KD-MVS: Knowledge Distillation Based Self-supervised Learning for Multi-view Stereo

Yikang Ding[1,2] Qingtian Zhu[1] Xiangyue Liu[1] Wentao Yuan[1]
Haotian Zhang[1⋆] Chi Zhang[1]

[1]Megvii Research [2]Tsinghua University

Abstract. Supervised multi-view stereo (MVS) methods have achieved remarkable progress in terms of reconstruction quality, but suffer from the challenge of collecting large-scale ground-truth depth. In this paper, we propose a novel self-supervised training pipeline for MVS based on knowledge distillation, termed *KD-MVS*, which mainly consists of self-supervised teacher training and distillation-based student training. Specifically, the teacher model is trained in a self-supervised fashion using both photometric and featuremetric consistency. Then we distill the knowledge of the teacher model to the student model through probabilistic knowledge transferring. With the supervision of validated knowledge, the student model is able to outperform its teacher by a large margin. Extensive experiments performed on multiple datasets show our method can even outperform supervised methods. Code is available at https://github.com/megvii-research/KD-MVS.

1 Introduction

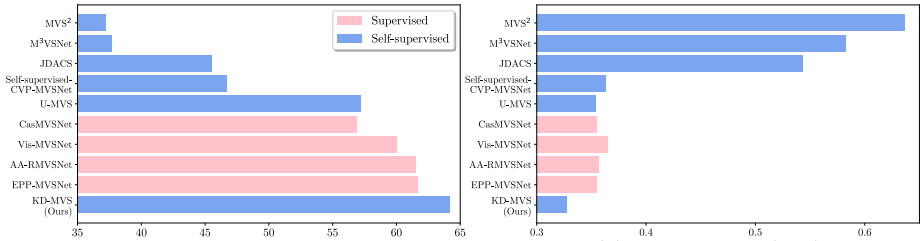
The task of multi-view stereo (MVS) is to reconstruct a dense 3D presentation of the observed scene using a series of calibrated images, which plays an important role in a variety of applications, e.g. augmented and virtual reality, robotics and computer graphics. Recently, learning-based MVS networks [44,45,11,7,6,21] have obtained impressive results. However, supervised methods require dense depth annotations as explicit supervision, the acquisition of which is still an expensive challenge. Subsequent attempts [18,39,38,42,14] have made efforts to train MVS networks in a self-supervised manner by using photometric consistency [18,4], optical flow [39] or reconstructed 3D models [14,42].

Though great improvement has been made, there is a significant gap in either reconstruction completeness or accuracy compared to supervised methods.

In this paper, we propose a novel self-supervised training pipeline for MVS based on knowledge distillation [13], named *KD-MVS*. The pipeline of KD-MVS mainly consists of (a) self-supervised teacher training and (b) distillation-based student training. In the self-supervised teacher training stage, the teacher model is trained by enforcing both the photometric consistency [18] and featuremetric consistency between the reference view and the reconstructed views, which

This work is done by the first four authors as interns at Megvii Research.

⋆ Corresponding author (zhanghaotian@megvii.com).



(a) Intermediate F-score (b) Overall error (mm)

Fig. 1: Visualized performance comparisons of state-of-the-art MVS methods on (a) Tanks and Temples benchmark [19] and (b) DTU dataset [2].

can be obtained via homography warping according to the estimated depth. Unlike the existing self-supervised MVS methods [18,42,4] that use only photometric consistency, we propose to use the internally extracted features to utilize the featuremetric consistency, which is different from the externally extracted features-based loss, e.g. perceptual loss [16]. We analyze and show that the proposed internal featuremetric loss is more suitable for MVS and is able to help the self-supervised teacher model yield relatively complete and accurate depth maps.

The distillation-based student training stage consists of two main steps: the pseudo probabilistic knowledge generation and the student training. We first use the teacher model to infer raw depth maps on unlabeled training data and perform cross-view check to filter unreliable samples. We then generate the pseudo probability distribution of the teacher model by probabilistic encoding. The probabilistic knowledge can be transferred to the student model by forcing the predicted probability distribution of the student model to be similar to the pseudo probability distribution. As a result, the student model can surpass its teacher and even outperform supervised methods. Extensive experiments on DTU dataset [2], Tanks and Temples benchmark [1] and BlendedMVS dataset [46] show that KD-MVS brings significant improvement to off-the-shelf MVS networks, even outperforming supervised methods, as is shown in Fig. 1. It is worth noting that applying with CasMVSNet [11], KD-MVS ranks 1^{st} among all submitted methods on Tanks and Temples benchmark [1].

Our main contributions are four-fold as follows:

- We propose a novel self-supervised training pipeline named KD-MVS based on knowledge distillation.
- We design an internal featuremetric consistency loss to perform robust self-supervised training of the teacher model.
- We propose to perform knowledge distillation to transfer validated knowledge from the self-supervised teacher to a student model for boosting performance.
- Our method ranks 1^{st} among all submitted methods (including supervised methods) on Tanks and Temples benchmark [1] and also achieves state-of-the-art performance on DTU [2] dataset and BlendedMVS [46] dataset.

2 Related Work

2.1 Learning-based MVS

Supervised MVS Learning-based methods for MVS have achieved impressive reconstruction quality. MVSNet [44] transforms the MVS task to a per-view depth estimation task and encodes camera parameters via differentiable homography to build 3D cost volumes, which will be regularized by a 3D CNN to obtain a probability volume for pixel-wise depth distribution. However, at cost volume regularization, 3D tensors occupy massive memory for processing. To alleviate this problem, some attempts [45,41,36] replace the 3D CNN by 2D CNNs and a RNN and some other methods [11,48,3,43] use a multi-stage approach and predict depth in a coarse-to-fine manner.

