Just a Few Points are All You Need for Multi-view Stereo: A Novel Semi-supervised Learning Method for Multi-view Stereo

Taekyung Kim[1], Jaehoon Choi[2], Seokeon Choi[1], Dongki Jung[3], Changick Kim[1]
[1]Korea Advanced Institute of Science and Technology
[2]University of Maryland
[3]NAVER LABS
{tkkim93, seokeon, changick}@kaist.ac.kr, kevchoi@umd.edu, dongki.jung@naverlabs.com

Abstract

While learning-based multi-view stereo (MVS) methods have recently shown successful performances in quality and efficiency, limited MVS data hampers generalization to unseen environments. A simple solution is to generate various large-scale MVS datasets, but generating dense ground truth for 3D structure requires a huge amount of time and resources. On the other hand, if the reliance on dense ground truth is relaxed, MVS systems will generalize more smoothly to new environments. To this end, we first introduce a novel semi-supervised multi-view stereo framework called a Sparse Ground truth-based MVS Network (SGT-MVSNet) that can reliably reconstruct the 3D structures even with a few ground truth 3D points. Our strategy is to divide the accurate and erroneous regions and individually conquer them based on our observation that a probability map can separate these regions. We propose a self-supervision loss called the 3D Point Consistency Loss to enhance the 3D reconstruction performance, which forces the 3D points back-projected from the corresponding pixels by the predicted depth values to meet at the same 3D coordinates. Finally, we propagate these improved depth predictions toward edges and occlusions by the Coarse-to-fine Reliable Depth Propagation module. We generate the spare ground truth of the DTU dataset for evaluation and extensive experiments verify that our SGT-MVSNet outperforms the state-of-the-art MVS methods on the sparse ground truth setting. Moreover, our method shows comparable reconstruction results to the supervised MVS methods though we only used tens and hundreds of ground truth 3D points.

1. Introduction

Multi-view Stereo (MVS) has been an important problem in computer vision, which reconstructs dense 3D geometry from multi-view images. The industrial applicability of 3D reconstruction such as autonomous driving and robotics has attracted extensive research for decades. Recent MVS studies [15, 16, 1, 17, 11] successfully incorporate traditional approaches to the learning-based methods and improve 3D reconstruction quality under the blessing of the MVS datasets [7, 10]. However, contrary to such increasing dependence on datasets, there have been fundamental difficulties in collecting dense ground truth 3D structures, which eventually hamper the generalization to unseen do-

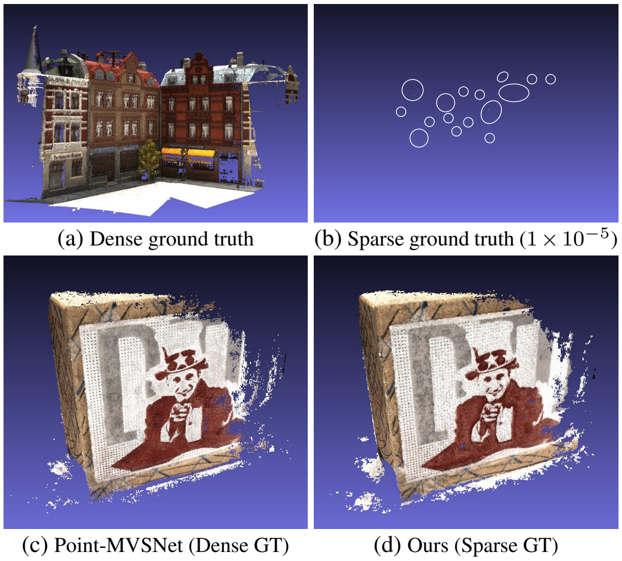

(a) Dense ground truth
(b) Sparse ground truth (1×10^{-5})
(c) Point-MVSNet (Dense GT)
(d) Ours (Sparse GT)

Figure 1: Visualization of a dense ground truth and our sparse ground truth of *scan14*, and multi-view reconstruction results of *scan13* of the DTU dataset [7] by Point-MVSNet [1] trained with dense ground truths and our SGT-MVSNet trained with sparse ground truths. This sparse ground truth generated by random sampling with 1×10^{-5} ratio contain around 30 to 40 3D points. Note that original dense ground truth 3D structures consist of approximately 3×10^6 points.

mains. Specifically, collecting an accurate and completed ground truth 3D structure generally takes several hours with a fixed active sensor. And the collection process even requires a subsequent labor-intensive post-process to remove outliers like dynamic objects which move through the field of view during the collection period [7, 10]. These harsh conditions are not available on dynamic places like a road. Thus, a semi-supervised multi-view stereo algorithm, which can be trained even with incomplete ground truth 3D structure, is necessary to ease the generalization of the model on the unseen environment.

In this paper, we first explore a novel semi-supervised MVS problem called a Sparse Ground truth-based MVS (SGT-MVS) problem, which assumes that only the sparse ground truth 3D structure is available for training. We first investigate its fundamental characteristics to discover key aspects for solving the SGT-MVS problem. Specifically, though the relatively scarce depth information inevitably degrades the 3D reconstruction quality overall, the systematic depth reasoning principle of MVS enables the MVS networks to reasonably estimate depth values on the non-occluded region even with a few ground truth 3D points. Nevertheless, the depth reasoning principle fundamentally suffers from prediction difficulties in occluded pixels of the given multi-view images and edge pixels. Learning-based MVS methods are able to solve these difficulties using the contextual information of the nearby non-occluded regions for the occluded pixels and highly discriminative features for the edge pixels since they can directly supervise the exact depth values on the occluded regions or edges. Sparse ground truth basically cannot guarantee the contextual and highly discriminative features to such level.

