Semi-supervised Deep Multi-view Stereo

Hongbin Xu†
South China University of Technology
Guangzhou, China
hongbinxu1013@gmail.com

Weitao Chen†
Alibaba Group
Hangzhou, China
hillskyxm@gmail.com

Yang Liu
Alibaba Group
Hangzhou, China

Zhipeng Zhou
Alibaba Group
Hangzhou, China

Haihong Xiao
South China University of Technology
Guangzhou, China

Baigui Sun
Alibaba Group
Hangzhou, China

Xuansong Xie
Alibaba Group
Hangzhou, China

Wenxiong Kang‡‡
South China University of Technology
Guangzhou, China
auwxkang@scut.edu.cn

ABSTRACT
Significant progress has been witnessed in learning-based Multi-view Stereo (MVS) under supervised and unsupervised settings. To combine their respective merits in accuracy and completeness, meantime reducing the demand for expensive labeled data, this paper explores the problem of learning-based MVS in a semi-supervised setting that only a tiny part of the MVS data is attached with dense depth ground truth. However, due to huge variation of scenarios and flexible settings in views, it may break the basic assumption in classic semi-supervised learning, that unlabeled data and labeled data share the same label space and data distribution, named as semi-supervised distribution-gap ambiguity in the MVS problem. To handle these issues, we propose a novel semi-supervised distribution-augmented MVS framework, namely SDA-MVS. For the simple case that the basic assumption works in MVS data, consistency regularization encourages the model predictions to be consistent between original sample and randomly augmented sample. For further troublesome case that the basic assumption is conflicted in MVS data, we propose a novel style consistency loss to alleviate the negative effect caused by the distribution gap. The visual style of unlabeled sample is transferred to labeled sample to shrink the gap, and the model prediction of generated sample is further supervised with the label in original labeled sample. The experimental results in semi-supervised settings of multiple MVS datasets show the superior performance of the proposed method. With the same settings in backbone network, our proposed SDA-MVS† outperforms its fully-supervised and unsupervised baselines.

CCS CONCEPTS
• Computing methodologies → Reconstruction.

KEYWORDS
3D Reconstruction, multi-view stereo, neural networks, semi-supervised learning

ACM Reference Format:
Hongbin Xu, Weitao Chen[1], Yang Liu, Zhipeng Zhou, Haihong Xiao, Baigui Sun, Xuansong Xie, and Wenxiong Kang. 2023. Semi-supervised Deep Multi-view Stereo. In Proceedings of the 31st ACM International Conference on Multimedia (MM ’23), October 29-November 3, 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611931

1 INTRODUCTION
Multi-view Stereo (MVS) is one of the cornerstone problems in computer vision, which reconstructs dense 3D geometry from calibrated multi-view images. Stereoscopic vision for 3D reconstruction is on the cusp of many industrial applications such as autonomous driving, robotics, and virtual reality for decades. Recent MVS works [39, 47, 48] extend the traditional approaches to deep-learning based methods, and improve the 3D reconstruction performance with the blessing of large-scale MVS datasets [14, 19]. Despite their ideal performance, there have been non-negligible difficulties in collecting dense 3D ground truth annotations, which may hamper the generalization to new domains. Specifically, collecting accurate and complete 3D ground truth [14, 19] requires tedious collection process with a fixed active sensor, as well as labor-intensive post-processing procedures to remove outliers like moving objects in a static scene. Thus, unsupervised/self-supervised MVS methods are proposed to avoid the dependence on the expensive 3D ground truth, which build the depth estimation problem as an image reconstruction problem with photometric consistency [13, 17, 41, 42]. With the help of these methods, the perplexity of 3D annotations

†Both authors contributed equally to this work.
‡Corresponding author.
‡‡Also with Pazhou Laboratory.
can be relieved, meantime achieving amazing 3D reconstruction quality [41].

Rethinking the merits and demerits of unsupervised and supervised MVS compared with each other, we can have the following findings: 1) Considering 3D reconstruction completeness, unsupervised MVS performs better than supervised MVS. Since the self-supervision loss built on photometric consistency excavates supervision signals on all available pixels in the image, unsupervised MVS has more complete regions with valid supervision constraints compared with supervised MVS which only has limited label-intensive annotations.

2) Considering 3D reconstruction accuracy, supervised MVS performs better than unsupervised MVS. Different from the valid supervision in supervised MVS, the dense self-supervision loss is usually not accurate enough, because it may be invalid on many unexpected cases, such as color constancy ambiguity [41], textureless backgrounds [42] and occluded regions [7].

Instead of merely staring at the demerits of unsupervised and supervised MVS methods for improvements, we can see that they are complementary to each other on their respective merits of improving completeness and accuracy. In this paper, to combine the merits of unsupervised and supervised MVS, we explore a novel semi-supervised MVS problem, which assumes that only a tiny part of the MVS dataset has 3D annotations. Specifically, it has an intractable risk of breaking the basic assumption in the standard semi-supervised classification problem [12, 26, 40], that labeled and unlabeled data come from the same label space, following independently identical distribution (i.i.d.). As shown in Fig. 1 (a)), the inherent distribution gap among different scenes in MVS problem may confuse the learning process, namely semi-supervised distribution-gap ambiguity.