Self-supervised MVS The key of self-supervised MVS methods is how to make use of prior multi-view information and transform the problem of depth prediction into other forms of problems. Unsup-MVS [18] firstly handles MVS as an image reconstruction problem by warping pixels to neighboring views with estimated depth values. Given multiple images, MVS^2 [4] predicts each view's depth simultaneously and trains the model using cross-view consistency. M^{3}VSNet [14] makes use of the consistency between the surface normal and depth map to enhance the training pipeline and JDACS [38] proposes a unified framework to improve the robustness of self-supervisory signals against natural color disturbance in multi-view images. U-MVS [39] utilizes the pseudo optical flow generated by off-the-shelf methods to improve the self-supervised model's performance. [42] renders pseudo depth labels from reconstructed mesh models and continues to train the self-supervised model.

2.2 Knowledge Distillation

Knowledge distillation [13] aims to transfer knowledge from a teacher model to a student model, so that a powerful and lightweight student model can be obtained. [25,35,29,34,26] consider knowledge at feature space and transfer it to the student model's feature space. Born-Again Networks (BAN) [8] trains a student model similarly parameterized as the teacher model and makes the trained student be a teacher model in a new round. The self-training scheme [37] generates distillation labels for unlabeled data and trains the student model with these labels. Probabilistic knowledge transfer (PKT) [28,27] trains the student model via matching the probability distribution of the teacher model. Since labeled data are not required to minimize the difference of probability distribution, PKT can also be applied to unsupervised learning. In this work, we are inspired by PKT and offline distillation [30,47,15,24,20] and propose to transfer the response-based knowledge [10] by forcing the predicted probability distribution of the student model to be similar to the probability distribution of the teacher model in an offline manner.

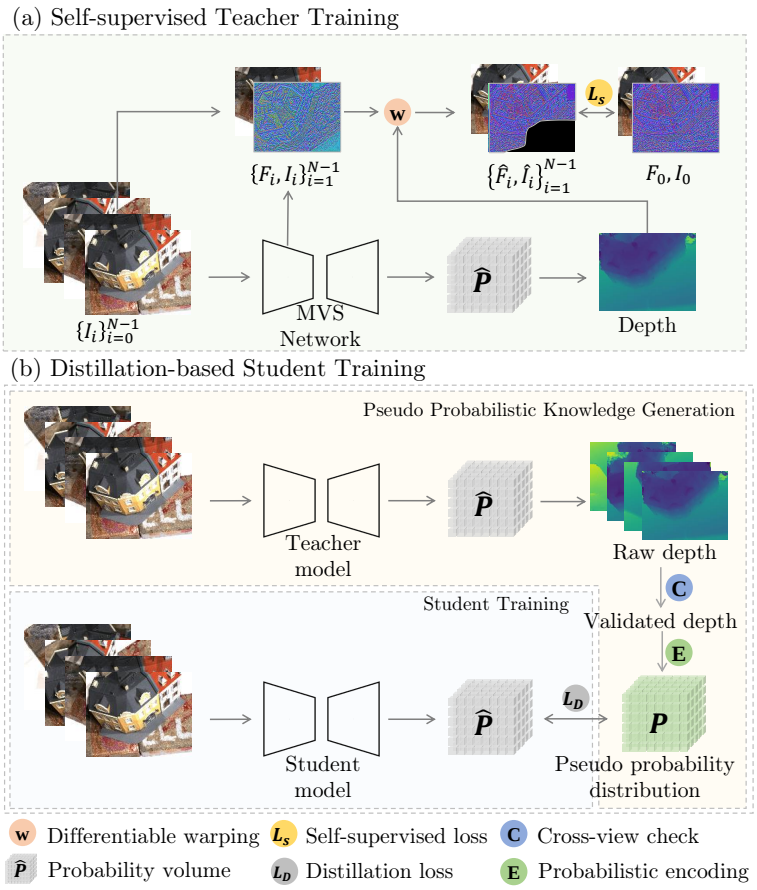


Fig. 2: Overview of KD-MVS. The first stage is self-supervised teacher training. The second stage is distillation-based student training, including pseudo probabilistic knowledge generation and student training.

3 Methodology

In this section, we elaborate the proposed training framework as illustrated in Fig. 2. KD-MVS mainly consists of self-supervised teacher training (Sec. 3.1) and distillation-based student training (Sec. 3.2). Specifically, we first train a teacher model in a self-supervised manner by using both the photometric and featuremetric consistency between the reference view and the reconstructed views. We then generate the pseudo probability distribution of the teacher model via cross-view check and probabilistic encoding. With the supervision of the pseudo probability, the student model is trained with distillation loss in an offline distillation manner. It is worth noting that the proposed KD-MVS is a general pipeline

for training MVS networks, it can be easily adapted to arbitrary learning-based MVS networks. In this paper, we mainly study KD-MVS with CasMVSNet [11].

3.1 Self-supervised Teacher Training

In addition to conventional photometric consistency [18] used in self-supervised MVS, we propose to use internal features and featuremetric consistency as an additional supervisory signal. Both the photometric and featuremetric consistency are obtained by calculating the distance between the reference view and the reconstructed views. The following is the introduction to view reconstruction and loss formulation.

View Reconstruction Given a reference image $\mathbf{I}_0$ and its neighboring source images $\{\mathbf{I}_i\}_{i=1}^{N-1}$, the common coarse-to-fine MVS network (e.g. CasMVSNet [11]) extracts features for all N images at three different resolution levels (1/4, 1/2, 1), denoted as $\{\mathbf{F}_i^{1/4}, \mathbf{F}_i^{1/2}, \mathbf{F}_i\}_{i=0}^{N-1}$, and estimates the depth maps at these three corresponding levels, as $\mathbf{D}_0^{1/4}$, $\mathbf{D}_0^{1/2}$ and $\mathbf{D}_0$.

Taking $\mathbf{F}_0$ and $\mathbf{D}_0$ as an example, the warping between a pixel $\mathbf{p}$ at the reference view and its corresponding pixel $\hat{\mathbf{p}}_i$ at the i-th source view under estimated depth $d = \mathbf{D}_0(\mathbf{p})$ is defined as:

$$\hat{\mathbf{p}}_i = \mathbf{K}_i[\mathbf{R_i}(\mathbf{K}_0^{-1}\mathbf{p}d) + \mathbf{t_i}], \tag{1}$$

where $\mathbf{R}_i$ and $\mathbf{t}_i$ denote the relative rotation and translation from the reference view to the i-th source view. $\mathbf{K}_0$ and $\mathbf{K}_i$ are the intrinsic matrices of the reference and the i-th source camera. According to Eq. (1), we are able to get the reconstructed images $\hat{\mathbf{I}}_i$ and features $\hat{\mathbf{I}}_i$ corresponding to the i-th source view. Fig. 4 shows a photometric warping process from the i-th source view to the reference view.