Based on our observations, we focus on improving the discriminability on the accurately predicted non-occluded regions while propagating the accurate depth values to edges, occlusions, and erroneous non-occluded regions. We use a probability map to detect theses erroneous regions since the MVS network cannot determine a certain depth value due to the fundamental prediction difficulty. Thus, by treating the probability like a confidence map, we first separate the accurately predicted regions and erroneous regions according to the probability value. Then, we apply loss function named 3D Point Consistency Loss to enhance the 3D reconstruction performance on the accurately predicted region by regressing the 3D points back-projected from the corresponding pixels to actually meet in the 3D world, where the back-projection means a transformation from a pixel in an image plane to a 3D point in a world frame. Since the corresponding pixels are likely to be back-projected into the distinctive 3D points due to the inaccurate depth values, it is reasonable to match them in the 3D world for better reconstruction quality. Finally, we propagate these improved predictions toward the low confidence regions through our Coarse-to-fine Reliable Depth Propagation module. To verify our method on the semi-supervised SGT-MVS problem, we generate sparse ground truth from the original dense 3D structures by randomly sampling at ratios of 1×10^{-5} and 1×10^{-4}. As shown in Fig. 1, while original dense ground truth 3D structures consist of approximately 3×10^6 points, our sparse ground truth 3D structures only consist of tens and hundreds of 3D points, respectively, for each 3D structure. We compared our SGT-MVSNet with the state-of-the-art MVS networks on the same sparse ground truth and confirm that our method can succesfully solve the SGT-MVS problem. Moreover, SGT-MVSNet matches the capacity of other state-of-the-art MVS networks, though we only used tens and hundreds of ground truth 3D points.

To summarize, our contributions are threefold:

- We first introduce a novel semi-supervised multi-view stereo problem called Sparse Ground truth-based MVS (SGT-MVS) problem.
- We introduce SGT-MVSNet, a semi-supervised MVS framework suitable for the sparse ground truth that consists of the 3D Point Consistency Loss and the Coarse-to-fine Reliable Depth Propagaion module.
- Extensive experiments verify that our method successfully solve the semi-supervised MVS problem with the sparse ground truth. The reconstruction performance of SGT-MVSNet is comparable to the supervised MVS methods even though we only used a few points.

2. Related Work

2.1. Learning-based Multi-view Stereo (MVS)

Recently, learning-based methods have been successfully applied to MVS reconstruction. SurfaceNet [8] and DeepMVS [5] prewarped the image features to the 3D voxelized space and used 3D CNNs to estimate the object surface. Due to the limitations of these voxel-based approaches, depth map-based approaches have been proposed to tackle large-scale reconstruction. Yao *et al.* [15] first proposed an end-to-end framework that constructs the cost volume by warping 2D image features from neighboring images. In addition, they applied the 3D CNN to regularize this cost volume and regressed a depth map. Most recent learning-based MVS algorithms [6, 16, 1, 17, 11] built upon depth map-based approaches which use a plane-sweeping algorithm [2] to compute a cost volume from the multi-view images and then estimate depth maps by via regression or classification. R-MVSNet [16] used the convolutional GRU to sequentially build a cost volume and reduced GPU memory consumption. To save memory consumption, Fast-MVSNet [17] used a sparse-to-dense strategy that refines

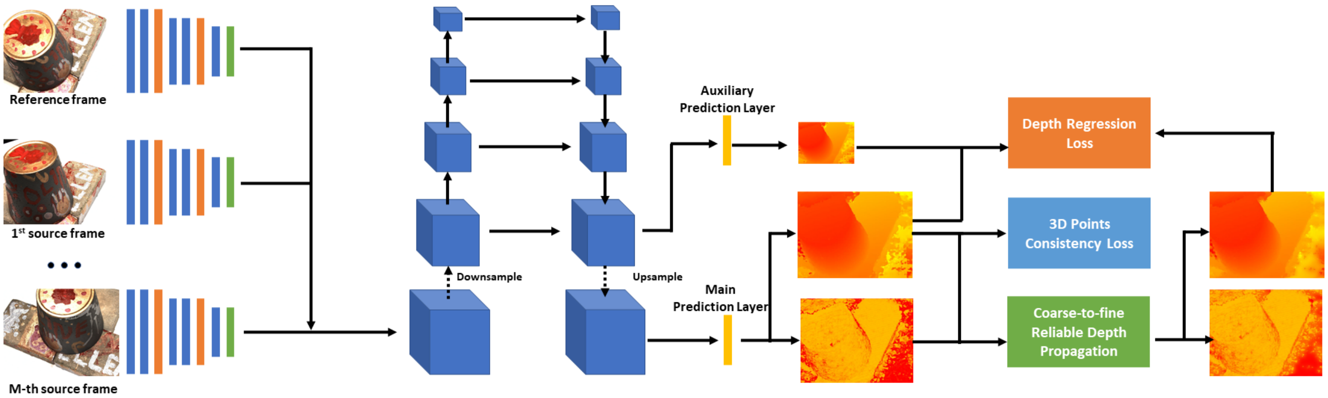


Figure 2: An overall framework of SGT-MVSNet. Our framework mainly consists of a feature extractor, a pyramidal cost volume regularization module, double prediction layers, the proposed Coarse-to-fine Reliable Depth Propagation module with a depth regression loss and the proposed 3D Point Consistency Loss.