To handle the problems, we propose a novel MVS framework, called semi-supervised distribution-augmented MVS framework (SDA-MVS). 1) The basic framework of SDA-MVS handles the labeled samples and unlabeled samples differently. The labeled samples are supervised under the common regime of supervision loss [47] measuring the difference between the prediction and ground truth. The basic photometric consistency loss [17] is used to supervise the unlabeled samples. No extra extensions [7, 41, 42] of the self-supervision loss are used to maintain a concise pipeline. 2) For the simple case that the assumption works, consistency regularization loss is used to minimize the difference of depth predictions with or without random data-augmentation. Following the low-density assumption [12], the low-density separation boundary among classes is enforced through the invariance against data-augmentations and proximity in latent space, meantime spreading the priors from labeled data to unlabeled data. 3) For further troublesome case that the assumption fails, we propose a style consistency loss consisting of a style translation module (STM) and geometry-preserving module (GPM). Taking inspiration from neural style transfer algorithms [11, 21], STM transfers the visual styles from unlabeled MVS images

Figure 1: Visualization proofs of the semi-supervised MVS problem.
to labeled MVS images. However, the style transfer algorithms may bring unexpected distortions in the generated images, which may corrupt the cross-view correspondence relationship in the MVS data (further discussed in Fig. 1 (b)). Consequently, GPM utilizes a spatial propagation network [21] to regularize the affinity of images, acting as an anti-distortion module towards unexpected distortions. The ground truth is then used to supervise the generated MVS images after style translation, diminishing the negative effect of distribution discrepancy between labeled and unlabeled MVS data.

In summary, our contributions are listed as follows: 1) In a semi-supervised setting which assumes only a small part of the MVS dataset are labeled, we firstly investigate a novel semi-supervised distribution-perturbed problem and propose a novel MVS framework named SDA-MVS; 2) To handle the natural distribution gap between labeled and unlabeled MVS data, we propose a style consistency loss to alleviate the problem. 3) For evaluation, the experimental results on DTU, BlendedMVS, GTA-SFM, and Tanks&Temples demonstrate the superior performance of the proposed method. We further extend the semi-supervised setting to the semi-supervised domain adaptation task in multiple MVS datasets and evaluate the effectiveness of SDA-MVS.

2 RELATED WORK

2.1 Fully-supervised Multi-view Stereo

Thanks to the bless of deep neural networks, learning-based methods have been successfully developed on MVS reconstruction. The pioneering work of MVSNet [47] proposes an end-to-end network for multi-view depth estimation. The multi-view feature maps are extracted with a shallow Convolutional Neural Network (CNN) with shared weights. Then, the cross-view features are projected to the reference frustum via differentiable homography warping to construct a cost volume following the regime of plane sweeping [27]. On each hypothetical depth plane, each element of cost volume represents the similarity score of the matching points among views. After regularizing by 3D CNN, the predicted probability volume is used to regress depth map via soft-argmin [16]. However, the huge computation consumption and memory footprint caused by cost volume and 3D CNN of MVSNet may limit the potential of handling high-resolution images and the performance of 3D reconstruction. Consequently, lots of efforts have been devoted to handling these issues, which can be divided into 2 categories: Recurrent-based methods [38, 45, 48] and Coarse-to-fine methods [5, 6, 24, 33, 46, 51, 53]. The recurrent-based methods replace the 3D CNN with recurrent neural networks (RNN) to regularize the cost volume. The coarse-to-fine methods separate the single-stage cost volume regularization process of MVSNet into multiple stages.

2.2 Self-supervised Multi-view Stereo

In aware of the expensive and time-consuming process for collecting ground truth depth maps in MVS tasks, a recent strand of work in unsupervised/self-supervised MVS methods strive to remove the reliance on ground truth and replace the depth loss regression with an image reconstruction loss built upon photometric consistency [17]. In Unsup_MVS [17], the predicted depth map of MVSNet is used to reconstruct the reference image from source images via homography warping. The self-supervised training follows the assumption that the reconstructed image from other views should be similar to the original image if the depth map is correct. Although the self-supervision loss provides a promising alternative to supervised loss, it is not accurate enough and may be confused by many unexpected problems, such as occlusion ambiguity [7], color constancy ambiguity [41], and textureless ambiguity [42].

2.3 Semi-supervised Learning and Multi-view Stereo

In recent years, immense progress has been witnessed in semi-supervised learning, especially in image classification. Following the continuity assumption of semi-supervised learning [1, 30], consistency regularization applies random data augmentation to semi-supervised learning by leveraging the idea that a classifier should output the same class distribution for an unlabeled example even after it has been augmented. The basic consistency loss [29] in semi-supervised frameworks, such as Π-model [28], Mean Teacher [34], Unsupervised Data Augmentation [40] and MixMatch [2] is the l-2 loss as follows:

\[
Ω(x; θ) = ||p_{model}(y|perturb(x); θ) − p_{model}(y|x; θ))||^2_2
\] (1)

Note that perturb \(x\) is a stochastic transformation, hence the two terms in Eq. 1 are not identical. Consistency regularization enforces the unlabeled example \(x\) to be classified the same as perturb \(x\), a random augmentation of itself. Whereas, different from the standard classification setting in semi-supervised learning, the Semi-MVS problem in this paper has to face huge variation of scenes in the MVS dataset, which may break the continuity assumption of labeled and unlabeled data distribution. Consequently, further improvements are required in Semi-MVS problem.

The most recent work with a similar topic of semi-supervised setting in MVS, is SGT-MVSNet [18]. However, Sparse Ground-Truth based MVS (SGT-MVS) problem has a basic premise that the labeled pixels and unlabeled pixels are picked from the same images, which can be assumed to follow the same data distribution. Different from SGT-MVS, the core problem of this work is how to propagate the supervision signal on labeled samples to unlabeled samples which may have a distribution gap.

