CL-MVSNet: Unsupervised Multi-view Stereo with Dual-level Contrastive Learning

Kaiqiang Xiong [1] Rui Peng [1] Zhe Zhang [1] Tianxing Feng [1]
Jianbo Jiao [4] Feng Gao [5] Ronggang Wang ✉[1,2,3]
[1]School of Electronic and Computer Engineering, Peking University
[2]Peng Cheng Laboratory [3]Migu Culture Technology Co., Ltd
[4]School of Computer Science, University of Birmingham [5]School of Arts, Peking University
xiongkaiqiang@stu.pku.edu.cn rgwang@pkusz.edu.cn
https://KaiqiangXiong.github.io/CL-MVSNet/

Abstract

Unsupervised Multi-View Stereo (MVS) methods have achieved promising progress recently. However, previous methods primarily depend on the photometric consistency assumption, which may suffer from two limitations: indistinguishable regions and view-dependent effects, e.g.*, low-textured areas and reflections. To address these issues, in this paper, we propose a new dual-level contrastive learning approach, named* ***CL-MVSNet****. Specifically, our model integrates two contrastive branches into an unsupervised MVS framework to construct additional supervisory signals. On the one hand, we present an image-level contrastive branch to guide the model to acquire more context awareness, thus leading to more complete depth estimation in indistinguishable regions. On the other hand, we exploit a scene-level contrastive branch to boost the representation ability, improving robustness to view-dependent effects. Moreover, to recover more accurate 3D geometry, we introduce an $\mathcal{L}0.5$ photometric consistency loss, which encourages the model to focus more on accurate points while mitigating the gradient penalty of undesirable ones. Extensive experiments on DTU and Tanks&Temples benchmarks demonstrate that our approach achieves state-of-the-art performance among all end-to-end unsupervised MVS frameworks and outperforms its supervised counterpart by a considerable margin without fine-tuning.*

1. Introduction

Multi-View Stereo (MVS) is a critical task in various applications, including robotics, self-driving, and VR/AR. The goal of MVS is to estimate a dense 3D reconstruction from multiple images captured from different views. Traditionally, it has been approached by computing dense correspondences between images, often based on hand-crafted similarity metrics and engineered regularizations [4, 33]. Recently, a surge of learning-based methods [7, 49–51] have been developed to advance the effectiveness of MVS, showing promising results in MVS benchmarks [1, 19]. However, most of them are supervised methods [3, 9, 27]. These methods heavily rely on large-scale ground-truth 3D training data, which are expensive to acquire.

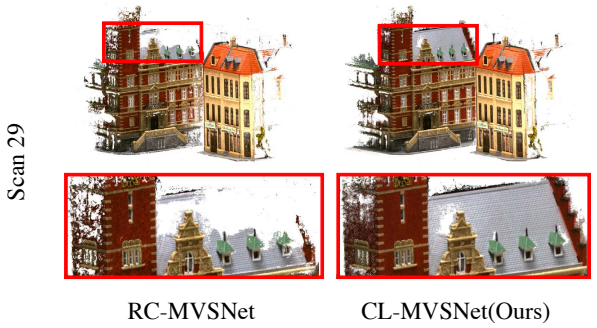


Figure 1. **Qualitative comparison of reconstruction quality with the SOTA method [5] on scan29 of DTU [1].** Our method performs better on repetitive patterns.

To tackle this problem, unsupervised MVS methods [5, 17, 41] have attempted to train MVS networks without annotations. Existing methods mainly depend on the hypothesis of photometric consistency, which states that the appearance of a point in 3D space is invariant across different views. However, this hypothesis may be ineffective owing to indistinguishable regions and view-dependent effects, *e.g.*, low texture, repetitive patterns, and reflections. Recently, the state-of-the-art (SOTA) method RC-MVSNet [5] adopts a rendering consistency network to address the ambiguity caused by view-dependent photometric effects and occlusions. While achieving promising results, this ap-

proach may suffer from significant performance degradation in indistinguishable regions, as shown in Fig. 1. To address the absence of valid supervisory signals, we propose to leverage contrastive learning to boost the robustness and generalizability of unsupervised MVS in various scenarios.

Contrastive learning is a widely used paradigm in unsupervised learning to construct additional supervisory signals. With the success of contrastive learning in images, [41, 53] have extended its application to unsupervised MVS with color fluctuation augmentations or part segmentations. However, there are still many remaining unresolved issues, *e.g.*, occlusions and low-textured surfaces. And the potential of incorporating contrastive learning into MVS remains largely under-explored. Moreover, recent studies [2, 55] showed that *hard positive samples are of benefit to boost the contrastive learning*. Inspired by this, we propose a dual-level contrastive learning approach, named **CL-MVSNet**, where image-level and scene-level contrastive learning are integrated into an unsupervised MVS framework.

