

MVS²: Deep Unsupervised Multi-view Stereo with Multi-View Symmetry

Yuchao Dai, Zhidong Zhu, Zhibo Rao, Bo Li

School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China
daiyuchao@gmail.com

Abstract

The success of existing deep-learning based multi-view stereo (MVS) approaches greatly depends on the availability of large-scale supervision in the form of dense depth maps. Such supervision, while not always possible, tends to hinder the generalization ability of the learned models in never-seen-before scenarios. In this paper, we propose the first unsupervised learning based MVS network, which learns the multi-view depth maps from the input multi-view images and does not need ground-truth 3D training data. Our network is symmetric in predicting depth maps for all views simultaneously, where we enforce cross-view consistency of multi-view depth maps during both training and testing stages. Thus, the learned multi-view depth maps naturally comply with the underlying 3D scene geometry. Besides, our network also learns the multi-view occlusion maps, which further improves the robustness of our network in handling real-world occlusions. Experimental results on multiple benchmarking datasets demonstrate the effectiveness of our network and the excellent generalization ability.

1. Introduction

Multi-view stereo (MVS) targets at reconstructing the observed 3D scene structure from its multi-view images, whereas both the intrinsic calibration and extrinsic calibration between cameras are available. Traditional geometry-based approaches exploit multi-view photometric consistency and various kinds of regularizations/priors [6]. Recently, the success of deep convolutional neural networks (CNNs) in monocular depth estimation [18, 9, 19] and binocular depth estimation [33, 34] has been extended to MVS. Existing deep CNNs based MVS approaches [30, 31, 11, 24] tend to represent MVS as an end-to-end regression problem. By exploiting large-scale ground truth 3D training data, these methods outperform traditional geometry-based approaches and dominate the leading boards on different benchmarking datasets [30, 31]. However, the success of these supervised MVS approaches strongly depends

on the availability of large-scale ground-truth 3D training data, which not only not always available but also may further hinder their generalization ability in never-seen-before open-world scenarios [34]. Thus it is highly desired to develop unsupervised learning based MVS approaches.

In this paper, we propose the first unsupervised deep MVS network as shown in Fig. 1, which could be learned in an end-to-end manner and without using ground-truth depth maps as the supervision signals. We demonstrate that the multi-view image warping errors (photometric consistency across different views) themselves are sufficient to drive a deep network to converge to the correct state that leads to superior MVS performance. Our network structure differs from existing MVS and simple extension of unsupervised binocular stereo matching in the following aspects:

- a) Our network is *symmetric* to all the views, *i.e.*, it treats each view equivalently and predicts the depth map for each view simultaneously. Existing supervised learning based MVS methods [30, 31, 11, 27] apply an “*asymmetric*” design and infer depth map for the reference image only. Thus, multiple depth maps estimated from different viewpoints do not comply with the same 3D geometry and 3D point clouds processing is required to derive a consistent 3D geometry. We would like to argue that this kind of “centralized” and “asymmetric” design has not fully exploited the multi-view relation encoded in the multi-view images.
- b) We propose a new cross-view consistency in depth maps building upon our multi-view symmetry network design. The underlying principle is that as the multi-view images observe the 3D scene structure from different viewpoints, the estimated depth maps from MVS network should be consistent in 3D geometry. As our experiments demonstrate, this consistency plays a key role in strengthening the image warping error and guiding the network to coverage to meaningful states.
- c) We integrate multi-view occlusion reasoning into our network, which enables us to detect occluded regions by using the cross-view consistency in depth maps. Under our framework, multi-view depth maps prediction and occlusion reasoning are alternatively updated.

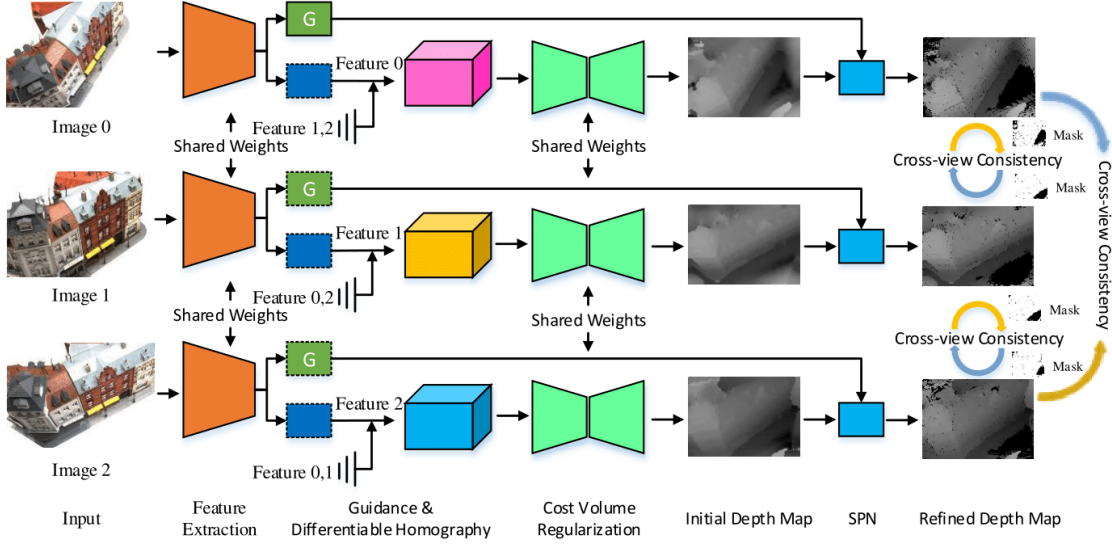


Figure 1. **Our unsupervised deep multi-view stereo network architecture.** Our network consists of five modules, namely, feature extraction, guidance and differentiable homography, cost volume regularization, spn, and cross-view consistency loss evaluation. Note that, under our symmetric design, our network outputs consistent depth map for each image.