3 METHOD

3.1 Problem Definition

Given a pair of multi-view images with \(N\) calibrated views, the reference image is denoted as \(I_1\) and the \(v\)-th source view is denoted as \(\{I_j\}_{j=2}^N\). The intrinsic and extrinsic parameters on view \(v\) are defined as \(\{K_v\}_{v=1}^N\) and \(\{T_{uv}\}_{u=1}^N\) respectively. The ground truth depth map on the reference view is noted as \(D\). A labeled sample is \(S^l = \{(I^l_j, K^l_v, T^l_{uv})_{j=1}^N, D^l\}\) and an unlabeled sample is \(S^u = \{(I^u_j, K^u_v, T^u_{uv})_{j=1}^N\}\). Assume that \(M\) samples are available in the whole MVS dataset, comprised of \(\mu\) labeled sample \(S^l\) and \((1 − \mu)\) unlabeled sample \(S^u\). Considering the difficulties in collecting dense depth ground truth, \(\mu\) is set to a small ratio of 0.1 in default, which creates a challenging task since only an extremely small ratio of ground truth is available.
### 3.2 Geometry-dropping Problem

As discussed in Sec. 1, the distribution gap may break the semi-supervised assumption due to the huge variation among scenarios. Taking inspiration from neural style transfer, we aim to transfer the visual style from unlabeled data to labeled data, trying to shrink this gap. However, another problem of losing 3D geometric details occurs when neural style transfer algorithms are directly applied to MVS images, as shown in Fig. 1(b). The results (2nd row) of STM embedded with standard neural style transfer algorithms lack geometric details on the zoomed region and the reconstructed 3D point cloud is much sparser compared with the original ones (1st row). Putting the cart before the horse, the lost details in STM may reversely degrade the performance in MVS problem. To handle this issue, we further utilize GPM to handle this problem as an image distortion problem.

### 3.3 Overall Pipeline

In Fig. 2, the overall framework of our proposed SDA-MVS is presented. As shown in Fig. 2(a), labeled sample $S^l$ and unlabeled sample $S^u$ are randomly selected from the labeled and unlabeled dataset respectively in the data preparation process. Then the labeled and unlabeled sample are fed to STM and GPM to generate style augmented sample $S^f$. Afterwards, as shown in Fig. 2(b), the labeled sample $S^l$ is supervised under standard supervision loss (Eq. 2). The unlabeled sample $S^u$ is supervised under unsupervised loss (Eq. 5) comprised of photometric consistency loss and consistency regularization loss. The style augmented sample $S^f$ is enforced to satisfy the style consistency regularization framework (Sec. 3.6).

### 3.4 Backbone

Arbitrary MVS network can be utilized as the backbone of the proposed semi-supervised framework, i.e. MVSNet [47], CasMVSNet [39], etc. In default, the representative CasMVSNet [39] is used. The MVS network requires $N$ multi-view images as input. The feature map extracted by CNN with shared weights on each view is reprojected to the same reference view with differentiable homography warping. The variance among the feature maps on different views is calculated to construct the cost volume, and a 3D U-Net is utilized to regularize the predicted probability volume $PV$. The predicted depth map $D_{pred}$ is finally regressed with soft-argmin operation.

### 3.5 Depth Consistency Regularization

#### 3.5.1 Supervised Loss

The labeled sample is denoted as $S^l = \{(I^l_v, K^l_v, T^l_v)_{v=1}^N, D^l_{gt}\}$. Following a standard supervised approach [47], the L2 loss between the predicted depth map $D_{pred}$ of the backbone network and the ground truth depth map $D_{gt}$ on all valid pixels is minimized:

$$L_{sup} = \frac{\sum_{i=1}^{HW} \mathbb{1}(D^l_{gt}(p_i) > 0) \|D_{pred}(p_i) - D^l_{gt}(p_i)\|^2_2}{\sum_{i=1}^{HW} \mathbb{1}(D^l_{gt}(p_i) > 0)} \tag{2}$$

where $i$ represents the index of available pixels in the $H \times W$ image, and $p_i$ is the pixel coordinate. $\mathbb{1}(D^l_{gt}(p_i) > 0)$ is the indicator function which represents whether valid depth ground truth exists in current pixel $p_i$. Note that all invalid pixels in the provided ground truth depth map are set to 0.

#### 3.5.2 Photometric Consistency Loss

The unlabeled sample is denoted as $S^u = \{(I^u_v, K^u_v, T^u_v)_{v=1}^N, I^u_{gt}\}$. With the homography warping function, pixel $p_i^u$ in the reference image $I^u_1$ corresponds to pixel $\hat{p}_i^v$ in the $v$-th source view image $I^u_v$.

$$D_v(\hat{p}_i^v) \hat{p}^v_i = D_v(I^u_1) I^u_1^{-1} D_{pred}(p_i^u) p_i^u \tag{3}$$

where $i(1 \leq i \leq HW)$ is the pixel index of $H \times W$ image. $D_v$ represents the depth value on view $v$, and $D_{pred}$ is the predicted depth map from unlabeled sample $S^u$. Since the $D_v(\hat{p}_i^v)$ is a scale term in homogeneous coordinates, we can normalize Eq. 3 to obtain the pixel coordinate $\tilde{p}_i^v$:

$$\tilde{p}_i^v = \pi(D_v(\hat{p}_i^v), \pi([x, y, z]^T) = [x/z, y/z, 1]^T \tag{4}$$