To resolve ambiguity from indistinguishable regions, we introduce an image-level contrastive learning strategy to encourage the model to be more context-aware. Specifically, for an image-level contrastive sample, all pixels in the source images are masked with independent and identically distributed Bernoulli probability, simulating the case that local photometric consistency fails. Following that, we maximize the similarity between the depth estimations of the regular sample and the image-level contrastive sample. The intuition is that the augmented images contain the same context information as the original ones, which can also be utilized to estimate complete depth maps as hard positive samples. In this way, the network is encouraged to exploit more contextual information instead of relying only on photometric consistency over small regions.

In addition, we propose a scene-level contrastive learning branch to alleviate the view-dependent photometric effects. Due to severe occlusions, reflections, and illumination changes, source images with few overlaps are often infeasible to use in unsupervised MVS. However, from the perspective of contrastive learning, a scene-level contrastive sample containing randomly selected source images can be considered a natural hard positive sample. These hard samples are expected to produce the same 3D representation as the regular samples, as they are captured from identical scenes. Therefore, we enforce contrastive consistency between the scene-level contrastive branch and the regular branch, encouraging the model to learn more powerful 3D cost volume regularization for robust depth estimation.

Furthermore, we propose an $\mathcal{L}0.5$ photometric consistency loss to support the training of the CL-MVSNet. In an MVS system, after depth estimation, most points with undesirable depth predictions will be filtered out before depth fusion. Besides, these points often locate in occluded areas or useless backgrounds where photometric consistency enforcement may mislead the model. To this end, instead of using vanilla photometric consistency loss based on $\mathcal{L}1$ or $\mathcal{L}2$ norm, we propose to use the $\mathcal{L}0.5$ norm, which has larger gradients with regard to smaller errors. In this way, the model increases the penalty of accurate points, resulting in more accurate survival points.

In conclusion, our main contributions are:

- We present an image-level contrastive consistency loss, which encourages the model to be more context-aware and recover more complete reconstruction in indistinguishable regions.
- We propose a scene-level contrastive consistency loss, which boosts the representation ability to promote robustness to view-dependent effects.
- We propose an $\mathcal{L}0.5$ photometric consistency loss to further advance the contrastive learning framework, which enables the model to focus on accurate points, resulting in more accurate reconstruction.
- Experiments on DTU [1] and Tanks&Temples [19] benchmarks show that our method outperforms state-of-the-art end-to-end unsupervised models and surpasses its supervised counterpart.

2. Related Work

Supervised MVS. With the development of deep learning technique and large-scale 3D datasets, supervised MVS has achieved significant progresses in recent years [23–25, 29, 35–37, 39, 44, 45, 48–51, 54]. MVSNet [50] proposes a popular MVS pipeline, which can be summarized as four steps: feature extraction, cost aggregation, cost volume regularization, and depth regression. Recent works make efforts to relieve the huge memory and computation cost, by introducing a multi-stage architecture [7, 13, 27, 49], RNN [51] and some other methods [35, 37]. However, all the above approaches depend on labeled training data, which are expensive to obtain in practice.

End-to-end Unsupervised and Multi-stage Self-supervised MVS. Following [5], we categorize current multi-view stereo methods trained without any annotations into two dominant groups: the end-to-end unsupervised MVS and the multi-stage self-supervised MVS. The end-to-end unsupervised MVS methods [5, 8, 15, 17, 21, 41] primarily rely on the hypothesis of photometric consistency. Concretely, Unsup-MVSNet proposes the first end-to-end unsupervised MVS framework. It enforces photometric consistency between the reconstructed images and the reference image. In addition, it utilizes structured similarity