Our main contributions are summarized as follow:

- 1) We present the first deep unsupervised MVS approach, which naturally fills the gap between traditional geometry-based approaches and deep supervised MVS methods. Our proposed unsupervised method avoids the necessity of large-scale 3D training data.
- 2) We introduce the cross-view consistency in depth maps and propose a loss function to measure the consistency. We demonstrate that this kind of consistency could be utilized to guide the training of a deep neural network.
- 3) Expensive experiments conducted on the SUN3D, RGB-D, DTU and Scenes11 benchmarking datasets demonstrate the effectiveness and the excellent generalization ability of our method.

2. Related Work

MVS has been an active research topic in geometric vision. Existing methods can be roughly classified into two categories: 1) Geometry-based MVS and 2) Supervised learning based MVS. We will also discuss related work in unsupervised monocular and binocular depth estimation.

Geometry-based Multi-view Stereo: Traditional MVS methods focus on designing neighbor selection and photometric error measures for efficient and accurate reconstruction [5, 8, 4]. Furukawa *et al.* [3] adopted geometric structures to reconstruct textured regions and applied Markov random fields to recover per-view depth maps. Langguth *et al.* [17] used the shading-aware mechanism to improve the robustness of view selection. Wu *et al.* [28] utilized the lighting and shadows information to enhance the perfor-

mance of the ill-posed region. Michael *et al.* [10] chose images to match (both at a per-view and per-pixel level) for addressing the dramatic changes in lighting, scale, clutter, and other effects. Schonberger *et al.* [22] proposed the COLMAP framework, which applied photometric and geometric priors to optimize the view selection and used geometric consistency to refine the depth map.

Supervised Deep Multi-view Stereo: Different from the above geometry-based methods, learning-based approaches adopt convolution operation which has powerful feature learning capability for better pair-wise patch matching [32, 12, 14]. Ji *et al.* [12] pre-warped the multi-view images to 3D space, then used CNNs to regularize the cost volume. Huang *et al.* [11] proposed DeepMVS, which aggregates information through a set of unordered images. Abhishek *et al.* [14] directly leveraged camera parameters as the projection operation to form the cost volume, and achieved an end-to-end network. Yao *et al.* [30] adopted a variance-based cost metric to aggregate the cost volume, then applied 3D convolutions to regularize and regress the depth map. Im *et al.* [24] applied a plane sweeping approach to build a cost volume from deep features, then regularized the cost volume via a context-aware aggregation to improve depth regression. Very recently, Yao *et al.* [31] introduced a scalable MVS framework based on the recurrent neural network to reduce the memory-consuming.

Unsupervised Geometric Learning: Unsupervised learning has been developed in monocular depth estimation and binocular stereo matching by exploiting the photometric consistency and regularization. Xie *et al.* [29] proposed Deep3D to automatically convert 2D videos and images

to stereoscopic 3D format. Zhou *et al.* [35] proposed an unsupervised monocular depth prediction method by minimizing the image reconstruction error. Mahjourian *et al.* [21] explicitly considered the inferred 3D geometry of the whole scene, where consistency of the estimated 3D point clouds and ego-motion across consecutive frames are enforced. Zhong *et al.* [33, 34] used the image warping error as the loss function to derive the learning process for estimating the disparity map.

3. Our Network

In this section, we present our unsupervised learning based multi-view stereo network, MVS², which could be learned without the need of ground truth 3D data. We represent MVS as the task of predicting a depth map for each view simultaneously such that the estimated multiple depth maps comply with the underlying 3D geometry. Our network structure follows the MVSNet model proposed in [30] but with significant modifications to achieve unsupervised MVS with multi-view symmetry, *i.e.*, MVS².

3.1. Multi-view Symmetric Network Design

Under the MVS configuration, each image observes the underlying 3D scene structure from different viewpoints. Therefore, the estimated depth maps from MVS network should be consistent in 3D geometry and each depth map estimation is not independent. However, existing deep MVS networks [30, 11, 31] generally apply an “asymmetric” design and infer depth map for each image (termed as “reference image”) individually. Thus, multiple depth maps estimated from different viewpoints do not necessarily comply with the same underlying 3D geometry.

In this paper, we propose a de-centralized and multi-view symmetric network structure for MVS as illustrated in Fig. 1. Our network is *symmetric* to all the views, *i.e.*, it treats each view equivalently and predicts the depth map for each view simultaneously. Our unsupervised deep MVS network consists of five modules, namely, multi-scale feature extraction, cost volume construction, cost volume regularization, depth map refinement through spatial propagation network, and unsupervised loss evaluation. We briefly describe each module with focus on how to achieve multi-view symmetry and how to enforce multi-view consistency.

3.1.1 Cost Volume Reconstruction

Under our multi-view symmetry configuration, we need to estimate a depth map for each input view. Following the MVSNet network, a cost volume has to be constructed for each input view. Denote the feature map extracted by feature extraction module for each view as $\mathcal{F}_i \in \mathbb{R}^{H \times W \times F}$, where H, W, F denote the image height, image width and feature dimension correspondingly. We adopt the classical

plane sweeping based stereo pipeline and use differentiable homography matrix to warp the current image into each of the remaining images as shown in Fig. 2.

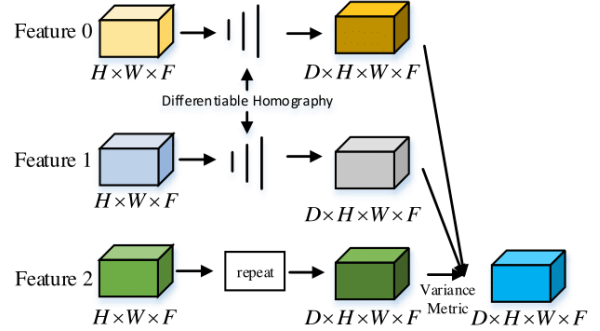


Figure 2. **Cost Volume Construction.** The per-image cost volume is constructed by calculating the variance of the warped feature maps and the reference feature maps.