With the correspondence relationship determined by Eq. 4, the image on the reference view can be reconstructed via images on source view $v$: $I^u_{gt-1}(p_i^1) = I^u_{gt}(\tilde{p}_i^1)$. Thus, the reconstructed image $I^u_{gt-1}$ is enforced to be the same as original image $I^u_1$ following photometric consistency:

$$L_{photo} = \sum_{v=2}^V \sum_{i=1}^{HW} \mathbb{1}(1 \leq \tilde{p}_i^v \leq [H, W]) \|I^u_{gt-1}(p_i^1) - I^u_{gt}(\tilde{p}_i^v)\|^2_2 \sum_{i=1}^{HW} \mathbb{1}(1 \leq \tilde{p}_i^v \leq [H, W]) \tag{5}$$
where \( I(1 ≤ p_i ≤ [H, W]) \) indicates whether the current pixel \( p_i \) can find valid pixel \( \tilde{p}_i \) in other view.

### 3.5.3 Consistency Regularization

The general form of consistency regularization compute the divergence between the two predicted outputs of original sample and perturbed sample. Denote that the perturbed version of unlabeled images \( \tilde{I}_{uv}^s = \phi(I_{uv}^s, \epsilon) \) by injecting a small noise \( \epsilon \). In MVS, the noise \( \epsilon \) can be applied as hyperparameters controlling various data augmentation transformations like color jitting, gamma correction, image blurring and etc. Similar as VAT [26], we aim to minimize the KL divergence between the predicted distributions on an unlabeled sample \( \{I_{uv}^s\}_{uv=1}^N \) and an augmented unlabeled sample \( \{\tilde{I}_{uv}^s\}_{uv=1}^N \).

As a re-parameterizing trick, the soft-argmin operation [16] in the backbone network actually convert the discrete output of probability volume \( PV \) into a continuous depth map by weighted summing it with all depth hypotheses. Conversely, we can also treat the depth regression task in MVS as a classification task whose predicted classes are predefined depth space. Assume that \( K \) depth hypotheses are predefined in the MVS task, and the probability volume \( PV \) with resolution of \( H \times W \times K \) can be separated into \( HW \) logits with \( K \) categories. In this way, we can simplify the dense depth regression problem into a per-pixel classification problem with \( K \) predefined depth hypothesis(categories), and the probability volume is comprised of the predicted logits, which can be further used in the KL divergence based constraints as follows:

\[
L_{\text{consistency}} = \frac{1}{HW} \sum_{(i,j)} \mathbb{D}_{KL}(PV(p_i)||\tilde{PV}(p_i)) \tag{6}
\]

where \( \mathbb{D}_{KL} \) represents the KL divergence. \( i \) is the index of all \( HW \) pixels in the image, and \( p_i \) is the corresponding pixel coordinate. \( PV \) is the predicted probability volume of unlabeled sample \( \{I_{uv}^s\}_{uv=1}^N \) and \( \tilde{PV} \) is the predicted probability volume of augmented unlabeled sample \( \{\tilde{I}_{uv}^s\}_{uv=1}^N \). For multi-staging MVS network, only the initial stage is used to calculate the consistency regularization loss due to dynamic depth hypotheses in different stages.

### 3.6 Style Consistency Regularization

#### 3.6.1 Style Translation Module (STM)

Based on aforementioned discussions, we aim to transfer the visual style of unlabeled image to labeled image, and shrink the distribution gap. The basic assumption of neural style transfer [11] is that the visual style is encoded by a set of Gram matrices \( \{G^{la}_{1 \leq i \leq N}\} \), where \( G^{la} \in \mathbb{R}^{C_l \times C_a} \) is derived from the feature map \( F^{la} \) of layer \( la \) in a CNN by computing the correlation between activation channels: \( (G^{la}(F^{la}))_{ij} = \sum_k F^{la}_{ik} F^{la}_{jk} \).

The Gram matrix captures semantic information which is irrelevant to position, and more likely to represent semantic visual styles [11]. For simplicity, we refer to a classic method called Whitening and coloring transform (WCT [21]) in STM. WCT solve the style transfer problem with linear transforms on feature maps derived from Gram matrix, which can also be viewed as an eccentric covariance matrix.

Denote that the unlabeled sample image \( I^u \) is viewed as style image and the labeled image \( I^l \) is treated as content image. Then the content feature map on layer \( la \) of VGG [32] is \( F^{la}_{l} = F^{la}(I^l) \) and the style feature map is \( F^{la}_{s} = F^{la}(I^u) \). The general form of WCT is defined as follows:

\[
\tilde{F}^{la}_{cs} = (E_{Dx}D_x^{-1} E_{Dx}^T)(E_{Dx}D_x^{-1} E_{Dx}^T)F^{la}_{s} \tag{7}
\]

where \( E_{Dx}D_x^{-1} E_{Dx}^T \) is called coloring transform and \( E_{Dx}D_x^{-1} E_{Dx}^T \) is called whitening transform. \( D_x \) and \( E_x \) are respectively the diagonal matrix with eigenvalues and the corresponding orthogonal matrix with eigenvectors of covariance matrix \( F^{la}_{cs} F^{la}_{cs}^T = E_x D_x E_x^T \). In analogy, \( D_c \) and \( E_c \) represent eigenvalues and eigenvector of covariance matrix \( F^{la}_{cs} F^{la}_{cs}^T = E_c D_c E_c^T \). The intuition of whitening transform is to peel off the visual style defined by normalizing the content feature map \( F^{la}_{cs} \) while preserving the global content structure. The intuition of coloring transform is the inverse process of whitening transform, and the visual styles of \( F^{la}_{cs} \) are appended to the whitened feature map whose visual style is peeled off in whitening transform. By training an autoencoder on the images with the loss in Eq. 8, the decoder is responsible for inverting transformed features back to the RGB space.