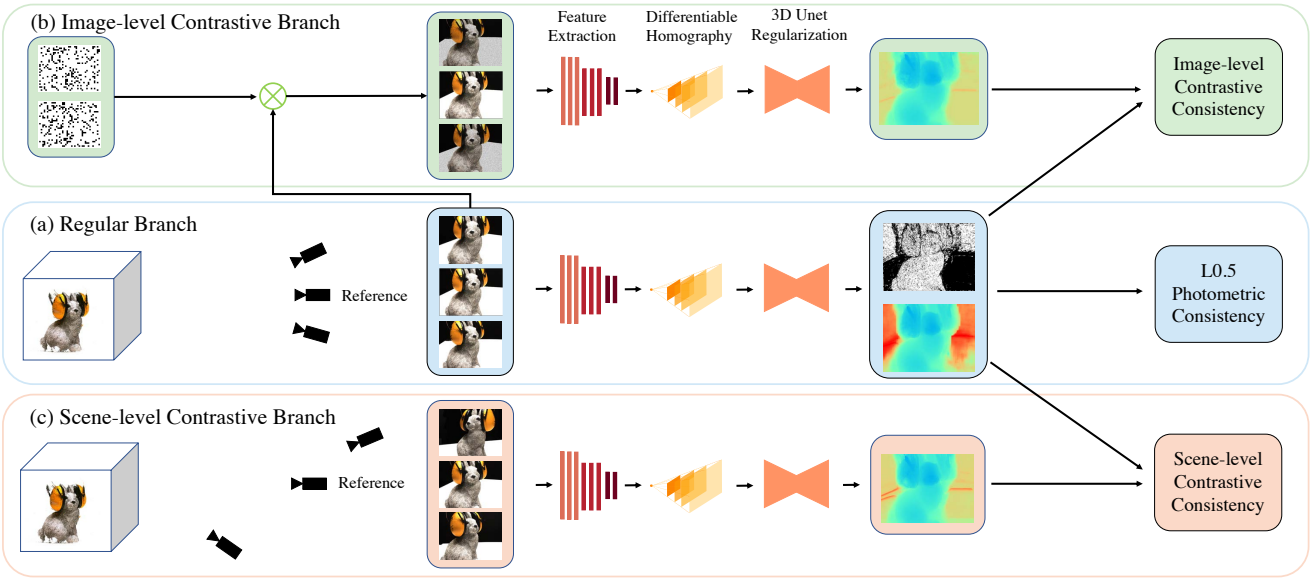


Figure 2. **The framework of CL-MVSNet.** The framework consists of: (a) a Regular Branch with a regular sample similar to CasMVSNet [13], (b) an Image-level Contrastive Branch with the image-level contrastive sample , (c) a Scene-level Contrastive Branch with the scene-level contrastive sample. To pull positive pairs close, we enforce contrastive consistency between the regular branch and two contrastive branches, with the confidence mask estimated from the regular branch. Moreover, our proposed $\mathcal{L}0.5$ photometric consistency is enforced between the reconstructed images and the input reference image on the regular branch for more accurate reconstruction.

loss and depth smoothness loss to further improve the performance. However, photometric consistency is ineffective in many challenging areas. Thus, recent researchers have attempted to incorporate pretext tasks into training to find reliable supervisory signals, *e.g.*, normal-depth consistency [15], semantic consistency [41], and rendering consistency [5]. However, there is still a performance gap between unsupervised methods and supervised SOTA methods.

While multi-stage self-supervised methods [10, 30, 42, 47] aim to obtain reliable pseudo-labels by filtering and processing inferred depth maps, which are used to supervise model training in subsequent stages. However, these methods cannot be trained in an end-to-end manner since they command complex pre-processing and fine-tuning. Additionally, the generated pseudo-labels require extra storage space, and iterative self-training can take significant time.

Contrastive Learning. Contrastive learning [6, 12, 14, 28] is a popular paradigm in unsupervised learning that encourages the model to be invariant to multiple transformations of a single sample. Specifically, contrastive learning pulls positive sample pairs close while pushing negative sample pairs away. This approach has achieved remarkable success in narrowing the performance gap between unsupervised and supervised models. Recently, hard positive samples have been confirmed to be beneficial for boosting contrastive learning [2, 55]. For this purpose, some studies have explored pixel- [40], image- [14], and object-level [20] contrastive learning. Most image-level contrastive learning methods rely on well-designed augmentation procedures. However, pixel- and object-level supervision provides a way to directly define positive and negative samples. In this work, we propose to leverage a dual-level contrastive learning strategy to boost the unsupervised MVS, which includes image-level and scene-level contrastive learning.

3. Method

In this section, the main contributions of CL-MVSNet will be elaborated. We firstly depict the unsupervised backbone (in Sec. 3.1), then we describe the proposed image-level contrastive consistency (in Sec. 3.2), the scene-level contrastive consistency (in Sec. 3.3), and the $\mathcal{L}0.5$ photometric consistency (in Sec. 3.4). Finally, we introduce the overall loss function during training (in Sec. 3.5). An overview of our architecture is shown in Fig. 2. Note that CL-MVSNet is a general framework suitable for arbitrary learning-based MVS. And we take the representative CasMVSNet [13] as our backbone in this work.

3.1. Unsupervised Multi-view Stereo

To begin with, we adopt the same view-selection strategy as CasMVSNet [13] to construct a regular sample. For a given regular sample comprised of 1 reference image I_1, $N-1$ source images $\{I_i\}_{i=2}^{N}$ and their camera parameters $\{K_i, T_i\}_{i=1}^{N}$, our goal is to estimate the corresponding depth map with the backbone network. Specifically, the backbone consists of four steps: feature extraction, cost volume construction, cost volume regularization, and depth regression.