Loss Formulation Our self-supervised training loss consists of two components: photometric loss $\mathcal{L}_{\text{photo}}$ and featuremetric loss $\mathcal{L}_{\text{fea}}$. Following [18], the $\mathcal{L}_{\text{photo}}$ is based on the ℓ-1 distance between the raw RGB reference image and the reconstructed images. However, we find that the photometric loss is sensitive to lighting conditions and shooting angles, resulting in poor completeness of predictions. To overcome this problem, we use the featuremetric loss to construct a more robust loss function. Given the extracted features $\{\mathbf{F}_i\}_{i=0}^{N-1}$ from the feature net of MVS network, and the reconstructed feature maps $\hat{\mathbf{F}}_i$ generated from the i-th view, our featuremetric loss between $\hat{\mathbf{F}}_i$ and $\mathbf{F}_0$ is obtained by:

$$\mathcal{L}_{\text{fea}}^{(i)} = \|\hat{\mathbf{F}}_i - \mathbf{F}_0\|. \tag{2}$$

It is worth noting that we put forward to use the internal features extracted by the internal feature net of the online training MVS network instead of the external features (e.g. extracted by a pre-trained backbone network [16]) to

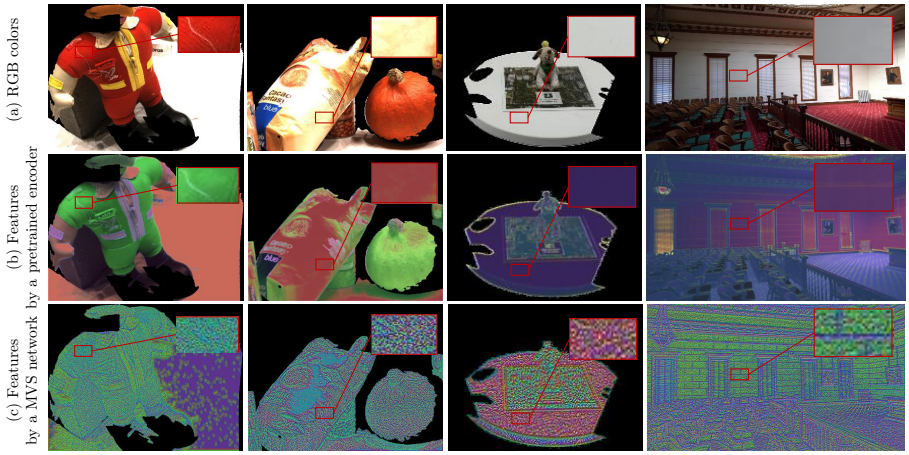


Fig. 3: Visualized examples of RGB colors (photometric) and extracted features (featuremetric). Dimension reduction of features is done by PCA. Features in (b) are extracted by a pre-trained ResNet-18 [12] and those in (c) are the features of the online training MVS network. Pretrained backbones tend to neglect the pixel-wise differences within intra-class regions while the online training MVS network is able to extract locally distinguishable features, which are more beneficial to downstream feature matching.

compute featuremetric loss. Our insight is that the nature of MVS is multi-view feature matching along epipolar lines, so the features are supposed to be locally discriminative. The pre-trained backbone networks, e.g. ResNet [12] and VGG-Net [32], are usually trained with image classification loss, so that their features are not locally discriminative. As shown in Fig. 3, we compare the features extracted by an external pre-trained backbone (ResNet [12]) and by the internal encoder of the MVS network during online self-supervised training. These two options lead to completely different feature representation and we study it in Sec. 4.4 with experiments.

To summarize, the final loss function for self-supervised teacher training is

$$\mathcal{L}_S = \frac{1}{|\mathbf{V}|} \sum_{\mathbf{p} \in \mathbf{V}} \sum_{i=1}^{N-1} (\lambda_{\text{fea}} \mathcal{L}_{\text{fea}}^{(i)} + \lambda_{\text{photo}} \mathcal{L}_{\text{photo}}^{(i)}), \tag{3}$$

where $\mathbf{V}$ is the valid subset of image pixels. λ_{fea} and λ_{photo} are the two manually tuned weights, and in our experiments, we set them as 4 and 1 respectively. For coarse-to-fine networks, e.g. CasMVSNet, the loss function is applied to each of the regularization steps.

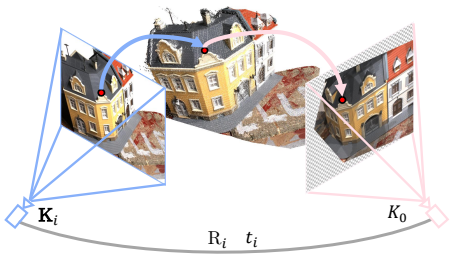


Fig. 4: Photometric warping.

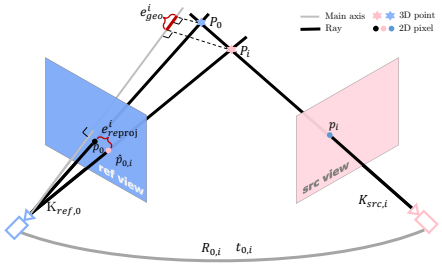


Fig. 5: Cross-view check.

3.2 Distillation-based Student Training

To further stimulate the potential of the self-supervised MVS network, we adopt the idea of knowledge distillation and transfer the probabilistic knowledge of the teacher to a student model. This process mainly consists of two steps, namely pseudo probabilistic knowledge generation and student training.