In this way, we obtain $N - 1$ warped feature volumes for each depth value d . We add the current feature volume into the group of warping feature volumes. Denote D as the depth sample number, then we obtain D groups of multiple feature volumes $\{V_{ij}\}_{j=1, \dots, N}$. Finally, the multiple feature volumes are aggregated to one cost volume $\mathcal{C}_i \in \mathbb{R}^{D \times H \times W \times F}$ by using the variance operation [30], which has been shown to be better than other operations such as mean or sum operation.

3.1.2 Cost Volume Regularization

The raw cost volume \mathcal{C}_i aggregated by the variance-based cost metric could be noise-contaminated, so we utilize 3D CNN to regularize each raw cost volume to generate a probability volume. After that, we apply the ArgMin operation to regress the depth map for the current view. The cost volume regularization process is illustrated in Fig. 3.

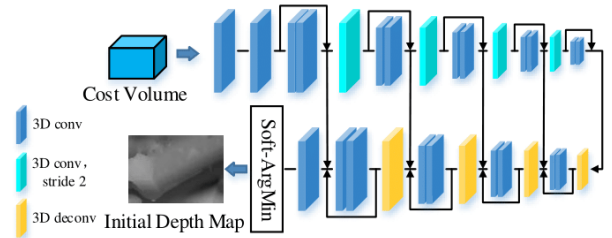


Figure 3. **The cost volume regularization module.** It takes a cost volume as input, and is followed by a series of 3D CNNs.

As shown in Fig. 3, we apply multi-scale 3D CNN to regularize the cost volume. The multi-scale 3D CNN consists of four-scale, where each convolutional operation is followed by a BN layer and a ReLU layer. On this base, we pass the feature maps between the same scale to form

a residual architecture for avoiding losing the critical information. The output of our regularization module is a 1-channel volume \mathcal{V} with dimension $D \times 1/4H \times 1/4W$.

Finally, we adopt the regression way to obtain an initial depth map \mathcal{D}_{init} . We first use the softmax function along the depth dimension to convert volume \mathcal{V} to a probability map \mathcal{P} . Then, we apply the ArgMin operation to regress the depth map. The whole process is expressed as:

$$\mathcal{D}_{init} = \sum_{d=d_{min}}^{d_{max}} d \times \mathcal{P}(d) = \sum_{d=d_{min}}^{d_{max}} d \times \text{softmax}(\mathcal{V}_d), \quad (1)$$

where d_{min}, d_{max} denote the min and max depth value.

3.1.3 Depth Map Refinement

Even though the initial depth map is already a qualified output, the reconstruction boundaries of the object may suffer from over-smoothing due to up-sampling. To tackle this problem and improve the performance, we apply the spatial propagation network (SPN) [20] to refine the initial depth map. In this step, we obtain the guidance from the feature extraction module, and it could produce the affinity matrix which is spatially dependent on the input image. Then, we adopt the affinity matrix to guide the refinement process.

3.2. Multi-view Occlusion reasoning

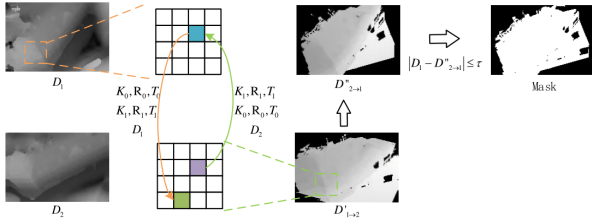


Figure 4. **The cross-view depth consistency check process.** First, we warp the refined depth map D_1 to other view. Then, we re-warp the warping depth $D'_{1 \rightarrow 2}$ map to current view. Final, we compare the refined depth map D_1 and warping depth map $D''_{2 \rightarrow 1}$ to obtain the mask.

Occlusion is inevitable to MVS, thus we have to decide the occlusion mask to avoid the occluded points from participating in the loss evaluation. Different from the occlusion mask detection based on forward-backward consistency check [23][36], we exploit pixel-wise cross-view depth consistency to obtain the occlusion mask. Specifically, given a pair of estimated depth maps D_i and D_j , we can synthesize two versions of D_i by using the depth maps and the warping relations $W_{i \rightarrow j}$ and $W_{j \rightarrow i}$. The first order synthesized depth map $D'_{i \rightarrow j}$ is generated by D_i and $W_{i \rightarrow j}$. The second synthesized depth map $D''_{j \rightarrow i}$ is generated by $D'_{i \rightarrow j}$ and $W_{j \rightarrow i}$. The cross-view depth consistency

check is illustrated in Fig. 4. Given perfect depth maps D_i and D_j , D_i and $D''_{j \rightarrow i}$ should be the same up to occlusion. Therefore, we mark points which satisfy the constraint $|D_i - D''_{j \rightarrow i}| > \tau$ as invalid, where we set the threshold $\tau = 5$. For the sake of robustness, we use the cross-view depth consistency rather than the brightness consistency. In Fig. 5, we present a visualization of the evolution of the occlusion mask, its corresponding source images and synthesized image. It can be observed that with the increase of iterations the occlusion mask becomes more and more accurate.