During feature extraction, images $\{I_i\}_{i=1}^{N}$ are fed into a shared Feature Pyramid Network [22] to generate 2D pixel-

wise features $\{F_i\}_{i=1}^N$ at three stages with incremental resolutions. At the coarsest stage, with initial depth hypothesis $\{d_{min}, \dots, d_{max}\}$, 3D feature volumes $\{V_i\}_{i=1}^N$ are built from features $\{F_i\}_{i=1}^N$ via differentiable homography.

As for cost volume aggregation, we adopt group-wise correlation metric following [27, 35, 45, 54]. We divide the N feature volumes $\{V_i\}_{i=1}^N$ of N_C channels into N_G groups, then construct the raw cost volume C as below:

$$\mathbf{C} = \frac{1}{(N-1)N_C/N_G} \sum_{i=2}^{N-1} \langle V_1^g, V_i^g \rangle_{g=1}^{N_G}. \quad (1)$$

Then the raw cost volume $\mathbf{C}$ undergoes a regular 3D U-Net and a softmax, resulting in a probability volume P_v. Finally, the depth map D is obtained by weighted sum:

$$D = \sum_{d=d_{min}}^{d_{max}} d \times P_v(d). \quad (2)$$

In a coarse-to-fine fashion, the coarse depth map D^1 is used to generate the depth hypothesis of the next stage. With features $\{F_i^2, F_i^3\}_{i=1}^N$ at the larger resolution, finer depth maps $\{D^2, D^3\}$ will be estimated iteratively.

Depth estimation with Confidence Mask. During the depth estimation, there is a byproduct probability volume P_v, measuring the pixel-wise confidence of the depth hypothesis. Then a probability map P_m can be acquired by taking the probability sum over the four nearest depth hypotheses with regard to depth estimation [50]. We generate a binary confidence mask M_c that indicates whether the model is confident about the pixel-wise depth estimation:

$$M_c = P_m > \gamma, \quad (3)$$

where the γ is set to 0.95 in our implementation. The depth estimation D_R from the regular branch and the confidence mask M_c will be used to regularize the outputs of the two proposed contrastive branches (in Sec. 3.2 and Sec. 3.3).

Vanilla Photometric Consistency Loss. Previous unsupervised MVS methods [5, 8, 10, 15, 17, 30, 41, 42, 47] use a photometric consistency loss to train an unsupervised MVS network without any ground truth depth, as shown in Fig. 3.

Specifically, given a reference image I_1, a source image I_i, associated intrinsic K, relative transformation T, and an inferred depth map D_R, the source image I_i is warped to reconstruct reference image $\hat{I}_i$. For a specific pixel p in reconstructed reference image $\hat{I}_i$, its coordinate p' in source image I_i can be calculated via inverse warping:

$$p' = KT(D_R(p) \cdot K^{-1}p). \quad (4)$$

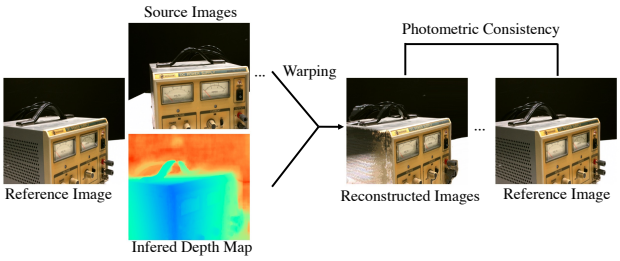


Figure 3. **Photometric consistency.** The source images are warped to reconstruct the reference image with the inferred depth map on the reference view. Then consistency is enforced between the reconstructed images and the reference image.

Then the reconstructed image $\hat{I}_i$ can be obtained via differentiable bilinear sampling $\hat{I}_i(p) = I_i(p')$.

Along with the inverse warping process, a binary valid mask M_i is generated, indicating valid pixels in the reconstructed image $\hat{I}_i$. In previous unsupervised MVS methods, all source images $\{I_i\}_{i=2}^N$ are warped to the reference view according to the inferred reference depth D_R, then the vanilla photometric consistency loss can be computed as:

$$\mathrm{L}_{PC} = \sum_{i=2}^{N} \frac{\|(\hat{I}_i - I_1) \odot M_i\|_\theta + \|(\nabla \hat{I}_i - \nabla I_1) \odot M_i\|_\theta}{\|M_i\|_1}, \quad (5)$$

where ∇ refers to the gradient operator, $\odot$ is element-wise product, θ=1 or 2, which denotes the $\mathcal{L}1$ or $\mathcal{L}2$ norm.