Pseudo Probabilistic Knowledge Generation We consider the knowledge transfer is done through the probability distribution as is done in [15,30,35]. However, we face two main problems when applying distillation in MVS. (a) The raw per-view depth generated from the teacher model contains a lot of outliers, which is harmful to training student model. Thus we perform cross-view check to filter outliers. (b) The real probabilistic knowledge of the teacher model cannot be used directly to train the student model. That is because the depth hypotheses in the coarse-to-fine MVS network need to be dynamically sampled according to the results of the previous stage, and we cannot guarantee that the teacher model and student model always share the same depth hypotheses. To solve this problem, we propose to generate the pseudo probability distribution by probabilistic encoding.

Cross-view Check is used to filter outliers in the raw depth maps, which are inferred by the self-supervised teacher model on the unlabeled training data. Naturally, the outputs of the teacher model are per-view depth maps and the corresponding confidence maps. For coarse-to-fine methods, e.g. CasMVSNet [11], we multiply confidence maps of all three stages to obtain the final confidence map and take the depth map in the finest resolution as the final depth prediction.

We denote the final confidence map of reference view as $\mathbf{C}_0$ and the final depth prediction as $\mathbf{D}_0$, the depth maps of source views as $\{\mathbf{D}_i\}_{i=1}^{N-1}$. As is illustrated in Fig. 5, considering an arbitrary pixel $\mathbf{p}_0$ in the reference image coordinate, we cast the 2D point $\mathbf{p}_0$ to a 3D point $\mathbf{P}_0$ with the depth value $\mathbf{D}_0(\mathbf{p}_0)$. We then back-project $\mathbf{P}_0$ to i-th source view and obtain the point $\mathbf{p}_i$ in the source view. Using its estimated depth $\mathbf{D}_i(\mathbf{p}_i)$, we can cast the $\mathbf{p}_i$ to the 3D point $\mathbf{P}_i$. Finally, we back project $\mathbf{P}_i$ to the reference view and get $\hat{\mathbf{p}}_{0,i}$. Then the reprojection error at $\mathbf{p}_0$ can be written as $e^{\mathrm{i}}_{\mathrm{reproj}} = \|\mathbf{p}_0 - \hat{\mathbf{p}}_{0,i}\|$. A geometric error

$e^{\mathrm{i}}_{\mathrm{geo}}$ is also defined to measure the relative depth error of $\mathbf{P}_0$ and $\mathbf{P}_i$ observed from the reference camera as $e^{\mathrm{i}}_{\mathrm{geo}} = \|\tilde{D}_0(\mathbf{P}_0) - \tilde{D}_0(\mathbf{P}_i)\|/\tilde{D}_0(\mathbf{P}_0)$, where the $\tilde{D}_0(\mathbf{P}_0)$ and $\tilde{D}_0(\mathbf{P}_i)$ are the projected depth of $\mathbf{P}_0$ and $\mathbf{P}_i$ in the reference view. Accordingly, the validated subset of pixels with regard to the i-th source view is defined as

$$\{\mathbf{p}_0\}_i = \{\mathbf{p}_0 | \mathbf{C}_0(\mathbf{p}_0) > \tau_{\mathrm{conf}}, e^{\mathrm{i}}_{\mathrm{reproj}} < \tau_{\mathrm{reproj}}, e^{\mathrm{i}}_{\mathrm{geo}} < \tau_{\mathrm{geo}}\}, \tag{4}$$

where τ represents threshold values, we set τ_{conf}, τ_{reproj} and τ_{geo} to 0.15, 1.0 and 0.01 respectively. The final validated mask is the intersection of all $\{\mathbf{p}_0\}_i$ across N-1 source views. The obtained $\{\tilde{D}_0(\mathbf{P}_i)\}_{i=0}^{N-1}$ and validated mask will be further used to generate the pseudo probability distribution.

Probabilistic Encoding uses the $\{\tilde{D}_0(\mathbf{P}_i)\}_{i=0}^{N-1}$ to generate the pseudo probability distribution $P_{\mathbf{p}_0}(d)$ of depth value d for each validated pixel $\mathbf{p}_0$ in reference view. We model $P_{\mathbf{p}_0}$ as a Gaussian distribution with a mean depth value of $\mu(\mathbf{p}_0)$ and a variance of $\sigma^2(\mathbf{p}_0)$, which can be obtained by maximum likelihood estimation (MLE):

$$\mu(\mathbf{p}_0) = \frac{1}{N}\sum_{i=0}^{N-1}\tilde{D}_0(\mathbf{P}_i), \quad \sigma^2(\mathbf{p}_0) = \frac{1}{N}\sum_{i=0}^{N-1}\Big(\tilde{D}_0(\mathbf{P}_i) - \mu(\mathbf{p}_0)\Big)^2. \tag{5}$$

The $\mu(\mathbf{p}_0)$ fuses the depth information from multiple views, while the $\sigma^2(\mathbf{p}_0)$ reflects the uncertainty of the teacher model at $\mathbf{p}_0$, which will provide probabilistic knowledge for the student model during distillation training.

Student Training With the pseudo probability distribution P, we are able to train a student model from scratch via forcing its predicted probability distribution $\hat{P}$ to be similar with P. For the discrete depth hypotheses $\{d_k\}_{k=0}^{D}$, we obtain their pseudo probability $\{P(d_k)\}_{k=0}^{D}$ on the continuous probability distribution P and normalize $\{P(d_k)\}_{k=0}^{D}$ using SoftMax, taking the result as the final discrete pseudo probability value. We use Kullback–Leibler divergence to measure the distance between the student model's predicted probability and the pseudo probability. The distillation loss $\mathcal{L}_D$ is defined as

$$\mathcal{L}_D = \mathcal{L}_{KL}(P||\hat{P}) = \sum_{\mathbf{p}\in\{\mathbf{p}_v\}}\Big(P_{\mathbf{p}} - \hat{P}_{\mathbf{p}}\Big)\, log\left(\frac{P_{\mathbf{p}}}{\hat{P}_{\mathbf{p}}}\right), \tag{6}$$

where $\{\mathbf{p}_v\}$ represents the subset of valid pixels after cross-view check.

In experiments, we find that the trained student model also has the potential of becoming a teacher and further distilling its knowledge to another student model. As a trade-off between training time and performance, we perform the process of knowledge distillation once more. More details can be found in Sec. 4.4.

