62176096) and in part by the Natural Science Foundation of Hubei Province under grant 2020CFA025.

Data Availability Statement: The DTU dataset can be accessed at https://roboimagedata.compute.dtu.dk/ (accessed on 23 April 2016), and the Tanks and Temples benchmark can be accessed at https://www.tanksandtemples.org/ (accessed on 20 July 2017).

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

References

1. Cheng, S.; Xu, Z.; Zhu, S.; Li, Z.; Li, L.E.; Ramamoorthi, R.; Su, H. Deep stereo using adaptive thin volume representation with uncertainty awareness. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2524–2534.
2. Gonçalves, G.; Gonçalves, D.; Gómez-Gutiérrez, Á.; Andriolo, U.; Pérez-Alvárez, J.A. 3D reconstruction of coastal cliffs from fixed-wing and multi-rotor uas: Impact of sfm-mvs processing parameters, image redundancy and acquisition geometry. *Remote Sens.* **2021**, *13*, 1222. [CrossRef]
3. Fuhrmann, S.; Langguth, F.; Moehrle, N.; Waechter, M.; Goesele, M. MVE—An image-based reconstruction environment. *Comput. Graph.* **2015**, *53*, 44–53. [CrossRef]
4. Cernea, D. OpenMVS: Multi-View Stereo Reconstruction Library. 2020, Volume 5, p. 7. Available online: https://cdcseacave.github.io/openMVS (accessed on 20 May 2015).
5. 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.
6. Schönberger, J.L.; Zheng, E.; Frahm, J.M.; Pollefeys, M. Pixelwise view selection for unstructured multi-view stereo. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 501–518.
7. Ji, M.; Gall, J.; Zheng, H.; Liu, Y.; Fang, L. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2307–2315.
8. Zhong, Y.; Li, H.; Dai, Y. Open-world stereo video matching with deep rnn. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 101–116.
9. Zhang, X.; Zhao, Y.; Wang, H.; Zhai, H.; Sun, H.; Zheng, N. End-to-end learning of self-rectification and self-supervised disparity prediction for stereo vision. *Neurocomputing* **2022**, *494*, 308–319. [CrossRef]
10. Khot, T.; Agrawal, S.; Tulsiani, S.; Mertz, C.; Lucey, S.; Hebert, M. Learning unsupervised multi-view stereopsis via robust photometric consistency. *arXiv* **2019**, arXiv:1905.02706.
11. Dai, Y.; Zhu, Z.; Rao, Z.; Li, B. Mvs2: Deep unsupervised multi-view stereo with multi-view symmetry. In Proceedings of the 2019 International Conference on 3D Vision (3DV), Quebec, QC, Canada, 16–19 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–8.
12. Huang, B.; Yi, H.; Huang, C.; He, Y.; Liu, J.; Liu, X. M3VSNet: Unsupervised multi-metric multi-view stereo network. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 3163–3167.
13. Xu, H.; Zhou, Z.; Qiao, Y.; Kang, W.; Wu, Q. Self-supervised multi-view stereo via effective co-segmentation and data-augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; Volume 35, pp. 3030–3038.
14. 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.
15. Xu, H.; Zhou, Z.; Wang, Y.; Kang, W.; Sun, B.; Li, H.; Qiao, Y. Digging into uncertainty in self-supervised multi-view stereo. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 6078–6087.
16. Lhuillier, M.; Quan, L. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 418–433. [CrossRef] [PubMed]
17. Furukawa, Y.; Ponce, J. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1362–1376. [CrossRef] [PubMed]
18. Hane, C.; Zach, C.; Cohen, A.; Angst, R.; Pollefeys, M. Joint 3D scene reconstruction and class segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 97–104.
19. Shen, S. Accurate multiple view 3d reconstruction using patch-based stereo for large-scale scenes. *IEEE Trans. Image Process.* **2013**, *22*, 1901–1914. [CrossRef] [PubMed]
20. Zheng, E.; Dunn, E.; Jojic, V.; Frahm, J.M. Patchmatch based joint view selection and depthmap estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1510–1517.
21. Galliani, S.; Lasinger, K.; Schindler, K. Massively parallel multiview stereopsis by surface normal diffusion. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 873–881.
22. Li, Z.; Wang, K.; Meng, D.; Xu, C. Multi-view stereo via depth map fusion: A coordinate decent optimization method. *Neurocomputing* **2016**, *178*, 46–61. [CrossRef]
23. Xu, Q.; Tao, W. Multi-scale geometric consistency guided multi-view stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5483–5492.
24. Zhou, L.; Zhang, Z.; Jiang, H.; Sun, H.; Bao, H.; Zhang, G. DP-MVS: Detail Preserving Multi-View Surface Reconstruction of Large-Scale Scenes. *Remote Sens.* **2021**, *13*, 4569. [CrossRef]
25. Stathopoulou, E.K.; Battisti, R.; Cernea, D.; Remondino, F.; Georgopoulos, A. Semantically derived geometric constraints for MVS reconstruction of textureless areas. *Remote Sens.* **2021**, *13*, 1053. [CrossRef]
26. Bleyer, M.; Rhemann, C.; Rother, C. Patchmatch stereo-stereo matching with slanted support windows. In Proceedings of the BMVC, Dundee, UK, 29 August–2 September 2011; Volume 11, pp. 1–11.

27. Yao, Y.; Luo, Z.; Li, S.; Fang, T.; Quan, L. Mvsnet: Depth inference for unstructured multi-view stereo. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 767–783.
28. Xue, Y.; Chen, J.; Wan, W.; Huang, Y.; Yu, C.; Li, T.; Bao, J. Mvscrf: Learning multi-view stereo with conditional random fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4312–4321.
29. Luo, K.; Guan, T.; Ju, L.; Huang, H.; Luo, Y. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10452–10461.
30. Xu, Q.; Tao, W. Learning inverse depth regression for multi-view stereo with correlation cost volume. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12508–12515.
31. Yao, Y.; Luo, Z.; Li, S.; Shen, T.; Fang, T.; Quan, L. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5525–5534.
32. Yan, J.; Wei, Z.; Yi, H.; Ding, M.; Zhang, R.; Chen, Y.; Wang, G.; Tai, Y.W. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 674–689.
33. Wei, Z.; Zhu, Q.; Min, C.; Chen, Y.; Wang, G. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 6187–6196.
34. Chen, R.; Han, S.; Xu, J.; Su, H. Point-based multi-view stereo network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1538–1547.
35. Gu, X.; Fan, Z.; Zhu, S.; Dai, Z.; Tan, F.; Tan, P. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2495–2504.
36. Yang, J.; Mao, W.; Alvarez, J.M.; Liu, M. Cost volume pyramid based depth inference for multi-view stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4877–4886.
37. Cao, C.; Ren, X.; Fu, Y. MVSFormer: Multi-View Stereo by Learning Robust Image Features and Temperature-based Depth. *arXiv* **2023**, arXiv:2208.02541.
38. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
39. Aanæs, H.; Jensen, R.R.; Vogiatzis, G.; Tola, E.; Dahl, A.B. Large-scale data for multiple-view stereopsis. *Int. J. Comput. Vis.* **2016**, *120*, 153–168. [CrossRef]
40. Knapitsch, A.; Park, J.; Zhou, Q.Y.; Koltun, V. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graph. (ToG)* **2017**, *36*, 1–13. [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.