3.2. Image-level Contrastive Consistency

As mentioned before, the local photometric consistency fails to offer valid supervisory signals in indistinguishable regions, *e.g.*, areas with low texture or repetitive patterns. To overcome this problem, [11] explicitly introduces a patch-wise photometric consistency loss to enhance the matching robustness. However, the model may suffer notable performance degradation for inappropriate handcrafted patch size, *e.g.*, the large patch may lead to the loss of accuracy in rich-textured areas. In this work, we propose an alternative to implicitly encourage the model to be context-aware by introducing an image-level contrastive branch. The image-level contrastive sample and image-level contrastive consistency loss will be elaborated on next.

Image-level Contrastive Sample. The image-level contrastive sample is generated by applying transformation on a given regular sample $\{I_i\}_{i=1}^N$ artificially. For a source image I_i with the size of $H \times W \times C$, we set an occlusion rate α to construct a binary pixel-wise mask with the size of $H \times W$. Concretely, any element in the mask will draw a value of 1 according to the given occlusion rate α:

$$M_{o(i,j)} \sim B(\alpha), \quad (6)$$

Figure 4. **Image-level contrastive sample.** (a) a source image of the regular sample. (b) a source image of the image-level contrastive sample. A Bernoulli-distributed binary mask is used to simulate the failure case of local photometric consistency in (b).

where B denotes the Bernoulli distribution, the occlusion mask is denoted as M_o. Then M_o is used to mask the image I_i as shown in Fig. 4. Simply processing all source images in this way, an image-level contrastive sample $\{I_1, I_2', I_3', \ldots, I_N'\}$ is generated. Then the hard positive sample will be fed to the net to obtain the depth estimation D_{IC}. To ensure training stability, we employ a curriculum learning strategy to gradually increase the occlusion rate α, which grows from 0 to 0.1 in our implementation.

Image-level Contrastive Consistency Loss. To enforce consistency between the regular branch and the image-level contrastive branch, we compute the image-level contrastive consistency loss L_{ICC} as:

$$L_{ICC} = \frac{\|(D_R - D_{IC}) \odot M_c\|_1}{\|M_c\|_1}, \tag{7}$$

where D_R and D_{IC} denote the inferred depth of the regular branch and the image-level contrastive branch respectively.

3.3. Scene-level Contrastive Consistency

To the best of our knowledge, all unsupervised MVS methods including our regular sample apply photometric and geometric priors for view selection following MVSNet [50]. Specifically, images with more overlaps with the reference image will get higher scores, and the $N-1$ images with the highest scores will be selected as source images in the regular sample. Note that some supervised methods [4, 34] randomly select source views to improve the robustness. Due to severe view-dependent effects and occlusions where local photometric consistency fails (in Fig. 5), directly using this strategy in the regular branch of unsupervised MVS will lead to worse performance. However, from a contrastive learning perspective, these images can be used to construct hard positive samples. Below, we will introduce how to generate the scene-level contrastive sample and enforce the scene-level contrastive consistency in detail.

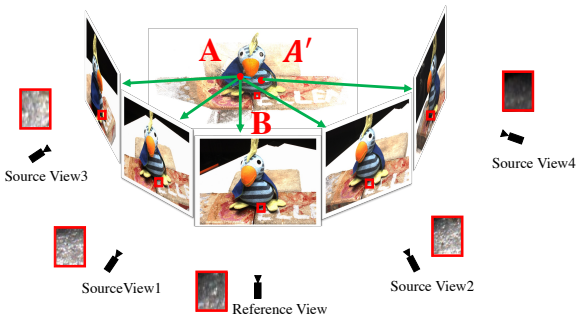


Figure 5. **View-dependent effects and occlusions.** From source view4, point A is occluded and only point A' is visible along the line of sight. Besides, The appearance of the identical region B differs in different views due to variations in illumination, camera exposure, and reflections.

Scene-level Contrastive Sample. For a given reference image I_1, the scene-level contrastive sample can be constructed by combining the reference image $I_1$with $N-1$ randomly selected image $\{I_i''\}_{i=2}^N$ of the same scene. Afterward, a depth map D_{SC} will be inferred for the contrastive sample. It is worth noting that the cost volume built with the scene-level contrastive sample represents the identical 3D scene as the regular ones. Thus, the network is supposed to gain the same depth estimation from the scene-level contrastive sample as the regular ones.