Table 1: Quantitative results on DTU evaluation set [2] (**lower is better**). Sup. indicates whether the method is supervised or not.

Method	Sup.	Acc.	Comp.	Overall
Gipuma [9]	-	0.283	0.873	0.578
COLMAP [31]	-	0.400	0.664	0.532
MVSNet [44]	✓	0.396	0.527	0.462
AA-RMVSNet [36]	✓	0.376	0.339	0.357
CasMVSNet [11]	✓	0.325	0.385	0.355
UCS-Net [3]	✓	0.338	0.349	0.344
Unsup_MVS [18]	✗	0.881	1.073	0.977
MVS^2 [4]	✗	0.760	0.515	0.637
M^3VSNet [14]	✗	0.636	0.531	0.583
JDACS [38]	✗	0.571	0.515	0.543
Self-supervised-CVP-MVSNet [42]	✗	**0.308**	0.418	0.363
U-MVS+MVSNet [39]	✗	0.470	0.430	0.450
U-MVS+CasMVSNet [39]	✗	0.354	0.354	0.354
Ours+MVSNet	✗	0.424	0.426	0.425
Ours+CasMVSNet	✗	0.359	**0.295**	**0.327**

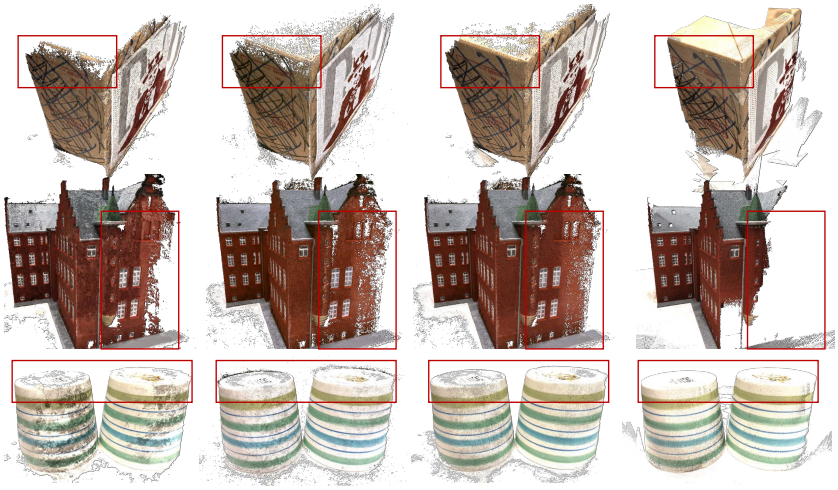
4 Experiments

4.1 Datasets

DTU dataset [2] is captured under well-controlled laboratory conditions with a fixed camera rig, containing 128 scans with 49 views under 7 different lighting conditions. We split the dataset into 79 training scans, 18 validation scans, and 22 evaluation scans by following the practice of MVSNet [44]. BlendedMVS dataset [46] is a large-scale dataset for multi-view stereo and contains objects and scenes of varying complexity and scale. This dataset is split into 106 training scans and 7 validation scans. Tanks and Temples benchmark [19] is a public benchmark acquired in realistic conditions, which contains 8 scenes for the intermediate subset and 6 for the advanced subset.

4.2 Implementation Details

At the phase of self-supervised teacher training on DTU dataset [2], we set the number of input images $N = 5$ and image resolution as 512×640. For coarse-to-fine regularization of CasMVSNet [11], the settings of depth range and the number of depth hypotheses are consistent with [11]; the depth interval decays by 0.25 and 0.5 from the coarsest stage to the finest stage. The teacher model is trained with Adam for 5 epochs with a learning rate of 0.001. At the phase of distillation-based student training, we train the student model with the pseudo probability distribution for 10 epochs. Model training of all experiments is carried out on 8 NVIDIA RTX 2080 GPUs.



(a) CasMVSNet (b) U-MVS (c) **Ours** (d) Ground Truth

Fig. 6: Comparison of reconstructed results with the supervised baseline [11] and the state-of-the-art self-supervised method U-MVS [39] on DTU test set [2].

4.3 Experimental Results

DTU Dataset We evaluate KD-MVS, applied to MVSNet [44] and CasMVS-Net [11] on DTU dataset [2]. We set $N = 5$ and input resolution as 864×1152 at evaluation. Quantitative comparisons are shown in Tab. 1. Accuracy, Completeness and Overall are the three official metrics from [2]. Our method outperforms all self-supervised methods by a large margin and even the supervised ones. Fig. 6 shows a visualization comparison of reconstructed point clouds. Our method achieves much better reconstruction quality when compared with the baseline network and the state-of-the-art self-supervised method.

Tanks and Temples Benchmark We test our method on Tanks and Temples benchmark [19] to demonstrate the ability to generalize on varying data. For a fair comparison with state-of-the-art methods, we fine-tune our model on the training set of the BlendedMVS dataset [46] using the original image resolution (576×768) and $N = 5$. More details about the fine-tuning process can be found in supp. materials. Similar to other methods [11,39], the camera parameters, depth ranges, and neighboring view selection are aligned with [45]. We use images of the original resolution for inference. Quantitative results are shown in Tab. 2 and Tab. 3, and the qualitative comparisons ares shown in Fig. 7. When applied on CasMVSNet [11], our method ranks 1^{st} among all submitted methods (including supervised methods) on intermediate set of Tanks and Temples online benchmark [19] by Mar. 5, 2022.

Table 2: Quantitative results on the intermediate set of Tanks and Temples benchmark [1]. Sup. indicates whether the method is supervised or not. **Bold** and underlined figures indicate the best and the second best results.