a sparse depth map by introducing a differentiable Gauss-Newton layer. Chen *et al.* [1] proposed a new approach that improves depth prediction by refining the point cloud in the 3D space in a coarse-to-fine manner. Att-MVSNet [11] leveraged the attention modules [4] to improve the MVS performance. However, these methods highly rely on dense ground truth 3D structures despite their difficulty in the collection. Therefore, we focus on alleviating the dependence of the MVS networks on the dense ground truth.

2.2. Unsupervised Multi-view Stereo

Most learning-based MVS methods are heavily reliant on dense ground truth depth maps. However, generating dense depth maps for large-scale datasets is expensive and time-consuming. To overcome this limitation, Tejas *et al.* [9] used the combination of both the photometric loss and the regression loss to train an MVS network without ground truth depth maps. Some methods [3, 18] adopt a depth consistency loss across views. However, despite the promising approaches for easy generalization on new environments, these methods are not quite competitive compared to the supervised MVS method. Besides, our method is comparable to the supervised MVS methods even with tens and hundreds of ground truth 3D points.

2.3. Multi-view Stereo Datasets

There are a number of datasets for evaluating MVS algorithms. The Middlebury dataset [14] is the first public benchmark for MVS evaluation. It consists of hundreds of low-resolution images with calibrated cameras in a controlled laboratory environment. The ETH3D dataset [13] includes high-resolution images of building facades models and 3D ground truth point clouds captured by a laser scanner. The DTU dataset [7] contains large amounts of images of real-world objects with surface point clouds, which are collected using a robotic arm. The DTU dataset provides diverse and well-textured scenes under different lighting conditions. The Tanks and Temples dataset [10] includes high-resolution video data and ground truth point clouds collected by a laser scanner. However, most of these datasets are collected through time-consuming and labor-intensive processes, which inspired the pursuit of our method.

3. Method

3.1. Problem Formulation

For a given reference frame I_0 and source frames $\{I_i\}_{i=1}^{N}$, our main objective is to estimate dense depth map D of a 3D structure from the reference view. The only difference with the supervised MVS problem is that we can only use sparsely collected 3D points $\{P_j\}_{j=1}^{M}$ of the original 3D structure. To practically use the ground truth 3D points for training, we compute depth values of each 3D point in the perspective of the reference view. In additional, K, R, and t denote for the intrinsic, rotation and translation parameters of a reference view, respectively.

3.2. Observations on the SGT-MVS problem

Some dense estimation tasks like semantic segmentation require a deep contextual understanding of each class to robustly estimate on the unseen environment, which require abundant pixel-level annotations. On the other hand, the fundamental depth reasoning mechanism of stereo matching and MVS is to search for an optimal consistency cost for each pixel. Even without the help of numerous ground truth, if the encoder is able to extract discriminative features from given multiple views, the estimation network can reasonably predict the depth values on non-occluded pixels of the reference frame. However, since occluded pixels have only a few or no corresponding pixels in the source frames, these pixels might not be photometrically consistent with other views and the multi-view consistency cost might not be optimal for the exact depth value. Thus, the erroneous estimation on occluded regions cannot be easily solved only with discriminability of the feature.

Besides the occluded pixel issue, the erroneous depth estimation on edge pixels is also an important issue in the SGT-MVS problem. Since the depth value tends to vary greatly at the boundary of the objects, the cost volume should change drastically at the edge pixels. Hence, these regions fundamentally require highly discriminative features that can differentiate the cost volume value across the edge.

Conventional MVS methods are able to address these issues since directly supervising the ground truth depth values enables the network to refer cost volume values of nearby non-occluded pixels on the occluded pixels and edge pixels. This approach can rarely hold in the SGT-MVS problem since sparse ground truth 3D structures can barely provide the exact depth values for such pixels.

3.3. Overall Pipeline

Based on the aforementioned observations, we aim at solving the semi-supervised multi-view stereo (SGT-MVS) problem by maximizing the discriminability of the feature in a self-supervised approach and propagating the accurate depth predictions to the fundamentally erroneous regions. To achieve this goal, we **first** employ a suitable prediction layer to maximize the sparse ground truth exploitation. **Second**, to address the degenerative 3D reconstruction quality, we design the **3D Point Consistency Loss** to regress the 3D points back-projected from the corresponding pixels to actually meet in the 3D world. Since this training method is vulnerable to wrong pixel correspondences, we set a firm criterion to filter them out. **Third**, to tackle the fundamental difficulties on edges and occlusions, we build the **Coarse-to-fine Reliable Depth Propagation** module that leverages nearby accurate predictions to revise the wrongly predicted depth values. The overall framework of our SGT-MVSNet is described in Fig. 2.