\[
I^u = \text{Dec}(F^{la}(I^u)), I^l = \text{Dec}(F^{la}(I^l)) \tag{8}
\]

The decoder of autoencoder pretrained on the dataset can reconstruct the transformed feature map back into the style augmented image \( I^g \):

\[
I^g = \text{Dec}(\tilde{F}^{la}_{cs}) \tag{9}
\]

#### 3.6.2 Geometry Preserving Module (GPM)

As shown in Fig. 1 (b), directly applying neural style transfer algorithm may lose geometric details which are important for modeling 3D consistency among views in MVS. The reason is that all operations of neural style transfer are processed on feature maps extracted by a VGG network, which is usually over 16 times smaller than the original image. The detailed information modeling the local regions may be lost under such a small resolution, thus unexpected distortions may occur [21]. Consequently, to handle this issue, we utilize the spatial propagation network (SPN) [22] to filter the distortions in the image. SPN is a generic framework that can be applied to many affinity-related tasks. Here, we utilize SPN to model local pixel pairwise relationships, defined by the original image. SPN has 2 branches: propagation network and guidance network. In intuition, the weights of filters are learned through the CNN guidance network, which are further fed to propagation network to filter the distortions. (Please refer to appendix for more details).

The training of the SPN requires original image \( I = (I^u, I^l) \) and reconstructed image with unexpected distortion \( \text{Dec}(F^{la}(I)) \).

The original image is treated as a prior of local affinity and fed to the guidance network, while the distorted image \( \text{Dec}(F^{la}(I)) \) is fed to the propagation module in SPN. The training loss for SPN is shown as follows:

\[
L_{\text{spn}} = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{HW} \sum_{i=1}^{HW} ||I_n(p_i) - \tilde{I}_n(p_i)||_2^2 + \frac{1}{|P_{\text{sparse}}|} \sum_{p_j \in P_{\text{sparse}}} ||I_l(p_j) - \tilde{I}_{l-1}(p_j)||_2^2 \tag{10}
\]

where the style transferred image is calculated by: \( \tilde{I} = \text{SPN(Dec}(F^{la}(I)), I) \). \( P_{\text{sparse}} \) is the sparse point cloud extracted with COLMAP [31] among the multi-view images. Utilizing the sparse 3D points, corresponding on pixel \( I_o \) is back-projected to pixel \( p_j \) in reference
view following homography warping function (Eq. 3). The sparse correspondence among views is enforced to retain the 3D consistency. After training with Eq. 10, the SPN is used to filter the style transferred image generated by Eq. 9:

\[
\hat{D} = \text{SPN} (\text{Dec}(F_c I^a), I^f)
\]

3.6.3 Style Consistency Loss. With the aforementioned modules, the visual style of unlabeled sample \(S^u = \{(I^u_v, K^u_v, T^u_v, t^u_v)^N_{v=1}\}\) is transferred to labeled sample \(S^l = \{(I^l_v, K^l_v, T^l_v, t^l_v)^N_{v=1}\}, D^l\), and the generated sample is noted as \(S^g = \{(\hat{I}^g_v, \hat{K}^g_v, \hat{T}^g_v, \hat{t}^g_v)^N_{v=1}, \hat{D}^l\}\). The camera parameters and ground depth value of \(S^g\) are shared with the original labeled sample \(S^l\). Following Eq. 7, Eq. 9 and Eq. 11, the generated image \(\hat{I}^g_v\) on each view \(v\) is calculated by utilizing unlabeled image \(I^u_v\) as style image and labeled image \(I^l_v\) as content image. Then the style augmented samples are fed to the backbone network and return the predicted depth map \(D^l_{\text{pred}}\). The style consistency loss requires the output depth map \(D^l_{\text{pred}}\) of the style transferred samples \(S^g\) to be the same as the ground truth \(D^l\):

\[
L_{\text{style}} = \sum_{i=1}^{H \times W} \frac{1}{\sum_{i=1}^{H \times W} 1(D^g_{gt}(p_i) > 0)} ||D^l_{\text{pred}}(p_i) - D^l_{gt}(p_i)||_2^2
\]

3.7 Overall Loss

As shown in Fig. 2, the overall loss is the sum of all aforementioned terms:

\[
L_{\text{overall}} = L_{\text{sup}} + L_{\text{photo}} + \lambda_1 * L_{\text{consis}} + \lambda_2 * L_{\text{style}}
\]

where \(L_{\text{sup}}\) (Eq. 2) is the basic supervision loss on labeled sample \(S^l\). On unlabeled sample \(S^u\), \(L_{\text{photo}}\) (Eq. 5) is the basic photometric consistency loss in unsupervised MVS, and \(L_{\text{consis}}\) (Eq. 6) is the consistency regularization loss. \(L_{\text{style}}\) (Eq. 12) is the style consistency calculated on style augmented sample \(S^g\). In default, \(\lambda_1\) is set to 0.1, and \(\lambda_2\) is set to 1.0.