Scene-level Contrastive Consistency Loss. Then a scene-level contrastive consistency loss L_{SCC} will be applied on the scene-level contrastive branch. The loss is used to pull the scene-level contrastive sample closer to the regular sample, improving the robustness to view-dependent effects:

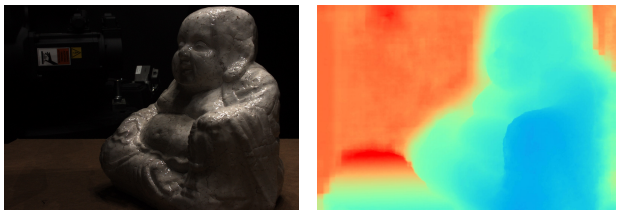
$$L_{SCC} = \frac{\|(D_R - D_{SC}) \odot M_c\|_1}{\|M_c\|_1}, \tag{8}$$

where D_R and D_{SC} refer to the depth map from the regular branch and the scene-level contrastive branch respectively.

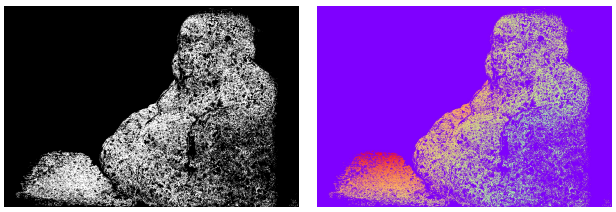
3.4. $\mathcal{L}0.5$ Photometric Consistency

In the generic pipeline of MVS, a depth filtering process is applied before depth fusion, in order to mask out most undesirable points, as shown in Fig. 6. For clarity, we divide the points in the inferred depth map into three categories:

- *accurate point*: the pixel with accurate prediction, which will survive the depth filtering and contribute to the final reconstruction.
- *ordinary point*: the pixel close to the accurate point, which will be filtered out.



Reference Image | Infered Depth Map



Geometric Filtering Mask | Filtered Depth Map

Figure 6. **Depth filtering.** Most points with inaccurate depth will be filtered out before depth fusion. It can be observed that these points usually locate in occluded regions or useless backgrounds, where photometric consistency fails.

- *terrible point*: the pixel with pretty erroneous prediction, most of which locate in useless backgrounds or occluded areas and will be filtered out. Moreover, the photometric consistency enforced on these points may mislead the model.

In the MVS system, an ideal depth map is supposed to contain more *accurate points*. Moreover, a loss function that imposes more constraints on *accurate points* will be desirable due to this observation. We perform an analysis of the $\mathcal{L}1$-norm and $\mathcal{L}2$-norm, which are broadly used in vanilla photometric consistency loss (Eq. (5)). We first write their expressions and gradient formulas:

$$\begin{cases} l_1(e) = \|e\|_1 = (\sum_{i=1}^{n} e_i), \dfrac{\partial l_1(e)}{\partial e_i} = 1, \\ l_2(e) = \|e\|_2 = (\sum_{i=1}^{n} e_i^2)^{\frac{1}{2}}, \dfrac{\partial l_2(e)}{\partial e_i} = k_1 \cdot e_i, \end{cases} \tag{9}$$

where $e = |x - x'|$, denotes the distance between the reconstructed image and original image at a specific pixel during photometric consistency enforcement; $k_1 = (\sum_{i=1}^{n} e_i^2)^{-\frac{1}{2}}$, which can be considered as a constant for a specific e_i.

During training, the parameters of the network are updated in a back-propagation manner, using gradients computed with respect to the loss function. According to Eq. (9), we can draw a conclusion the $\mathcal{L}1$ norm is a fair norm that treats all points equally, aiming to obtain depth maps with low mean absolute error; while the $\mathcal{L}2$ norm concentrates more on the pixels with larger error, to reduce the outliers, *i.e.*, *terrible points*. However, these two norms cannot focus on *accurate points* directly. Therefore, we propose a photometric consistency loss based on the $\mathcal{L}0.5$ norm. The

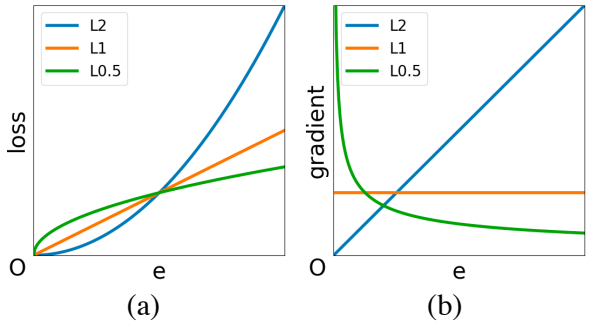


Figure 7. **Comparison of different norms.** (a) Different norm-based losses about e. (b) The gradient of different norm-based losses with regard to e. The horizontal axis e denotes the pixel-wise distance of image reconstruction with the estimated reference image depth. The $\mathcal{L}2$ norm aims to reduce outliers, the $\mathcal{L}1$ norm treats all points equally, while the $\mathcal{L}0.5$ norm concentrates more on accurate points.