3.4. Network Architecture

Our feature extractor and cost volume regularizer share similar structures with the recent learning-based MVS networks [15, 16, 1, 17]. We use 8-layer 2D CNNs for feature extraction and a 3-level pyramidal module with 10-layer 3D CNNs for cost volume regularization. Inspired from the efficient cost volume regularization process of the Fast-MVSNet [17], we downsample the base cost volume by half before regularization. However, unlike the conventional dense ground truth-based MVS networks [15, 16, 1, 17], we construct double prediction layers including a main prediction layer and a half resolution auxiliary prediction layer. We design this structure based on our empirical observation that it is desirable to use contextual information for textureless 3D structures. This also performs slightly better than a single prediction layer in the scarce ground truth setting. We use the regularized cost volume of size $\frac{1}{8}H \times \frac{1}{8}W \times 8$ for the auxiliary prediction, and upsample to the base cost volume size of $\frac{1}{4}H \times \frac{1}{4}W \times 8$ for main prediction. Note that our double prediction layers do not correspond to a complex auxiliary inference structure discussed in Section 3.2, so the risk of discriminability contamination is low.

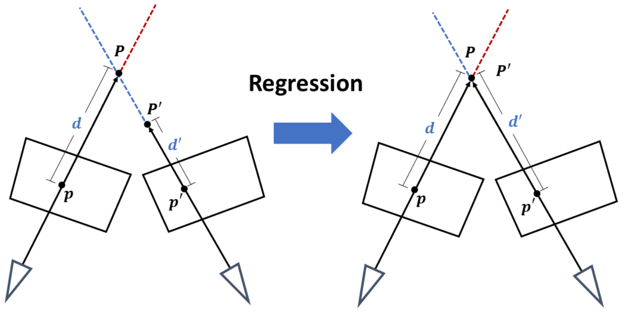


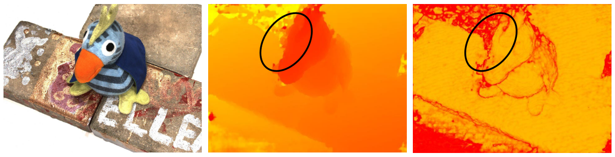
Figure 3: Description of the 3D Point Consistency Loss. Though the estimated corresponding pixel p' of p has inaccurate depth values, our 3D Point Consistency Loss can bring the 3D points P of p and P' of p' together so that the network can correctly estimate depth values on the pixel p'.

3.5. 3D Point Consistency Loss

Though MVS networks can reasonably predict depth values on the non-occluded regions of the given multi-view images, the 3D reconstruction results inevitably suffer from degraded performance compared to the dense ground truth-based MVS networks. Therefore, even though corresponding pixels of each view theoretically originate from an identical 3D point, it is likely that the back-projected 3D points of these pixels will not meet at the same position in the 3D world frame due to the inaccurate depth predictions. Here, the back-projection means a transformation from a pixel in an image plane to a 3D point in a world frame. To solve this degenerative 3D points reconstruction quality issue, we define a **3D Point Consistency Loss** to regress the back-projected 3D points of the corresponding pixels to actually meet at the same 3D coordinate and eventually form a correct correspondence in the 3D world, as shown in Fig. 3.

However, pixels on edges and occlusions are deterministically projected onto nearby wrong pixels in other views. Moreover, even non-occluded pixels can be matched with wrong pixels in other view if the predicted depth values are inaccurate. Regressing 3D world distances between these inaccurate 3D points can rather form false correspondences. To tackle these challenges, we need a solid criterion to filter out these potential false correspondences and search for the reliable pixels that are likely to match in the 3D world.

In this state, we observed that a probability map of the



(a) Reference view (b) Depth prediction (c) Probability map

Figure 4: Visualization of the erroneous prediction and the corresponding probability map. The edges and occlusions tend to have a low probability value. We treat the probability map like a confidence map for the reliable depth prediction.

depth prediction can approximately detect edges, occlusion, and erroneous non-occluded regions that could potentially form wrong correspondences, as shown in Fig. 4. Thus, we design a criterion with two conditions based on the probability map: i) the probability value of the pixels should exceed a certain confidence threshold ϵ_h; ii) The distance between the 3D points should not exceed a certain distance threshold ϵ_w. Then, we formulate the 3D Point Consistency Loss for a pixel p as follows:

$$\begin{aligned} p' &= K'(R'R^{-1}(D(p)K^{-1}p - t) + t') \\ P &= D(p)K^{-1}p - t \\ P' &= D(p')K'^{-1}p' - t' \\ L_{con}(p) &= \begin{cases} \|P - P'\|_2, & \text{if } \min\limits_{q\in\{p,p'\}} C(q) > \epsilon_h \\ & \text{and } \|P - P'\|_2 < \epsilon_w \\ 0, & \text{otherwise} \end{cases}, \end{aligned} \tag{1}$$

where p' denotes the estimated corresponding pixels of p, P and P' denote the 3D points back-projected from p and p', $D(p)$ and $C(p)$ denote a predicted depth and probability values of the pixel p, ϵ_h represents a threshold for highly reliable depth prediction, and ϵ_w represents a threshold for filtering wrongly matched pixels.