4 EXPERIMENT

4.1 Dataset and Evaluation Metric

DTU (DTU) [14], Tanks&Temples (T&T) [19], BlendedMVS (BLD) [49] and GTA-SFM (GTA) [37] are utilized for evaluation in this paper. For quantitative evaluation on DTU benchmark, we calculate the accuracy and the completeness predefined by official protocol. The overall score takes the average of mean accuracy and mean completeness as the reconstruction quality. For quantitative evaluation on T&T benchmark, we evaluate the intermediate and advanced set according to their online benchmark. For the BLD and GTA datasets, we implement the quantitative evaluation ourselves which also supports GPU parallel computation. The ground truth point clouds are sampled from the provided meshes, and the evaluation protocol follows the metric of precision, recall, and f-score defined in T&T.

4.2 Implementation Detail

Training and testing: In default, we utilize CasMVSNet [39] as the backbone network. The split of training, valid and test sets in each dataset follows the official configuration in DTU, BLD and GTA. On each dataset, the model is trained on the training set first and tested on the test set for ablation study and evaluation. 8 NVIDIA V100 GPUs are used for training. The batch size on each GPU is set to 1 and the training procedure requires 16 epochs. The number of views is set to 3 following the default setting of CasMVSNet [39]. The hyper-parameter setting of backbone networks follows the official CasMVSNet as well. In default, Whitening and Coloring Transform (WCT [21]) is selected as the backbone of STM module. The backbone of GPM is a spatial propagation network [22], a generic framework that can be applied to many affinity-related tasks, including but not limited to image matting, segmentation and etc. STM and GPM are further trained on each dataset following Eqs. 8 and 10 respectively. For evaluation, the model is tested on the evaluation or test set of DTU, T&T, BLD, and GTA.

Semi-supervised settings: 2 different semi-supervised settings are adopted in experiments: (1) labeled and unlabeled data separated by scenes; (2) labeled and unlabeled data separated by views. From the visualization results in Figure 1(a), it is verified that apparent gap of distribution exists among different scenes. The former setting based on scenes aims at exploring the performance of our proposed method under this challenging case. For the latter one, we aim to simulate the common case that labeled and unlabeled data are separated by the multi-view image pairs on different views.

4.3 Ablation Study

4.3.1 Loss terms of SDA-MVS. To validate the effectiveness of each loss term in the proposed method, we conduct ablation study about the loss terms in this section. Experiments on DTU, BLD and GTA are conducted on the model trained with the training set of each dataset and tested on the corresponding test set. The backbone model is CasMVSNet [39]. The quantitative ablation results are shown in Tab. 1. Different loss terms are used respectively for experiments, including \(L_{\text{photo}}, L_{\text{consis}}, L_{\text{style}}, \text{ and } L_{\text{sup}}\). In the table, the first column means the utilized setting of 3D annotations. 0% and 100% respectively represent the unsupervised baseline and the fully-supervised baseline. Following the 2 semi-supervised settings defined in Sec. 4.2, we randomly select 10% samples with 3D ground truth from the whole dataset and the remaining 90% are unlabeled samples. 10%-S means 10% labeled samples are selected based on scenes, 10%-V means 10% labeled sample are selected based on views. From the table, we can find that our semi-supervised method achieves competitive performance on the Overall score of DTU, the F-score of BLD and GTA, compared with the supervised baselines. As shown, our method under the setting of 10%-V achieves Overall metric of 0.3337, which is better than the score of 0.3673 in unsupervised baseline and the score of 0.355 in fully-supervised baseline. Similar results are also provided under the setting of 10%-S, where our method shows competitive performance with the Overall metric of 0.3504 compared to the supervised baseline and unsupervised baseline. Furthermore, qualitative results of the ablation experiments under the settings of 10%-V and 10%-S are respectively shown in Fig. 3 and Fig. 4.

4.3.2 Percentage of Labels. The percentage of labeled samples in the semi-supervised setting is an important factor to demonstrate the effectiveness of the proposed method. To explore the effectiveness of our proposed method under different percentage of labeled samples, we conduct experiments under the labeled percentage...
Table 1: Ablation study of the proposed method on DTU, BlendedMVS and GTASFM datasets. ↑ means the higher the better, and ↓ means the lower the better.

| Dataset  | Acc. | Comp. | Overall | DTU Dataset | BLD Dataset | GTA Dataset | GT/A Dataset |
|---------|------|-------|---------|-------------|-------------|-------------|--------------|
| 0%      | ✓    | ✓     | ✓       | L_photo     | L_consis    | L_style     | L_sup        |
|         | 0.3748 | 0.3598 | 0.3673  | 0.3528 | 0.1575 | 0.1796 | 0.4055 | 0.4441 | 0.4170 |
| 10%-S   | ✓    | ✓     | ✓       | L_photo + L_consis | 0.3572 | 0.3603 | 0.3588 | 0.4104 | 0.4135 | 0.3632 | 0.4264 | 0.4392 | 0.4255 |
|         | 0.3447 | 0.3560 | 0.3504  | 0.4217 | 0.4328 | 0.3806 | 0.4523 | 0.5286 | 0.4807 |
| 10%-V   | ✓    | ✓     | ✓       | L_photo + L_consis + L_style | 0.3498 | 0.3480 | 0.3489 | 0.4269 | 0.4475 | 0.3857 | 0.4870 | 0.6089 | 0.5351 |
|         | 0.3305 | 0.3369 | 0.3337  | 0.4446 | 0.4614 | 0.4033 | 0.5354 | 0.6261 | 0.5711 |
| 100%    | ✓    | ✓     | ✓       | L_sup     | 0.325  | 0.383  | 0.355   | 0.3644 | 0.4491 | 0.3760 | 0.5239 | 0.4794 | 0.4796 |

Figure 3: Qualitative ablation study of our proposed method under the setting of 10%-V.