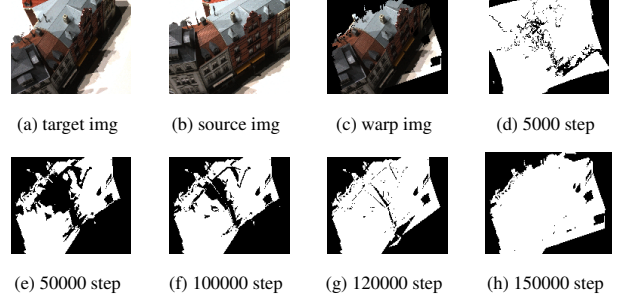


Figure 5. **The visualization and iteration of the occlusion mask.** From left to right: target image, source image, warped image and the occlusion mask of various stages of the training process.(black means invalid points). Best Viewed on Screen.

3.3. Loss Functions for Unsupervised MVS

In this paper, we target to develop an unsupervised learning framework to estimate a fine and smooth depth map for each input image. For optimizing the quality of depth map, we adopt two aspects of loss functions: view synthesis loss which includes unary term loss and smoothness loss, and cross-view consistency loss. Given multi-view images (I_1, I_2, \dots, I_n) , we first obtain the corresponding estimated depth maps during the training process. With the estimated depth map D_i of the i_{th} view and the given camera pose between i_{th} and j_{th} view (I_i and I_j), we can produce the synthesized view $I'_{j \rightarrow i}$ of i_{th} using the pixel of j_{th} view and the mapping relations $W_{j \rightarrow i}$ between them. Similarly, we can also obtain the synthesized image $I'_{i \rightarrow j}$ and the mapping relations $W_{i \rightarrow j}$. According to the bilateral mapping relations, we can produce the secondary synthesized image $I''_{j \rightarrow i}$ using $I'_{i \rightarrow j}$ and similarly the $I''_{i \rightarrow j}$ using $I'_{j \rightarrow i}$ with the bilinear sampler methods.

Our overall loss function can be formulated as follows:

$$\mathcal{L} = \sum_{i=1}^{v-1} \sum_{j=i+1}^v \mathcal{L}_{synthesis}^{i,j} + \mathcal{L}_{consistency}, \quad (2)$$

where v denotes the total amount of selected views. Apart from this, $\mathcal{L}_{synthesis}^{i,j}$ and $\mathcal{L}_{consistency}$ stand for the synthesized image loss between I_i and I_j and the cross-view consistency loss.

3.3.1 View Synthesis Loss

The view synthesis loss $\mathcal{L}_{synthesis}^{i,j}$ between I_i and I_j is defined as:

$$\mathcal{L}_{synthesis}^{i,j} = \omega_u(\mathcal{L}_u^{i,j} + \mathcal{L}_u^{j,i}) + \omega_s\mathcal{L}_s^i, \quad (3)$$

where $\mathcal{L}_u^{i,j}$ denotes the unary term loss and \mathcal{L}_s^i denotes the depth field smoothness regularization loss.

Unary term loss. During the reconstruction process, we would like to minimize the discrepancy between the source image and the reconstructed image. Our loss consists of not only the L_1 distance between images and their gradients, but also the structure similarity SSIM. In order to further improve the robustness in brightness, we also exploit the Census transformation to measure the difference. Thus, our unary term loss is defined as follow:

$$\mathcal{L}_u^{i,j} = \frac{1}{|M|} \sum \left(\lambda_1 \cdot \varphi(I_i - I'_{j \rightarrow i}) + \lambda_2 \cdot \varphi(\nabla I_i - \nabla I'_{j \rightarrow i}) + \lambda_3 \cdot \frac{1 - \mathcal{S}(I_i, I'_{j \rightarrow i})}{2} + \lambda_4 \cdot \varphi(C(I_i) - C(I'_{j \rightarrow i})) \right) \cdot M, \quad (4)$$

where M is the unoccluded mask for obtaining the valid points. $\mathcal{S}(\cdot)$ denotes the structure similarity SSIM. $\varphi(s) = \sqrt{s^2 + 0.001^2}$ can elevate the robustness of our loss. $\nabla(\cdot)$ denotes the gradient operator and $C(\cdot)$ denotes the Census transformation of image. In this paper, we set $\lambda_1 = 0.5$, $\lambda_2 = 0.8$, $\lambda_3 = 0.5$, $\lambda_4 = 0.2$.

Smoothness regularization term loss. To encourage the smoothness in the predicted depth map, the depth smoothness term is defined as:

$$\mathcal{L}_s^i = \frac{1}{N} \sum \left(e^{-\alpha_1 |\nabla I_i|} |\nabla D_i| + e^{-\alpha_2 |\nabla^2 I_i|} |\nabla^2 D_i| \right), \quad (5)$$

where $\alpha_1 = 0.5$, $\alpha_2 = 0.5$. N denotes the total number of the pixels.

3.3.2 Cross-view Consistency Loss

Besides the above brightness constancy loss, we also apply a new cross-view consistency loss by considering the consistency between the images and depth maps for these views. We introduce the following two losses: cross-view consistency loss \mathcal{L}_c and multi-view brightness consistency loss \mathcal{L}_b ,

$$\mathcal{L}_{consistency} = \sum_{i=1}^{v-1} \sum_{j=i+1}^v (\mathcal{L}_c^{i,j} + \sum_{k \neq j}^v \mathcal{L}_b^{i,j,k}). \quad (6)$$

The cross-view consistency loss consists of image consistency loss \mathcal{L}_m based on images and depth consistency loss \mathcal{L}_d based on depth maps. It can be formulated as:

$$\mathcal{L}_c^{i,j} = \lambda_5 \cdot (\mathcal{L}_m^{i,j} + \mathcal{L}_m^{j,i}) + \lambda_6 \cdot (\mathcal{L}_d^{i,j} + \mathcal{L}_d^{j,i}), \quad (7)$$

where $\lambda_5 = 0.3$, $\lambda_6 = 0.3$.