3.6. Coarse-to-fine Reliable Depth Propagation

The fundamental errors on the occlusions and edges still remain unresolved since the 3D Point Consistency Loss mainly focuses on enhancing the performance on the non-occluded regions. Thus, we aim to explicitly propagate accurate predictions to erroneous regions. Inspired by the observations on the probability map, we build a propagation module called a **Coarse-to-fine Reliable Depth Propagation** module that modifies uncertain depths by referring to nearby reliable predictions with a high probability value while preserving reliable predictions.

In addition, our module works in a coarse-to-fine manner, considering that these erroneous values usually appear at the patch-level region rather than a single pixel, as shown in Fig 4. (b). Moreover, we refer to not only the surrounding pixels but also farther pixels to reflect the contextual tendency of the nearby depth values. For a given pixel position p, our propagation strategy is to utilize the surrounding neighbor $N_{d_k}(p)$ with $d_1 = 1$ and the neighbors $N_{d_k}(p)$, $(k = 2, ..., K)$ with certain dilation values d_k to modify the depth values $D(p)$ on each coarse-to fine level as follows:

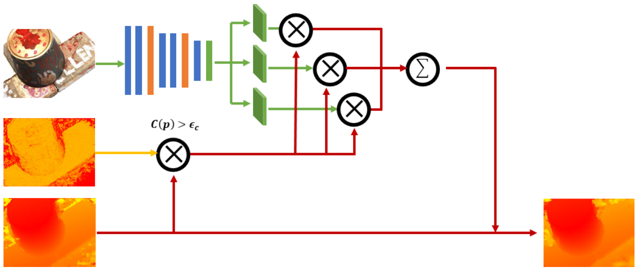


Figure 5: Description of the Coarse-to-fine Reliable Depth Propagation process in each coarse-to-fine level for the depth map. We update the uncertain depth prediction based on the feature similarity between the nearby pixels. We enlarge the scope through dilated convolution to reflect the contextual tendency of the depth values. The confidence map is also updated in the same procedure.

$$D'(p) = \begin{cases} \sum_{q\in\bigcup_{k=1}^{K} N_{d_k}(p)} D(q)\cdot w_{p,q}, & \text{if } C(p) > \epsilon_c \\ D(p), & \text{otherwise} \end{cases}, \tag{2}$$

where $C(p)$ is a probability value for the pixel position p, and ϵ_c is a threshold for the probability value. The computation process is described in Fig. 5. In the test phase, we also update the probability value on p to determine whether to proceed the propagation on p at the next level or not.

$$C'(p) = \begin{cases} \sum_{q\in\bigcup_{k=1}^{K} N_{d_k}(p)} C(q)\cdot w_{p,q}, & \text{if } C(p) > \epsilon_c \\ C(p), & \text{otherwise} \end{cases}, \tag{3}$$

Empirically, we employ two additional dilated neighbors (K=3) with $d_2 = 3$ and $d_3 = 6$.

4. Training Loss

Our training loss mainly consists of the 3D Point Consistency Loss and depth regression losses for double prediction layers and the Coarse-to-fine Reliable Depth Propagation module. We employ the mean absolute difference for depth regression losses. Suppose we have a ground truth depth map $\hat{D}$ and a half-sized map $\hat{D}_{aux}$ for the reference view and the source views v_1 and v_2 are the randomly sampled source views. Then, the training loss can be formulated as follows:

$$Loss = \sum_{p \in \mathbf{p}_{sparse}} \|D(p) - \hat{D}(p)\| + \|D_{aux}(p) - \hat{D}_{aux}(p)\|$$
$$+ \lambda_{pcl} \cdot \sum_{v \in \{v_1, v_2\}} \sum_{\substack{C_{ref}(p) > \epsilon_h \\ C_v(p'_v) > \epsilon_h \\ \|P - P_v\|_2 < \epsilon_w}} \|P - P_v\|_2$$
$$+ \lambda_{crdp} \cdot \sum_{p \in \mathbf{p}_{sparse}} \|D'(p) - \hat{D}(p)\|, \tag{4}$$

where λ_{pcl} and λ_{crdp} are weights for the 3D Point Consistency Loss and the Coarse-to-fine Reliable Depth Propagation module. Note that all the previously defined notations with a subscript v represent that of the v-th view.

5. Experiments

5.1. Experimental Setup

Datasets We used the *DTU* and *Tanks and Temples* datasets [7, 10] for evaluation. The *DTU* dataset [7] is a large-scale MVS dataset which consists of 128 different 3D structures with each scene captured from 49 or 64 fixed viewpoints on seven light conditions. The dataset originally provides ground truth point cloud and normal surfaces, and Yao *et al*. [15] generated the depth maps by rendering the mesh surface to each viewpoint. We followed the same configuration of training, validation, and evaluation splits by the previous learning-based MVS methods for a fair comparison. The *Tanks and Temples* dataset is a large-scale MVS dataset for the outdoor scenes that consists of intermediate and advanced sequences. Following the previous MVS methods [15, 16, 1, 17], we only use the intermediate sequence for evaluation.