$\mathcal{L}0.5$ norm and its gradient formula can be expressed as:

$$l_{0.5}(e) = \|e\|_{\frac{1}{2}} = (\sum_{i=1}^{n} e_i^{\frac{1}{2}})^2, \frac{\partial l_{0.5}(e)}{\partial e_i} = k_2 \cdot e_i^{-\frac{1}{2}}, \tag{10}$$

in which $k_2 = \sum_{i=1}^{n} e_i^{\frac{1}{2}}$. And the comparison of different norms can be seen in Fig. 7 more intuitively.

Unlike $\mathcal{L}1$-norm or $\mathcal{L}2$-norm, the $\mathcal{L}0.5$-norm pays more attention to *accurate points*. Owing to its gradient property, applying this norm to the photometric consistency loss can make *accurate points* more accurate, turn *ordinary points* into *accurate points*, and pay less attention to *terrible points*. According to the above analyses, we propose the $\mathcal{L}0.5$ photometric consistency loss as:

$$L_{0.5PC} = \sum_{i=2}^{N} \frac{\|(\hat{I}_i - I_1) \odot M_i\|_{\frac{1}{2}} + \|(\nabla \hat{I}_i - \nabla I_1) \odot M_i\|_{\frac{1}{2}}}{\|M_i\|_1}. \tag{11}$$

3.5. Overall Loss

The overall loss function of our proposed framework is constructed as follows:

$$\begin{aligned} L = &\lambda_1 L_{0.5PC} + \lambda_2 L_{ICC} + \lambda_3 L_{SCC} + \\ &\lambda_4 L_{SSIM} + \lambda_5 L_{Smooth}. \end{aligned} \tag{12}$$

L_{SSIM} [38] and L_{Smooth} [26] denote structure similarity loss and depth smooth loss respectively, which are used broadly in previous unsupervised MVS methods. The balancing weights are empirically set as λ_1 = 0.8, λ_3=0.01, λ_4= 0.2, λ_5= 0.0067, and the weight for image-level contrastive consistency is initialized with λ_2= 0.01 and doubles every two epochs. Note that the color fluctuation augmentation used by [5, 41, 42] is also applied in the image-level contrastive branch.

4. Experiment

4.1. Datasets

DTU Dataset [1] is a popular indoor benchmark comprising 79 training scans and 22 testing scans, all taken under 7 different lighting conditions. The DTU dataset provides 3D point clouds captured by structured-light sensors, with each view consisting of an image and its calibrated camera parameters. Following common practices, we perform training on the provided training dataset, while evaluation is conducted on the designated evaluation dataset.

Tanks&Temples [19] is a large-scale dataset collected under realistic lighting conditions, consisting of an intermediate subset and an advanced subset. The intermediate subset includes 8 scenes, and the advanced subset includes 6 scenes. All scenes vary in terms of scale, surface reflection, and exposure conditions. These two subsets are widely used to verify the generalization performance of MVS methods.

4.2. Implementation Details

Training. Our CL-MVSNet is trained on the DTU training set for 16 epochs in an end-to-end unsupervised learning manner. Following previous methods, the input image number N is set to 5, and images are resized and cropped to 512×640. Our backbone is similar to CasMVSNet with 3 stages, and the depth hypotheses for each stage are 48, 32, and 8 respectively. We adopt Adam [18] to optimize our model with an initial learning rate 0.0005, which is decayed by 2 after 10, 12, and 14 epochs. The network is implemented in Pytorch and trained on 8 NVIDIA Tesla V100s.

Testing. On the DTU testing set, the input image number N is set to 5 as [5, 10], and the resolution is 1184×1600 following [13, 30, 47, 49, 50]. The inferred depth maps are filtered with photometric and geometric consistencies and then fused to a point cloud following [5]. On Tanks&Temples, the input image number N is set to 7, and the resolution is 1024×1920 or 768×576. Note that our model trained on DTU is directly used to test on Tanks&Temples without finetuning on BlendedMVS [52] as [10, 41] or training on Tanks&Temples training set as [47]. The inferred depth maps are filtered with photometric and geometric consistencies and then fused to a point cloud with the same strategy as [5]. The number of depth hypotheses in the coarsest stage is set to 64, and the corresponding depth interval is set to 3 times as the interval of [50]. And only the regular branch works for testing. More details can be found in the supplementary material.