Method	Sup.	Mean	Family	Francis	Horse	L.H.	M60	Panther	P.G.	Train
COLMAP [31]	-	42.14	50.41	22.25	26.63	56.43	44.83	46.97	48.53	42.04
ACMM [40]	-	57.27	69.24	51.45	46.97	63.20	55.07	57.64	60.08	54.48
AttMVS [22]	-	60.05	73.90	62.58	44.08	**64.88**	56.08	59.39	**63.42**	56.06
MVS^2 [4]	✗	37.21	47.74	21.55	19.50	44.54	44.86	46.32	43.38	29.72
M^3VSNet [14]	✗	37.67	47.74	24.38	18.74	44.42	43.45	44.95	47.39	30.31
JDACS [38]	✗	45.48	66.62	38.25	36.11	46.12	46.66	45.25	47.69	37.16
Self-supervised-CVP-MVSNet [42]	✗	46.71	64.95	38.79	24.98	49.73	52.57	51.53	50.66	40.52
U-MVS+CasMVSNet [39]	✗	57.15	76.49	60.04	49.20	55.52	55.33	51.22	56.77	52.63
CasMVSNet [11]	✓	56.84	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51
Vis-MVSNet [48]	✓	60.03	77.40	60.23	47.07	63.44	62.21	57.28	60.54	52.07
AA-RMVSNet [36]	✓	61.51	77.77	59.53	51.53	64.02	64.05	59.47	60.85	54.90
EPP-MVSNet [23]	✓	61.68	77.86	60.54	52.96	62.33	61.69	60.34	62.44	55.30
Ours+CasMVSNet	✗	**64.14**	**80.42**	**67.42**	**54.02**	64.52	**64.18**	**61.60**	62.37	**58.59**

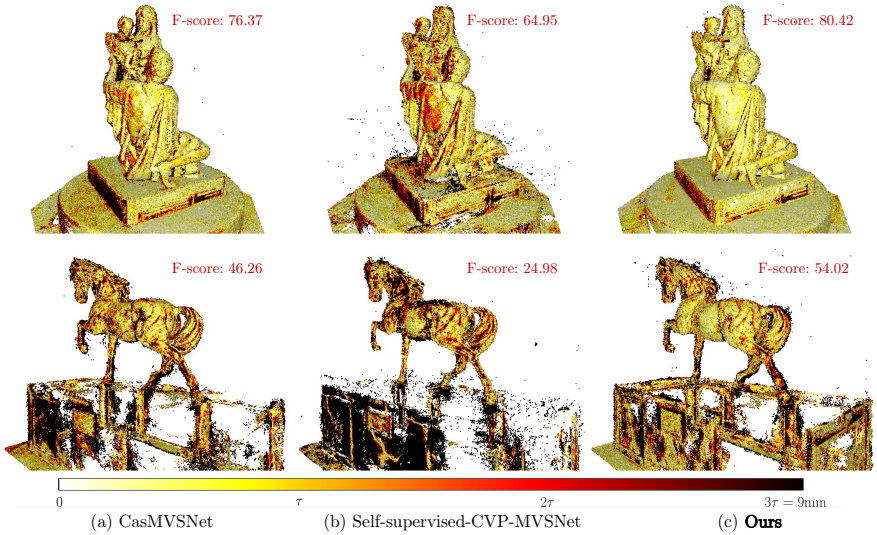
Table 3: Quantitative results on the advanced set of Tanks and Temples benchmark [1].

Method	Sup.	Mean	Auditorium	Ballroom	Courtroom	Museum	Palace	Temple
COLMAP [31]	-	27.24	16.02	25.23	34.70	41.51	18.05	27.94
ACMM [40]	-	34.02	23.41	32.91	**41.17**	48.13	23.87	34.60
CasMVSNet [11]	✓	31.12	19.81	38.46	29.10	43.87	27.36	28.11
AA-RMVSNet [36]	✓	33.53	20.96	40.15	32.05	46.01	29.28	32.71
Vis-MVSNet [48]	✓	33.78	20.79	38.77	32.45	44.20	28.73	37.70
EPP-MVSNet [23]	✓	35.72	21.28	39.74	35.34	**49.21**	30.00	**38.75**
Ours+CasMVSNet	✗	**37.96**	**27.22**	**44.10**	35.53	49.16	**34.67**	37.11

BlendedMVS Dataset We further demonstrate the quality of depth maps on the validation set of BlendedMVS dataset [46]. The details of the training process can be found in supp. materials. We set $N = 5$, image resolution as 512×640, and apply the evaluation metrics described in [5] where depth values are normalized to make depth maps with different depth ranges comparable. Quantitative results are illustrated in Tab. 4. EPE stands for the endpoint error, which is the average ℓ-1 distance between the prediction and the ground truth depth; e_1 and e_3 represent the percentage of pixels with depth error larger than 1 and larger than 3.

4.4 Ablation Study

Implementation of Featuremetric Loss As analyzed in Sec. 3.1, we consider the nature of the MVS is multi-view feature matching along epipolar lines, where the features are supposed to be relatively locally discriminative. Tab. 5 shows the quantitative results of different settings. Compared with using photometric loss only, both internal featuremetric and external featuremetric loss can boost the



(a) CasMVSNet (b) Self-supervised-CVP-MVSNet (c) **Ours**

Fig. 7: Comparison of reconstructed results with the supervised baseline Cas-MVSNet [11] and the state-of-the-art self-supervised method [42] on Tanks and Temples benchmark [19]. $\tau = 3mm$ is the distance threshold determined officially and darker regions indicate larger error encountered with regard to τ.

performance. And our proposed internal featuremetric loss shows superiority over the external featuremetric loss with external features by a pre-trained ResNet. It is worth noting that it is not feasible to adopt our featuremetric loss alone. The reason is that the feature network is online trained within the MVS network and thus applying featuremetric loss alone will lead to failure of training where features tend to be a constant (typically 0).

Number of Self-training Iterations Given the scheme of knowledge distillation via generating pseudo probability, we can iterate the distillation-based student training for an arbitrary number of loops. Here we study the performance gain when the number of iterations increases in Tab. 6. As a trade-off of efficiency and accuracy, we set the number of iterations to be 2.

5 Discussion

5.1 Insights of Effectiveness

We attribute the effectiveness of KD-MVS to the following four parts. (a) The first one is multi-view consistency as introduced in Sec. 3.2, which can be used to filter the outliers in noisy raw depth maps. The remaining inliers are relatively accurate and are equivalent to ground-truth depth to a certain extent.