Sparse ground truth depth map generation The semi-supervised MVS problem aims to learn dense 3D reconstruction capacity only with the sparse ground truths so that the network can reconstruct the original 3D structure from multi-view images. Thus, to verify the effectiveness of our method on the semi-supervised MVS problem, we need a data split with sparse ground truth 3D structures for training and another data split with dense ground truth 3D structures for evaluation. Since the MVS datasets basically do not provide sparse ground truth depth maps, we newly generate them for the experiments. We randomly sampled 3D points from each ground truth 3D structure with a certain ratio. We used the sampling ratios of 1×10^{-5} and 1×10^{-4}, and these quantities imply tens and hundreds of points for each 3D structure in the *DTU* dataset, respectively.

Implementation details We trained our MVS network with the training data of the DTU dataset. Following the learning curriculum of the previous works [15, 16, 17, 1], we used the input size of 640 × 412 and 960 × 1280, and three and five views respectively for training and tests. We used 96 depth planes for training and 144 depth planes for testing. Following the consensus of the previous works [15, 16, 17, 1], We employed the RMSProp optimizer. We set the initial learning rate by 0.0005 and decayed it by 0.9 for each two epoch. We used 0.9, 10, and 0.5 for ϵ_h, ϵ_w, and ϵ_c. We set the batch size to 4. All experiments were conducted on four NVIDIA GTX Titan Xp GPUs using PyTorch [12].

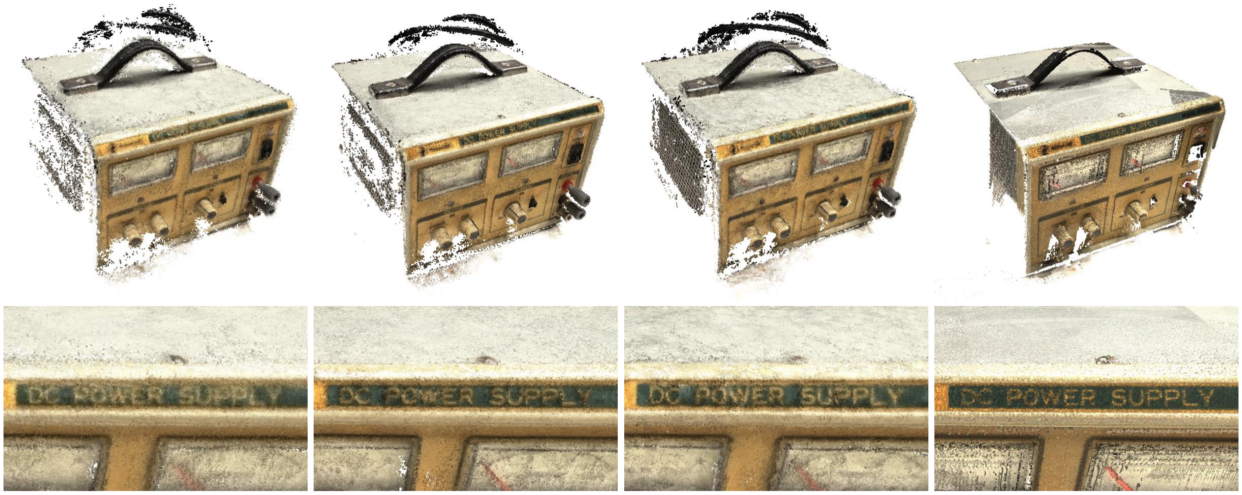
Ground-truth	Method	Acc. (mm)	Comp. (mm)	Overall (mm)
Sparse (1× 10^{-4})	MVSNet [15]	0.481	0.492	0.486
	PointMVSNet [1]	0.490	0.502	0.496
	Fast-MVSNet [17]	**0.355**	0.425	0.390
	Ours	0.421	**0.349**	**0.385**
Sparse (1× 10^{-5})	MVSNet [15]	0.537	0.494	0.516
	PointMVSNet [1]	0.603	0.621	0.612
	Fast-MVSNet [17]	**0.381**	0.481	0.431
	Ours	0.441	**0.381**	**0.411**

Table 1: Quantitative 3D reconstruction quality results of our method and the state-of-the-art supervised MVS methods in the sparse ground truth setting on the DTU evaluation dataset [7].

Moreover, since 3D reconstruction process fundamentally have a trade-off between accuracy and completeness Moreover, since the 3D reconstruction process fundamentally have a trade-off between accuracy and completeness that can be controlled by reconstruction parameters. occurs the sub-optimality on the accuracy metric is also

5.2. Comparisons on the DTU dataset

We trained the state-of-the-art MVS methods with the sparse ground truth 3D structures to verify the effectiveness of our method in the SGT-MVS problem. The 3D reconstruction process fundamentally has a trade-off between accuracy and completeness controlled by reconstruction parameters, and we determined the current trade-off for the overall performance. As shown in Table 1, though Fast-MVSNet [17] achieved the best performance in the accuracy metric, our method outperformed the state-of-the-art methods in the completeness and overall metrics in the sparse ground truth setting with both sampling ratios of 1×10^{-5} and 1×10^{-4}. Although the our performance is sub-optimal in the accuracy metric, the optimal accuracy performance can also be obtained with the loss of the completeness performance. Figure 6 (a) and (b) show qualitative results of Point-MVSNet [1] and our SGT-MVSNet at the 1×10^{-5} sampling ratio setting, and our results show better visibility on the letters. Moreover, our method shows visually fine reconstruction results even compared with the dense ground truth-based MVS methods and the ground



(a) Point-MVSNet (Sparse GT) (b) Ours (Sparse GT) (c) Point-MVSNet (Dense GT) (d) Ground truth

Figure 6: 3D reconstruction results of the Point-MVSNet trained with sparse and dense ground truth, and our SGT-MVSNet trained with sparse ground truth. We use the sparse ground truth with a sampling ratio of 1×10^{-5} for the visualization.