Given two images I_i and I_j , we can produce a synthesized image $I'_{j \rightarrow i}$ by using I_j , D_j and the relative pose between them. Naturally, we can also produce the secondary synthesized image $I''_{i \rightarrow j}$ using $I'_{j \rightarrow i}$, D_i and their relative pose. Suppose that the predicted depth maps are accurate, then the discrepancy between $I''_{i \rightarrow j}$ and I_j should be very small and vice versa. In order to alleviate the robustness of the consistency loss, we also introduce another term to access the difference between $I''_{i \rightarrow j}$ and I_j . The cross-view image consistency loss \mathcal{L}_m is defined as:

$$\mathcal{L}_m^{i,j} = \mathcal{L}_{synthesis}(I_j, I''_{i \rightarrow j}). \quad (8)$$

For the sake of robustness, we exploit the constraint between the predicted depth map D_i and the synthesized depth map $D'_{j \rightarrow i}$. Therefore, the cross-view depth consistency loss \mathcal{L}_d is defined as:

$$\mathcal{L}_d^{i,j} = \varphi(D_i - D'_{j \rightarrow i}) \cdot M, \quad (9)$$

Besides the above consistency loss, we also present the multi-view brightness consistency loss to enhance the relationship of other views relative to the reference view. Our multi-view brightness consistency loss is formulated as:

$$\mathcal{L}_b^{i,j,k} = \mathcal{L}_{synthesis}(I'_{j \rightarrow i}, I''_{k \rightarrow i}), \quad (10)$$

which evaluates the brightness constancy across views i, j, k , i.e., multi-view consistency.

4. Experimental Results

To evaluate the performance of our proposed network MVS², we conducted experiments on widely used multi-view stereo datasets, e.g. DTU [1], SUN3D, RGBD, MVS and Scenes11¹. To align with other related works, we only trained our network on the training set of the DTU dataset, and directly tested on other datasets.

4.1. Implementation Details

Dataset: The DTU dataset[1] is a large-scale multi-view stereo dataset, which consists of 128 scenes and each scene contains 49 images with 7 different lighting conditions. For a fair comparison, we follow the experimental setting in [30]. We generate the ground truth depth maps from the point cloud with the screened Poisson surface reconstruction method [15]. We choose scenes: 1, 4, 9, 10, 11, 12, 13, 15, 23, 24, 29, 32, 33, 34, 48, 49, 62, 75, 77, 110, 114, 118 as the testing set and the other scenes as training set. The RGBD, SUN3D, MVS and Scenes11 datasets contain more than 30000 different scenes in total, which are very different from the DTU dataset. We use these datasets to validate the powerful generalization ability of our network.

¹<https://github.com/lmb-freiburg/demon>

Table 1. **Quantitative results on the DTU’s evaluation set [1].** We evaluate all methods using both the distance metric [1] (lower is better), and the percentage metric [16] (higher is better) with 1mm thresholds

	Accuracy(mm)			Completeness(mm)			overall	Percentage (<1mm)		
	Mean.	Median.	Variance	Mean.	Median.	Variance		Acc.	Comp.	<i>f-score</i>
Camp [2]	0.835	0.482	1.549	0.554	0.523	4.076	0.695	71.75	64.94	66.31
Furu [6]	0.612	0.324	1.249	0.939	0.463	3.392	0.775	69.55	61.52	63.26
Tola [25]	0.343	0.210	0.385	1.190	0.492	5.319	0.766	90.49	57.83	68.07
SurfaceNet[12]	0.450	0.254	1.270	1.043	0.285	5.594	0.746	83.80	63.38	69.95
MVSNet [30]	0.396	0.286	0.436	0.741	0.399	2.501	0.592	86.46	71.13	75.69
MVS ² (ours)	0.760	0.485	1.791	0.515	0.307	1.121	0.637	70.56	66.12	68.27

Training Details: Our MVS² network is implemented in Tensorflow with an NVIDIA v100 GPU. We train our model on the DTU’s training set, but test it on the DTU’s test set and other datasets directly. The image resolution for the DTU dataset is 640×512 . The resolution of the predicted depth map is one-quarter of the original input due to down-sampling. The depth ranges are uniformly sampled from 425mm to 935mm with a resolution of 2.6mm and the depth sample number is $D = 192$. For other datasets, in order to align the depth range, we set the depth start from 0.5mm with the depth sample resolution of 0.25, and the number of depth sample is $D = 128$.

For the hyper-parameters, we set $\omega_w = 0.8$, $\omega_s = 0.1$ throughout the experiments. The batch size is set to 1 due to memory limit. The models are trained with RMSP optimizer for 10 epochs, with the learning rate of $2e-4$ for the first 2 epochs and decreased by 0.9 for every two epochs.

Error Metrics: We use the standard metrics used in a public benchmark suite for performance evaluation. These quantitative measures include absolute relative error (Abs Rel), absolute difference error (Abs diff), square relative error (Sq Rel), root mean square error and its log scale (RMSE and RMSE log) and inlier ratio ($\delta < 1.25^i$, $i = 1, 2, 3$).

4.2. Comparison with SOTA Methods

To verify the performance of our MVS², we tested it on the widely used DTU dataset. First, we conducted extensive quantitative comparisons with the state-of-the-art (SOTA) methods published recently. Performance comparison with other SOTA MVS methods is reported in Tab. 1. From Tab. 1, we can conclude that MVS² achieves higher completeness than other SOTA MVS methods while achieving comparable performance under other metrics. We applied a depth map fusion step to integrate the depth maps from different views to a unified point cloud representation. We chose the gipuma [7] to fuse our depth maps. The qualitative comparisons in 3D reconstruction are shown in Fig. 6, where MVS² achieves 3D reconstruction comparable with state-of-the-art supervised MVS method [30].

4.3. Ablation Studies

To analyze the contribution of different modules of our network model, we conduct three ablation studies on the DTU validation set with $W \times H \times D = 640 \times 512 \times 192$.