Figure 4: Qualitative ablation study of our proposed method under the setting of 10%-S.

ranging from 5% to 100%. As shown in Fig. 5, we report the per-point errors of reconstructed dense 3D point cloud under the percentage of 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100%. Each percentage setting is randomly selected from all available samples for a fair comparison. However, the random selection process may make the performance fluctuate to some extent, which is further discussed detaiilly in Sec. 4.3.3. Here, we can only summarize the tendency of increasing the percentage of labeled samples from the experiment results. Note that the 100% setting here is the combination of fully-supervised training and our proposed framework. Compared with the fully-supervised baseline whose Overall score is about 0.355, each percentage setting of our proposed framework in Fig. 5 and Fig. 5 achieves superior performance which is smaller than 0.355. It demonstrates the stable improvement of our proposed method compared to the basic fully-supervised baseline as well as the unsupervised one.
4.3.3 Random Sampling of Labeled and Unlabeled data. Due to the random sampling process of labeled and unlabeled data, the unexpected factors in random separation might be a confounder to the exact performance of our semi-supervised MVS method. To prove the exact performance under different random selected samples, we conduct random sampling repeatedly for 5 times and provide the benchmarking results on DTU in Tab. 2. As the results show, the Overall metric of 10%-V fluctuates around 0.33 and the Overall metric of 10%-S fluctuates around 0.352. These results prove that our semi-supervised method can achieve competitive and even better performance compared with unsupervised and fully-supervised baseline.

4.4 Benchmark Result

4.4.1 Results on DTU. We compare the proposed semi-supervised framework with other state-of-the-art methods quantitatively in Table 3 The 1-st column of the table means the annotation strategy of the method: 1) Trad. means traditional MVS methods without annotation; 2) 100% means fully-supervised MVS methods; 3) 0% means unsupervised MVS methods; 4) Sparse means sparse point annotations according to [18]; 5) 10%-S and 10%-V means the aforementioned semi-supervised settings (10% scenes / views) in Sec. 4.2. In the table, we compare our method with state-of-the-art traditional MVS methods (i.e. Furu [9], Gipuma [19], and Colmap [31]), learning-based MVS methods (i.e. MVSNet [47], R-MVSNet [48], Point-MVSNet [5], CasMVSNet [39], UCS-Net [6], and UGNet [33]), and unsupervised MVS methods (i.e. Unsup_MVS [17], M3VSNet [13], JDACS [41], and U-MVS [42]). Furthermore, we also compare the recent semi-supervised MVS methods, SGT-MVS [18]. As shown in the table, with limited 10% dense ground truth in the training set, our proposed method performs competitively compared with supervised MVS methods, achieving an overall score of 0.334. Furthermore, compared with the reported official supervised performance of the backbone network, CasMVSNet [39], our proposed method achieve better performance with much less dense 3D annotations. In addition, the proposed SDA-MVS outperforms previous state-of-the-art traditional MVS methods and unsupervised MVS methods are presented in the table.

4.4.2 Results on Tanks&Temples. Leveraging the model trained by the training set of DTU without finetuning, we test the performance of our proposed method on T&T. In Tab. 4, the results on the Intermediate and Advanced set of T&T are reported. The reported F-score on both the intermediate and advanced partitions are used in the table. We compare our proposed method with traditional methods (i.e. PMVS [9], SMVS [20], MVE [8], COLMAP [31]), fully-supervised MVS methods (i.e. MVSNet [47], R-MVSNet [48], CasMVSNet [39], UCS-Net [6], PatchmatchNet [36], MSTR [53]), and self-supervised MVS methods (i.e. MVS2 [7], M3VSNet [13], JDACS [41], U-MVS [42]). Note that our SDA-MVS framework only...
Table 4: Quantitative results of different methods on Tanks and Temples benchmark (higher is better). CasMSVNet represents the fully-supervised baseline.

| Method                  | F-Score | T&R Intermediate | Train Mean | Audt. Mean |
|-------------------------|---------|------------------|------------|------------|
| OpenMVG + PAVS          | 41.05   | 37.76            | 12.33      | 36.68      |
| OpenMVG + SMVS          | 31.91   | 19.92            | 15.02      | 39.38      |
| OpenMVG + MVE           | 49.91   | 28.19            | 20.75      | 43.15      |
| CORDAP [31]             | 42.14   | 30.41            | 22.25      | 25.63      |
| CasMSVNet [39]          | 55.99   | 28.53            | 20.70      | 39.79      |
| R-CASMSVNet [48]        | 69.60   | 46.65            | 32.59      | 42.95      |
| CIDER [43]              | 56.79   | 32.39            | 29.89      | 54.67      |
| Fast-MVSNet [51]        | 65.18   | 39.59            | 34.98      | 47.81      |
| CasMVSNet [70]          | 76.37   | 58.45            | 46.26      | 51.81      |
| UCS-Net [6]             | 76.09   | 53.16            | 43.03      | 54        |
| CVF-MVSNet [46]         | 76.57   | 47.74            | 36.34      | 55.12      |
| PVANet [50]             | 69.36   | 46.8             | 46.01      | 57.54      |
| PatchmatchNet [36]      | 68.99   | 52.64            | 43.24      | 54.87      |
| MVSTR [33]              | 76.92   | 59.82            | 50.14      | 56.73      |
| MVS [7]                 | 47.47   | 21.15            | 19.50      | 44.54      |
| MVSNet [13]             | 47.74   | 24.38            | 18.74      | 44.42      |
| JDACS [41]              | 66.62   | 38.25            | 36.11      | 46.12      |
| U-MVS [42]              | 76.49   | 60.04            | 49.20      | 55.52      |
| SDA-MVS (10% Scenes)    | 59.36   | 38.88            | 32.87      | 54.63      |
| SDA-MVS (10% Views)     | 77.90   | 55.55            | 52.59      | 56.56      |