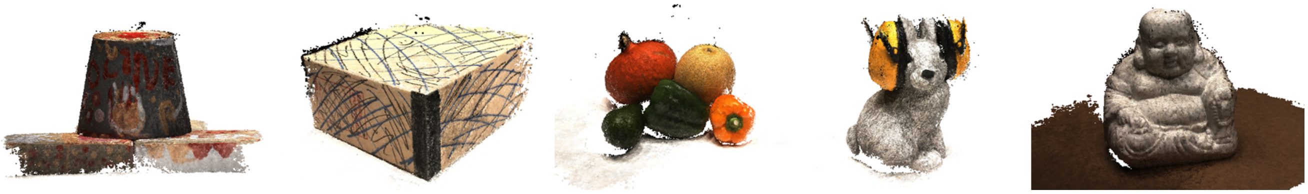


Figure 7: Qualitative results on *scan1*, *scan33*, *scan75*, and *scan114* of the DTU dataset [10]. The tested SGT-MVSNet is trained with the sparse ground truth of the DTU dataset [7] sampled with 1×10^{-5} ratio.

truth itself, as shown in Fig. 6. In addition, we visualized 3D reconstruction results for other 3D structures in Fig. 7 to verify the 3D reconstruction capability of our method.

5.3. Ablation study

To verify the effectiveness of the proposed method, we conducted quantitative and qualitative ablation study. Table 2 shows the improved reconstruction performances through the double prediction (DP) layers, the 3D Point Consistency Loss (PCL), and the Coarse-to-fine Reliable Depth Propagation module (CRDP). As we explained in Section 3.5, our method with the 3D Point Consistency Loss achieved the best performance in the accuracy metric, which verifies that our loss improved the depth prediction accuracy on the reliable non-occluded regions. Moreover, though our Coarse-to-fine Reliable Depth Propagation module slightly degrades the reconstruction accuracy, it significantly enhanced the quality in terms of completeness and eventually improved the performance in the overall metric. We visualized the 3D reconstruction results of each method with the error map in Fig. 8 to compare the reconstruction quality in detail. Interestingly, we observed that double prediction layers significantly alleviate the errors on the textureless region, which seems that the layers help to encode contextual information through the auxiliary loss. The 3D Point Consistency Loss further addressed the reconstruction errors, especially decreasing the overall error scale, as the color of the red regions becomes faded. Then, the Coarse-to-fine Reliable Depth further alleviated the erroneous red regions. From these results, we confirmed the effectiveness of our method in the SGT-MVS problem.

method	Acc.	Comp.	Overall
Baseline	0.460	0.422	0.441
Baseline + DP	0.448	0.419	0.434
Baseline + DP + PCL	**0.438**	0.396	0.417
Baseline + DP + PCL + CRDP	0.441	**0.381**	**0.411**

Table 2: Quantitative ablation study on the DTU evaluation dataset. Each of our methods contributes to the 3D reconstruction quality enhancement. Though the Coarse-to-fine Reliable Depth Propagation module slightly degrades the accuracy due to the blurring effect of propagation, it is complemented by significantly improved completeness quality, which refers to effective propagation toward erroneous regions.

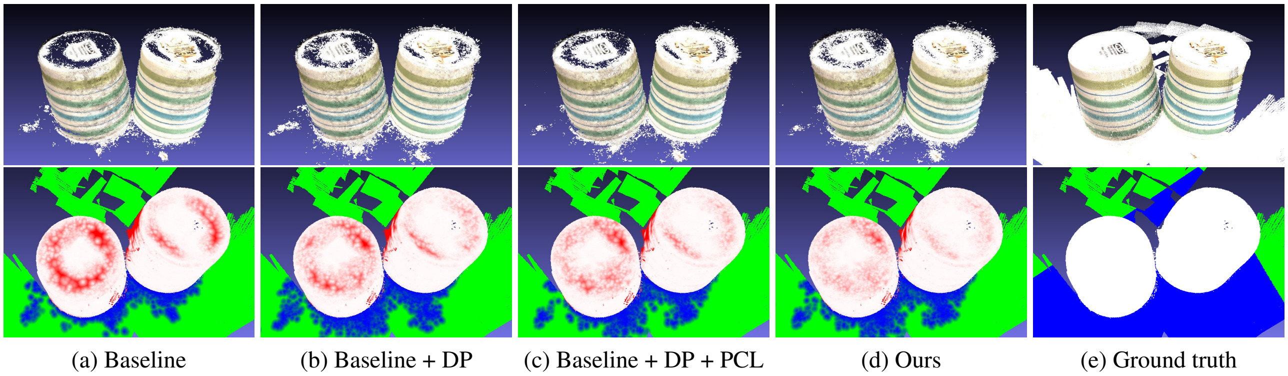

(a) Baseline (b) Baseline + DP (c) Baseline + DP + PCL (d) Ours (e) Ground truth

Figure 8: Qualitative ablation study on *scan48* of the DTU dataset. The 1st row shows the 3D reconstruction results of (a) baseline, (b) baseline with double prediction (DP) layers, (c) the aforementioned model with the 3D Point Consistency Loss (PCL), and (d) our integrated method with Coarse-to-fine Reliable Depth Propagation. The 2nd row shows the error maps of the reconstruction results and red color represents the severity of the errors.