Table 4: Quantitative results towards predicted depth maps on BlendedMVS validation set [46] (**lower is better**).

Method	Sup.	EPE	e_1	e_3
MVSNet [44]	✓	1.49	21.98	8.32
CVP-MVSNet [43]	✓	1.90	19.73	10.24
CasMVSNet [11]	✓	1.43	19.01	9.77
Vis-MVSNet [48]	✓	1.47	15.14	5.13
EPP-MVSNet [23]	✓	1.17	12.66	6.20
Ours	✗	**1.04**	**10.17**	**4.94**

Table 5: Ablation study on loss for self-supervised training stage (teacher model). $\mathcal{L}_{\text{fea}}$ and $\mathcal{L}^*_{\text{fea}}$ denotes feature-metric loss by the internal feature encoder and by an external pretrained encoder (ResNet-18 [12]) respectively.

$\mathcal{L}_{\text{photo}}$	$\mathcal{L}^*_{\text{fea}}$	$\mathcal{L}_{\text{fea}}$	Acc.	Comp.	Overall
✓			0.489	0.501	0.495
✓	✓		0.477	0.441	0.459
✓		✓	**0.457**	**0.399**	**0.428**

Table 6: Ablation study on the number of iterations for distillation training. Note that we consider the number of distillation rounds equal to the number of times fused depth is generated and verified.

#round(s)	Acc.	Comp.	Overall
1	0.387	0.334	0.361
2	0.359	**0.295**	**0.327**
3	**0.357**	0.298	**0.327**
4	**0**.358	0.297	0.328

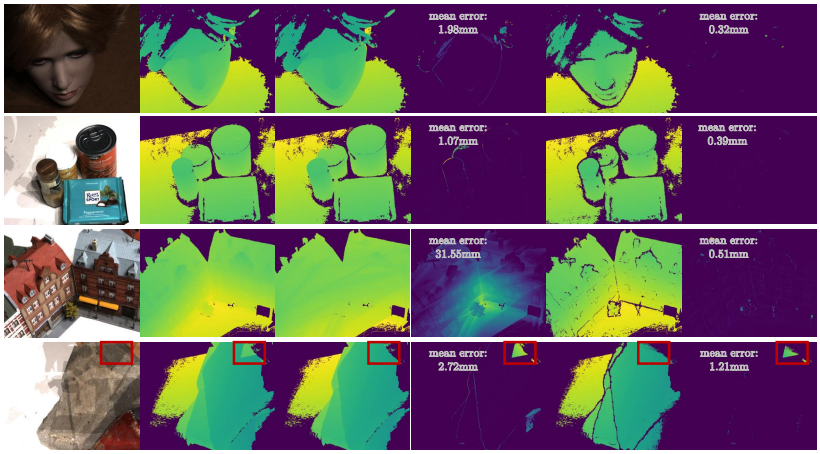
Table 7: Ablation study on the main factor of effectiveness. Mask indicates whether to use the validated mask. Depth indicates using ground truth depth or validated depth. Loss indicates which loss is used.

	Mask	Depth	Loss	Acc.	Comp.	Over.
(1)	✗	GT	ℓ-1	0.358	0.346	0.352
(2)	✓	GT	ℓ-1	**0.352**	0.334	0.343
(3)	✓	vali.	ℓ-1	0.361	0.331	0.346
(4)	✓	vali.	distill	0.359	**0.295**	**0.327**

(b) The probabilistic knowledge brings performance gain to the student model. Compared with using hard labels such as ℓ-1 loss and depth labels, applying soft probability distribution to student model brings additional inter-depth relationships and thus reduces the ambiguity of noisy 3D points. (c) The validated depth contains less perspective error than rendered ground truth labels. As shown in the last row of Fig. 8 (marked with a red box), there are some incorrect values in the ground-truth depth maps of DTU dataset [2] caused by perspective error, which is harmful to training MVS models. (d) The validated masks of the teacher model reduce the ambiguity of prediction by filtering the samples which are hard to learn, benefiting the convergence of the model. We perform an ablation study on these parts as shown in Tab. 7. (1) and (2) show that the validated mask is helpful, (3) and (4) show that enforcing the probability distribution can bring significant improvement. More details can be found in supp. materials.

5.2 Comparisons to SOTA Methods

U-MVS [39] leverages optical flow to compute a depth-flow consistency loss. To get reliable optical flow labels, U-MVS trains a PWC-Net [33] on DTU dataset [2] in a self-supervised manner, which costs additional training time and needs storage space for the pseudo optical flow labels (more than $120GB$).



(a) RGB (b) GT depth (c) Rendered depth (d) Rendered error (e) Our depth (f) Our error

Fig. 8: Visualization of depth maps and errors. (a) RGB reference images; (b) ground-truth depth maps; (c) rendered depth maps by [42]; (d) errors between (b) and (c); (e) pseudo labels in KD-MVS; (f) errors between (b) and (e). We apply the same mask on (b)(c)(e) for better visualization.

Self-supervised-CVP-MVSNet [42] renders depth maps from the reconstructed meshes, which brings in error during Poisson reconstruction [17]. We compare the rendered depth maps [42] and our validated depth maps in Fig. 8.

5.3 Limitations

- The quality of pseudo probability distribution highly depends on the cross-view check stage and relevant hyperparameters need to be tuned carefully.
- Knowledge distillation is known as data-hungry and it may not work as expected with a relatively small-scale dataset.

6 Conclusion

In this paper, we propose KD-MVS, which is a general self-supervised pipeline for MVS networks without any ground-truth depth as supervision. In the self-supervised teacher training stage, we leverage a featuremetric loss term, which is more robust than photometric loss alone. The features are yielded internally by the MVS network itself, which is end-to-end trained under implicit supervision. To explore the potential of self-supervised MVS, we adopt the idea of knowledge distillation and distills the teacher's knowledge to a student model by generating pseudo probability distribution. Experimental results indicate that the self-supervised training pipeline has the potential to obtain reconstruction quality equivalent to supervised ones.