(a) Ground Truth (b) Yao’s result [30] (c) Our result

Figure 6. **Qualitative comparison in 3D reconstruction between our MVS² and SOTA supervised MVS method [30] on the DTU dataset.** From left to right: ground-truth point clouds, Yao’s point clouds [30], our point clouds. Best Viewed on Screen.

Quantitative results are reported in Tab. 2.

SPN Refinement. Under our network model, we introduce the spatial propagation network (SPN) [20] to refine the initial depth map. To analyze the contribution of this module, we conduct experimental comparison with and without this module and the results are reported in Tab. 2. It can be observed that when the SPN module is removed, the performance consistently drops. For example the Abs Diff increases from 11.3912 to 13.0339, and the Abs Rel increases from 0.0147 to 0.0175, which clearly demonstrates the effectiveness of the SPN refinement module.

Cost Volume. In building the cost volume, we exploit both the feature for the current view and the variance-based feature [30]. To validate the effectiveness of our cost volume construction, we compare with a baseline implementation by using the variance-based feature only, which is used in [30]. As illustrated in Tab. 2, when the feature for the current view is excluded from the cost volume, the performance consistently drops. For example the Abs Rel jumps from 0.0147 to 0.0204 and the Abs Diff increases from 11.3912 to 15.1751. The experimental results prove the effectiveness of our proposed cost volume reconstruction.

Table 2. **Ablation Experiments.** (a) Without spatial propagation refine module, only compute self-supervised loss through initial depth map generated by inference network. (b) With cost volume generated just by the variance-based homography feature. (c) Without the view consistency self-supervised loss. (d) Our complete MVS². Datasets: DTU dataset

w/o	Error metric					Accuracy metric($\delta < \alpha^4$)			
	Abs Rel	Abs Diff	Sq Rel	RMSE	RMSE log	α	α^2	α^3	runtime
(a)SPN Refine	0.0175	13.0339	1.8440	30.4543	0.0187	0.9814	0.9992	1.0000	0.313s
(b)Cost(difference)	0.0204	15.1751	2.3806	34.6147	0.0251	0.9753	0.9986	1.0000	0.273s
(c)view consistency	0.0355	24.9464	5.2399	55.4236	0.0425	0.9482	0.9920	0.9998	0.322s
(d)MVS ² (ours)	0.0147	11.3912	1.5478	28.4428	0.0156	0.9900	1.0000	1.0000	0.325s

Table 3. **Ablation experiments with different combinations of consistency methods.** (a) With only the multi-view brightness consistency (BC) loss. (b) With only the cross-view consistency check (CC) loss. (c) With all loss. Datasets: DTU dataset

	Error metric				
	Abs Rel	Abs Diff	Sq Rel	RMSE	RMSE log
(a) BC	0.0180	13.4363	1.7042	30.2826	0.0190
(b) CC	0.0172	12.8649	1.6311	29.4134	0.0182
(c) BC and CC	0.0147	11.3912	1.5478	28.4428	0.0156

tion method in exploiting the feature of the current view.

Consistency Loss. In this paper, we have proposed a consistency loss to further constrain the multiple estimated depth maps, which is also a key contribution. To analyze the contribution of this consistency loss, we conducted experiments with and without this loss term and the results are reported in Tab. 2. When the cross-view consistency loss is removed from our unsupervised loss, the performance deteriorates sharply. For example the Abs Rel shoots up from 0.0147 to 0.0355 while the Abs Diff increases from 11.3912 to 24.9464 and the Sq Rel jumps from 1.5478 to 5.2399. The experimental results clearly demonstrate the significance of our proposed consistency loss.

Besides the above ablation studies in analyzing the contribution of our novel consistency loss term, as our consistency term actually consists of two terms (multi-view brightness consistency and cross-view consistency in depth maps), we also conducted two additional experiments to analyze the effectiveness of each term and the corresponding results are reported in Tab. 3. From Tab. 3, we could draw the following conclusions that: 1) Both the multi-view brightness consistency term (BC) and the cross-view consistency term (CC) are critical for achieving improved performance; 2) The cross-view consistency term (CC) plays a more important role than the multi-view brightness consistency term (BC) in depth map estimation.

4.4. Generalization Ability

As agreed in monocular depth estimation and binocular stereo matching, the supervised depth estimation methods strongly depend on the availability of large scale ground truth 3D data and the generalization ability could be hindered when evaluated on never-seen-before open-world scenarios. Here, we would like to verify the generalization ability of our unsupervised MVS network model.

We conducted experiments on SUN3D, RGBD, MVS and Scenes11 datasets using our pre-trained model without any fine tuning. In Table 4, we compare the performance of our MVS² with state-of-the-art traditional MVS methods and supervised MVS methods. We can conclude from Table 4 that: 1) Our MVS² outperforms state-of-the-art traditional geometry-based multi-view method COLMAP[22] with a wide margin, which shows the benefits in exploiting the large scale datasets; 2) Compared with supervised MVS methods trained on each dataset individually, our MVS², even only trained on the DUT training dataset, outperforms current state-of-the-art supervised MVS method DeepMVS on part of the error metrics. Qualitative comparison between our MVS² and competing MVS methods (COLMAP, DeMoN, DeepMVS) on the RGBD dataset is demonstrated in Fig. 8, where our method consistently achieves compared performance with SOTA supervised methods.

We also conducted experiments on the Tanks and Temples datasets without any fine tuning to validate the generalization ability of our network model. We choose $N = 3$, $W = 1920$, $H = 1024$ and $D = 192$ for our experiments. Qualitative point cloud results are presented in Fig. 7, where our MVS² could reconstruct very detailed 3D structures.