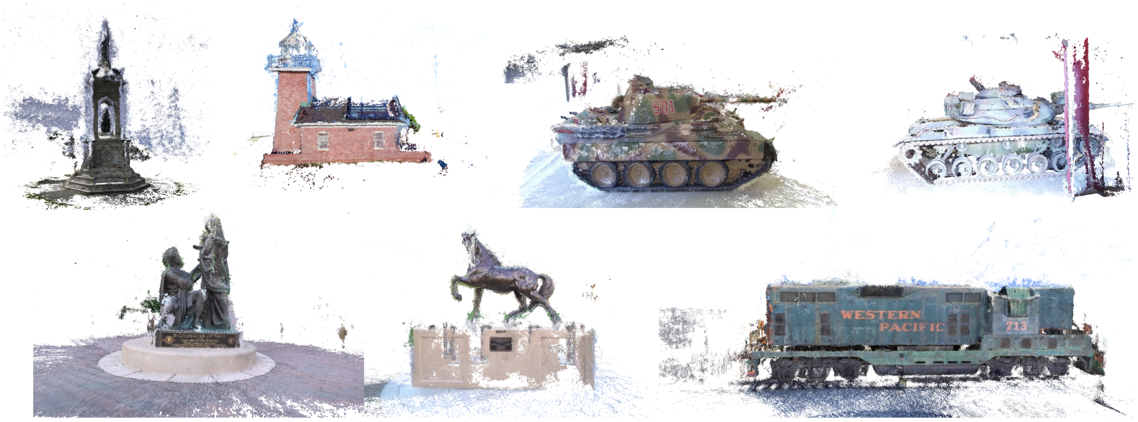


Figure 9: Qualitative results of the intermediate set in the *Tanks and Temples* dataset [10]. The tested SGT-MVSNet is trained with the sparse ground truth of the DTU dataset [7] sampled with a ratio of 1×10^{-5}.

5.4. Generalization on Tanks and Temples dataset

Following the previous methods [15, 16, 1, 17], we directly tested the model trained on the *DTU* dataset without any additional fine-tuning. We used five scenes with an input size of 1920×1056 and 144 depth planes for testing. As shown in Fig. 9, despite the sparce ground truth, our methods quite reasonably reconstruct 3D structures for a new domain, which shows the feasibility in generalization.

6. Conclusion

In this paper, we explored a novel semi-supervised multi-view stereo problem called a SGT-MVS problem. We observed the fundamentally imbalanced depth prediction performance between the accurate regions and the erroneous regions in the SGT-MVS problem. Based on the insights on the probability map, we divided the reliably predicted regions and erroneous regions. Then, we individually conquered these regions through the 3D Point Consistency Loss and the Coarse-to-fine Reliable Depth Propagation module. To verify the effectiveness, we generated sparse ground truth with 1×10^{-5} and 1×10^{-4} sampling ratios, and trained the state-of-the-art MVS networks with them. Experimental results showed that our SGT-MVSNet can be proper to the sparse ground truth setting, and the reconstructed results are also visually reliable. These results demonstrate that our method can be simply generalized to new environments only with easily collected sparse ground truth 3D structures.

Acknowledgement This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (2017-0-01772 and 2017-0-01781).

References

[1] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[2] Robert T Collins. A space-sweep approach to true multi-image matching. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 358–363. IEEE, 1996.

[3] Yuchao Dai, Zhidong Zhu, Zhibo Rao, and Bo Li. Mvs2: Deep unsupervised multi-view stereo with multi-view symmetry. In *2019 International Conference on 3D Vision (3DV)*, pages 1–8. IEEE, 2019.

[4] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[5] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018.

[6] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. DPSNet: End-to-end deep plane sweep stereo. In *International Conference on Learning Representations*, 2019.

[7] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413. IEEE, 2014.

[8] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2307–2315, 2017.

[9] Tejas Khot, Shubham Agrawal, Shubham Tulsiani, Christoph Mertz, Simon Lucey, and Martial Hebert. Learning unsupervised multi-view stereopsis via robust photometric consistency. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, 2019.

[10] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

[11] Keyang Luo, Tao Guan, Lili Ju, Yuesong Wang, Zhuo Chen, and Yawei Luo. Attention-aware multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[12] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[13] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. 2017.

[14] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Computer Vision and Pattern Recognition (CVPR)*, 2006.

[15] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. *European Conference on Computer Vision (ECCV)*, 2018.

[16] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[17] Zehao Yu and Shenghua Gao. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *CVPR*, 2020.

[18] Yiran Zhong, Yuchao Dai, and Hongdong Li. Self-supervised learning for stereo matching with self-improving ability. *ArXiv*, abs/1709.00930, 2017.