Table 5: Experimental results under the setting of unsupervised domain adaptation.

| Methods                | BLD → DTU | GTA → DTU |
|------------------------|------------|------------|
| Source Only            | Acc. | Comp. | Overall | Acc. | Comp. | Overall |

| CasMVSNet [39]         | 0.3609 | 0.4024 | 0.3818 | 0.3681 | 0.3488 | 0.3556 |
| H-model [28]           | 0.3630 | 0.3815 | 0.3726 | 0.3765 | 0.3499 | 0.4629 |
| Zhang et al. [52]      | 0.3619 | 0.3634 | 0.3627 | 0.3746 | 0.405  | 0.3908 |
| SDA-MVS                | 0.3681 | 0.3488 | 0.3556 | 0.3515 | 0.3901 | 0.3658 |

4.4.3 Unsupervised Domain Adaptation on MVS. Recent work [52] reveals that semi-supervised learning (SSL) can be simply extended to unsupervised domain adaptation (UDA) and achieve great performance. We further evaluate the proposed SDA-MVS by extending it to the unsupervised domain adaptation task and present the experimental results in Tab. 5. The CasMSVNet [39] is selected as the backbone network. We select BLD/GTA as the source domain, and DTU as the target domain, denoted as BLD/GTA → DTU. As shown in the table, our method can also achieve great performance when generalizing to the unsupervised domain adaptation MVS task. It reveals the extensive potential of our proposed SDA-MVS from SSL to UDA.

5 CONCLUSION

In this paper, we explore the semi-supervised MVS problem that assumes only part of the MVS dataset has dense depth annotations. It has an intractable risk of breaking the basic assumption in classic semi-supervised learning techniques, that labeled data and unlabeled data share same label space and data distribution, denoted as semi-supervised distribution-gap ambiguity. To handle this issue, we propose a novel semi-supervised distribution-augmented MVS framework, called SDA-MVS. For the case that the assumption works in the MVS data, consistency regularization based on the KL divergence between the predicted probability volumes with and without random data augmentation is enforced to train the model. For the case that the assumption fails in the MVS data because of distribution mismatch, style consistency regularization enforces the invariance between the style augmented sample and original labeled sample. The style augmented sample is generated by transferring visual styles from unlabeled data to labeled data, inherently shrinking the distribution gap. Experimental results show that our proposed SDA-MVS can handle the semi-supervised MVS effectively and be extended to an UDA MVS with great performance.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No.61976095) and the Natural Science Foundation of Guangdong Province, China (No.2022A1515010114). This work was also supported by Alibaba Group through Alibaba Research Intern Program.

REFERENCES

[1] Philip Bachman, Oaas Alsharif, and Doina Precup. 2014. Learning with pseudo-ensembles. Advances in neural information processing systems 27 (2014).
[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. Advances in Neural Information Processing Systems 32 (2019).
[3] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. 2017. Semi-supervised deep multi-view stereo with multi-view symmetry. In European Conference on Computer Vision. Springer, 766–779.
[4] Po-Heng Chen, Hsiao-Chen Yang, Kuan-Wen Chen, and Yong-Sheng Chen. 2020. Mvsnet++: Learning depth-based attention pyramid features for multi-view stereo. IEEE Transactions on Image Processing 29 (2020), 7261–7273.
[5] Rui Chen, Songfang Han, Jing Xu, and Hao Su. 2019. Point-based multi-view stereo network. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 1538–1547.
[6] Shao Cheng, Zexiang Xu, Shulin Zhu, Zhuan Li, LiErran Li, RaviRamamoorthi, and Hao Su. 2020. Deep stereo using adaptive thin volume representation with uncertainty awareness. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2524–2534.
[7] Yuchao Dai, Zhidong Zha, Zhubo Rao, and Bo Li. 2019. Mvs2: Deep unsupervised multi-view stereo with multi-view symmetry. In 2019 International Conference on 3D Vision (3DV). IEEE, 1–8.
Yves Grandvalet and Yoshua Bengio. 2004. Semi-supervised learning by entropy minimization. In Advances in neural information processing systems 17 (2004).

Baichuan Huang, Hongwei Yi, Can Huang, Yijie He, Jinglin Liu, and Xiao Liu. 2021. M³VSNNet: Unsupervised multi-metric multi-view stereo network. In Proceedings of the IEEE International Conference on Image Processing (ICIP). IEEE, 3163–3167.

Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. 2014. Large scale multi-view stereo evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 406–413.

Arno Knipitsch, Jasiq Park, Qian-Yi Zhou, and Vladlen Koltun. 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (ToG) 36, 4 (2017), 1–13.

Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. 2021. Patchmatchnet: Learned multi-view patchmatch stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 14194–14203.

Kazuhiro Wang and Shaojie Shen. 2020. Flow-motion and depth network for monocular stereo and beyond. IEEE Robotics and Automation Letters 5, 2 (2020), 3307–3314.

Zizhuang Wei, Qingshan Xu, and Wenbing Tao. 2020. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 12937–12947.

Chen Min, Yisong Chen, and Guoping Wang. 2020. Dense hybrid recurrent multi-view stereo network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 3307–3314.