5. Conclusions

In this paper, we have proposed the first unsupervised learning based MVS network, which learns the depth map for each view simultaneously without the need of ground truth 3D data. With our proposed multi-view symmetry network design, we can enforce the cross-view consistency of depth maps during training and testing. Our learned multi-view depth maps comply with the underlying 3D geometry. Our network learns multi-view occlusion maps in an alternative way. Experimental results on multiple benchmarking datasets demonstrate the effectiveness and excellent generalization ability of our network. In the future, we plan to extend the depth consistency beyond pairwise relation, such as consistency inside a clique. Extension to dynamic scenes [13] could be another interesting future direction.

Acknowledgement

This research was supported in part by the Natural Science Foundation of China grants (61871325, 61420106007,

Table 4. **Generalization Ability.** Multi-view stereo methods: COLMAP, DeepMVS, DeMoN, where Deep MVS and DeMoN are supervised methods and trained on these datasets correspondingly.

Datasets	Method	Abs Rel	Abs Diff	Error metric			Accuracy metric($\delta < \alpha^2$)		
				Sq Rel	RMSE	RMSE log	α	α^2	α^3
SUN3D	COLMAP[22]	0.6232	1.3267	3.2359	2.3162	0.6612	0.3266	0.5541	0.7180
	DeMoN[26]	0.2137	2.1477	1.1202	2.4212	0.2060	0.7332	0.9219	0.9626
	DeepMVS[11]	0.2816	0.6040	0.4350	0.9436	0.3633	0.5622	0.7388	0.8951
	MVS ² (ours)	0.3488	0.5956	0.4879	0.7525	0.3805	0.4930	0.7616	0.9100
RGBD	COLMAP [22]	0.5389	0.9398	1.7608	1.5051	0.7151	0.2749	0.5001	0.7241
	DeMoN [26]	0.1569	1.3525	0.5238	1.7798	0.2018	0.8011	0.9056	0.9621
	DeepMVS [11]	0.2938	0.6207	0.4297	0.8684	0.3506	0.5493	0.8052	0.9217
	MVS ² (ours)	0.4414	0.8698	0.9352	1.2853	0.4726	0.4657	0.6878	0.8057
MVS	COLMAP [22]	0.3841	0.8430	1.257	1.4795	0.5001	0.4819	0.6633	0.8401
	DeMoN[26]	0.3105	1.3291	19.970	2.6065	0.2469	0.6411	0.9017	0.9667
	DeepMVS [11]	0.2305	0.6628	0.6151	1.1488	0.3019	0.6737	0.8867	0.9414
	MVS ² (ours)	0.3729	0.8170	0.9135	1.3938	0.4921	0.5136	0.6952	0.9123
SCENES11	COLMAP[22]	0.6249	2.2409	3.7148	3.6575	0.8680	0.3897	0.5674	0.6716
	DeMoN [26]	0.5560	1.9877	3.4020	2.6034	0.3909	0.4963	0.7258	0.8263
	DeepMVS [11]	0.2100	0.5967	0.3727	0.5909	0.2699	0.6881	0.8940	0.9687
	MVS ² (ours)	0.5981	2.0848	3.3365	2.9477	0.4885	0.4695	0.6531	0.7879

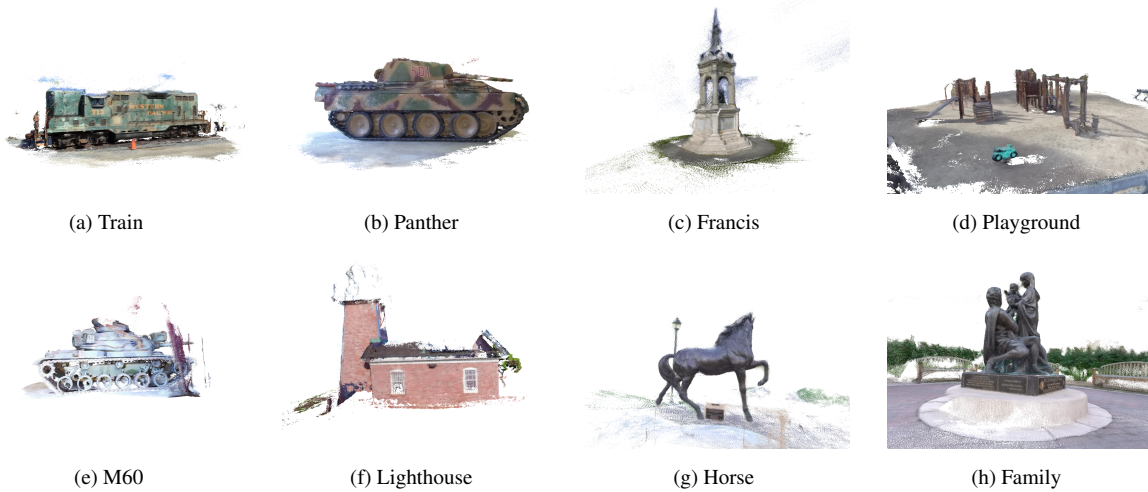


Figure 7. **3D point clouds generated by our MVS² without any finetuning on the Tanks and Temples dataset.**

61671387). We thank all anonymous reviewers for their valuable comments.