References

1. Tanks and temples benchmark. https://www.tanksandtemples.org 2, 11
2. Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. International Journal of Computer Vision **120**(2), 153–168 (2016) 2, 9, 10, 13
3. Cheng, S., Xu, Z., Zhu, S., Li, Z., Li, L.E., Ramamoorthi, R., Su, H.: Deep stereo using adaptive thin volume representation with uncertainty awareness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2524–2534 (2020) 3, 9
4. Dai, Y., Zhu, Z., Rao, Z., Li, B.: Mvs2: Deep unsupervised multi-view stereo with multi-view symmetry. In: 2019 International Conference on 3D Vision (3DV). pp. 1–8. IEEE (2019) 1, 2, 3, 9, 11
5. Darmon, F., Bascle, B., Devaux, J.C., Monasse, P., Aubry, M.: Deep multi-view stereo gone wild. arXiv preprint arXiv:2104.15119 (2021) 11
6. Ding, Y., Li, Z., Huang, D., Li, Z., Zhang, K.: Enhancing multi-view stereo with contrastive matching and weighted focal loss. arXiv preprint arXiv:2206.10360 (2022) 1
7. Ding, Y., Yuan, W., Zhu, Q., Zhang, H., Liu, X., Wang, Y., Liu, X.: Transmvsnet: Global context-aware multi-view stereo network with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8585–8594 (2022) 1
8. Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., Anandkumar, A.: Born again neural networks. In: International Conference on Machine Learning. pp. 1607–1616. PMLR (2018) 3
9. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 873–881 (2015) 9
10. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**(6), 1789–1819 (2021) 3
11. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2495–2504 (2020) 1, 2, 3, 5, 7, 9, 10, 11, 12, 13
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2016) 6, 13
13. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 **2**(7) (2015) 1, 3
14. Huang, B., Yi, H., Huang, C., He, Y., Liu, J., Liu, X.: Mˆ 3vsnet: Unsupervised multi-metric multi-view stereo network. arXiv preprint arXiv:2004.09722 (2020) 1, 3, 9, 11
15. Huang, Z., Wang, N.: Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint arXiv:1707.01219 (2017) 3, 7
16. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016) 2, 5
17. Kazhdan, M., Hoppe, H.: Screened poisson surface reconstruction. ACM Transactions on Graphics (ToG) **32**(3), 1–13 (2013) 14

18. Khot, T., Agrawal, S., Tulsiani, S., Mertz, C., Lucey, S., Hebert, M.: Learning unsupervised multi-view stereopsis via robust photometric consistency. arXiv preprint arXiv:1905.02706 (2019) 1, 2, 3, 5, 9
19. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (ToG) **36**(4), 1–13 (2017) 2, 9, 10, 12
20. Li, T., Li, J., Liu, Z., Zhang, C.: Few sample knowledge distillation for efficient network compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14639–14647 (2020) 3
21. Liao, J., Ding, Y., Shavit, Y., Huang, D., Ren, S., Guo, J., Feng, W., Zhang, K.: Wt-mvsnet: Window-based transformers for multi-view stereo. arXiv preprint arXiv:2205.14319 (2022) 1
22. Luo, K., Guan, T., Ju, L., Wang, Y., Chen, Z., Luo, Y.: Attention-aware multi-view stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1590–1599 (2020) 11
23. Ma, X., Gong, Y., Wang, Q., Huang, J., Chen, L., Yu, F.: Epp-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5732–5740 (2021) 11, 13
24. Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 5191–5198 (2020) 3
25. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3967–3976 (2019) 3
26. Passalis, N., Tefas, A.: Learning deep representations with probabilistic knowledge transfer. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 268–284 (2018) 3
27. Passalis, N., Tefas, A.: Learning deep representations with probabilistic knowledge transfer. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 268–284 (2018) 3
28. Passalis, N., Tzelepi, M., Tefas, A.: Probabilistic knowledge transfer for lightweight deep representation learning. IEEE Transactions on Neural Networks and Learning Systems **32**(5), 2030–2039 (2020) 3
29. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5007–5016 (2019) 3
30. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014) 3, 7
31. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision. pp. 501–518. Springer (2016) 9, 11
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 6
33. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8934–8943 (2018) 13
34. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. arXiv preprint arXiv:1910.10699 (2019) 3

35. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1365–1374 (2019) 3, 7
36. Wei, Z., Zhu, Q., Min, C., Chen, Y., Wang, G.: Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6187–6196 (2021) 3, 9, 11
37. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10687–10698 (2020) 3
38. Xu, H., Zhou, Z., Qiao, Y., Kang, W., Wu, Q.: Self-supervised multi-view stereo via effective co-segmentation and data-augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 2, p. 6 (2021) 1, 3, 9, 11
39. Xu, H., Zhou, Z., Wang, Y., Kang, W., Sun, B., Li, H., Qiao, Y.: Digging into uncertainty in self-supervised multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6078–6087 (2021) 1, 3, 9, 10, 11, 13
40. Xu, Q., Tao, W.: Multi-scale geometric consistency guided multi-view stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5483–5492 (2019) 11
41. Yan, J., Wei, Z., Yi, H., Ding, M., Zhang, R., Chen, Y., Wang, G., Tai, Y.W.: Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In: European Conference on Computer Vision. pp. 674–689. Springer (2020) 3
42. Yang, J., Alvarez, J.M., Liu, M.: Self-supervised learning of depth inference for multi-view stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7526–7534 (2021) 1, 2, 3, 9, 11, 12, 14
43. Yang, J., Mao, W., Alvarez, J.M., Liu, M.: Cost volume pyramid based depth inference for multi-view stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4877–4886 (2020) 3, 13
44. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 767–783 (2018) 1, 3, 9, 10, 13
45. Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L.: Recurrent mvsnet for high-resolution multi-view stereo depth inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5525–5534 (2019) 1, 3, 10
46. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1790–1799 (2020) 2, 9, 10, 11, 13
47. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016) 3
48. Zhang, J., Yao, Y., Li, S., Luo, Z., Fang, T.: Visibility-aware multi-view stereo network. British Machine Vision Conference (BMVC) (2020) 3, 11, 13