References

- [1] Henrik Aans, Rasmus Ramsbl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016.
- [2] Neill D. F. Campbell, George Vogiatzis, Carlos Hernandez, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conference on Computer Vision*, 2008.
- [3] Yasutaka Furukawa, Brian Curless, Steven M. Seitz, and Richard Szeliski. Manhattan-world stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- [4] Yasutaka Furukawa, Brian Curless, Steven M. Seitz, and Richard Szeliski. Towards internet-scale multi-view stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
- [5] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 9(1-2):1–148, 2015.
- [6] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009.
- [7] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *IEEE International Conference on Computer Vision*, 2015.
- [8] David Gallup, Jan Michael Frahm, and Marc Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
- [9] C. Godard, O. M. Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

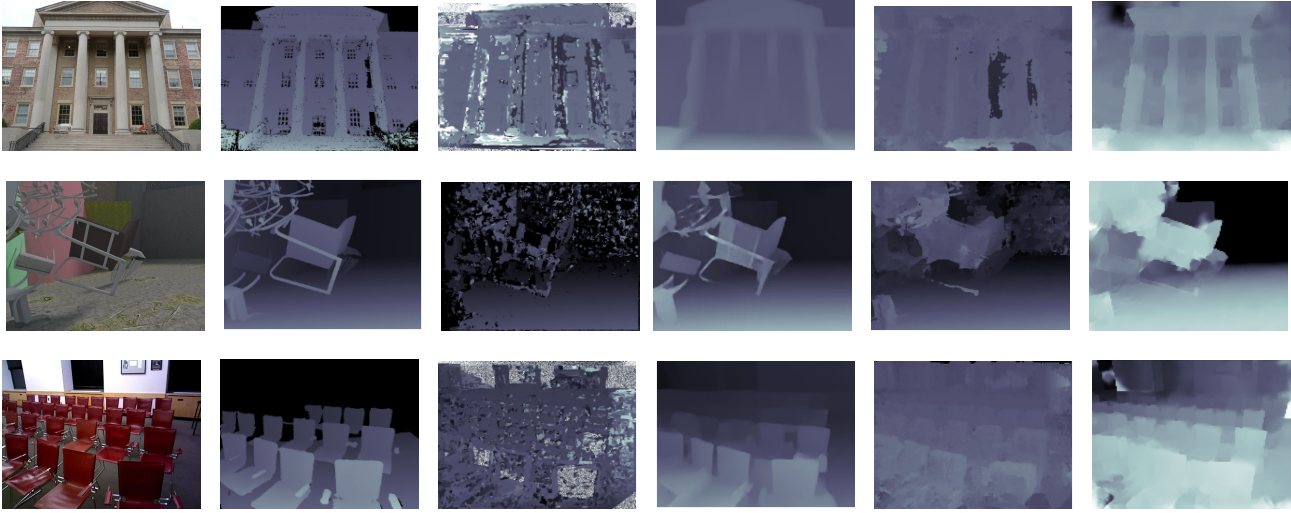


Figure 8. **Performance comparison on depth maps estimation with other supervised methods:** From Left to Right: reference image, ground truth depth map, depth map of COLMAP, depth map of DeMoN, depth map of DeepMVS and our result.

- IEEE Conference on Computer Vision and Pattern Recognition*, pages 6602–6611, July 2017.
- [10] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M. Seitz. Multi-view stereo for community photo collections. In *IEEE International Conference on Computer Vision*, pages 1–8, 2007.
 - [11] P. Huang, K. Matzen, J. Kopf, N. Ahuja, and J. Huang. Deepmvs: Learning multi-view stereopsis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018.
 - [12] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. SurfacerNet: An end-to-end 3d neural network for multiview stereopsis. In *IEEE International Conference on Computer Vision*, 2017.
 - [13] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):190–204, Jan 2019.
 - [14] Abhishek Kar, Christian Hne, and Jitendra Malik. Learning a multi-view stereo machine. In *Advances in Neural Information Processing Systems*, 2017.
 - [15] Michael M. Kazhdan, Matthew Bolitho, and Hugues Hoppe. Screened poisson surface reconstruction. *Acm Transactions on Graphics*, 32(3):1–13, 2013.
 - [16] Arno Knapitsch, Jaesik Park, Qian Yi Zhou, and Vladlen Koltun. Tanks and temples: benchmarking large-scale scene reconstruction. *Acm Transactions on Graphics*, 36(4):78, 2017.
 - [17] Fabian Langguth, Kalyan Sunkavalli, Sunil Hadap, and Michael Goesele. Shading-aware multi-view stereo. In *European Conference on Computer Vision*, pages 469–485, 2016.
 - [18] Bo Li, Yuchao Dai, and Mingyi He. Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference. *Pattern Recognition*, 83:328–339, 2018.
 - [19] Bo Li, Chunhua Shen, Yuchao Dai, Anton van den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015.
 - [20] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. In *Advances in Neural Information Processing Systems*, pages 1520–1530, 2017.
 - [21] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5667–5675, June 2018.
 - [22] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518, 2016.
 - [23] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *European Conference on Computer Vision*, 2010.
 - [24] Im Sunghoon, Jeon Hae-Gon, Lin Stephen, and Kweon In, So. Dpsnet: End-to-end deep plane sweep stereo. In *International Conference of Learning Representation*, 2019.
 - [25] Engin Tola, Christoph Strecha, and Pascal Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision & Applications*, 23(5):903–920, 2012.
 - [26] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox. Demon: Depth and motion network for learning monocular stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5622–5631, 2017.

- [27] Kaixuan Wang and Shaojie Shen. Mvdepthnet: Real-time multiview depth estimation neural network. In *International Conference on 3D Vision*, pages 248–257, 2018.
- [28] Chenglei Wu, Bennett Wilburn, Yasuyuki Matsushita, and Christian Theobalt. High-quality shape from multi-view stereo and shading under general illumination. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
- [29] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European Conference on Computer Vision*, pages 842–857, 2016.
- [30] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision*, 2018.
- [31] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.
- [32] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, and Thomas Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.
- [33] Yiran Zhong, Yuchao Dai, and Hongdong Li. Self-supervised learning for stereo matching with self-improving ability. In *arXiv preprint*, 2017.
- [34] Yiran Zhong, Hongdong Li, and Yuchao Dai. Open-world stereo video matching with deep rnn. In *European Conference on Computer Vision*, September 2018.
- [35] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [36] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *European Conference on Computer Vision*, 2018.