4.3. Benchmark Results on DTU

We compare the depth map evaluation results on the DTU evaluation set as shown in Tab. 1. Then we compare the quantitative reconstruction results on DTU as shown in Tab. 2, our CL-MVSNet architecture achieves the best completeness and overall score among all end-to-end unsupervised methods. It is worth noting that CL-MVSNet also outperforms its supervised counterpart CasMVSNet [13]. The qualitative results of the depth estimation shown in Fig. 8 and the 3D point cloud reconstruction shown in Fig. 9 also verify the advantages of our model.

Table 1. **Depth map evaluation results in terms of accuracy on DTU evaluation set(higher is better).** CL-MVSNet acquires the best depth estimation. All thresholds are given in millimeters.

Method	$\leq 2 \uparrow$	$\leq 4 \uparrow$	$\leq 8 \uparrow$
MVSNet [50]	0.704	0.778	0.815
Unsup MVSNet [17]	0.317	0.384	0.402
M3VSNet [15]	0.603	0.769	0.857
JDACS-MS [41]	0.553	0.705	0.786
RC-MVSNet [5]	0.730	0.795	0.863
CL-MVSNet(Ours)	**0.757**	**0.829**	**0.868**

Table 2. **Quantitative results on DTU evaluation set.** Best results in each category are in **bold**.

	Method	Acc.↓	Comp.↓	Overall↓
Supervised	SurfaceNet [16]	0.450	1.040	0.745
	MVSNet [50]	0.396	0.527	0.462
	CasMVSNet [13]	0.325	0.385	0.355
	PatchmacthNet [35]	0.427	**0.277**	0.352
	CVP-MVSNet [49]	**0.296**	0.406	0.351
	UCS-Net [7]	0.338	0.349	**0.344**
Multi-stage Self-supervised	Self-sup CVP [47]	**0.308**	0.418	0.363
	U-MVSNet [42]	0.354	0.3535	0.3537
	KD-MVS [10]	0.359	**0.295**	**0.327**
End-to-end Unsupervised	Unsup MVSNet [17]	0.881	1.073	0.977
	MVS2 [8]	0.760	0.515	0.637
	M3VSNet [15]	0.636	0.531	0.583
	DS-MVSNet [21]	**0.374**	0.347	0.361
	JDACS-MS [41]	0.398	0.318	0.358
	RC-MVSNet [5]	0.396	0.295	0.345
	CL-MVSNet(Ours)	0.375	**0.283**	**0.329**

4.4. Benchmark Results on Tanks&Temples

In line with existing methods, we test our method on the Tanks&Temples benchmark to verify the generalization ability. The quantitative results on Tanks&Temples are reported in Tab. 3. Our method achieves SOTA performance among all existing end-to-end unsupervised MVS methods. Our method also surpasses the supervised counterpart CasMVSNet [13]. The qualitative reconstruction results are visualized in Fig. 10. Fig. 11 shows a qualitative comparison of reconstruction quality with other methods. The performance on Tanks&Temples shows the generalizability and robustness of our model.

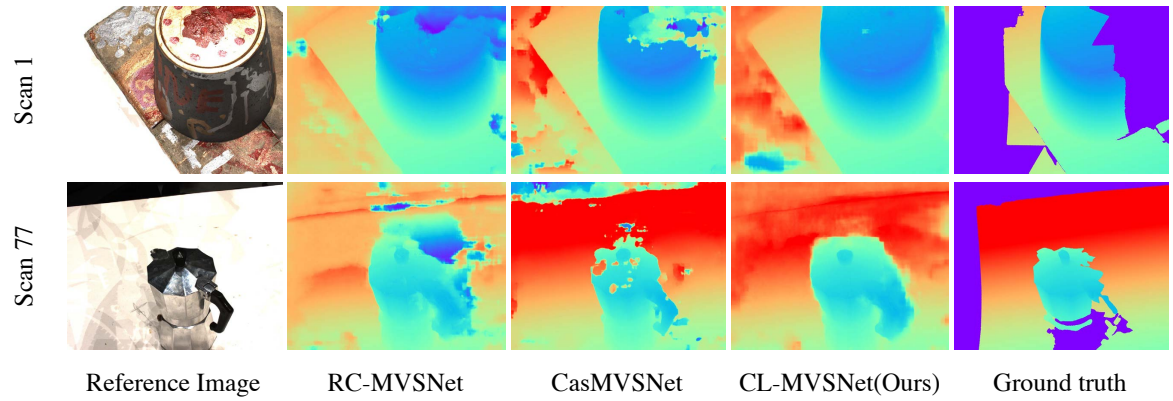


Figure 8. **Inferred depth map comparison on DTU [1].** Compared with the SOTA unsupervised method [5] and the supervised counterpart [13], CL-MVSNet gains more accurate depth maps.

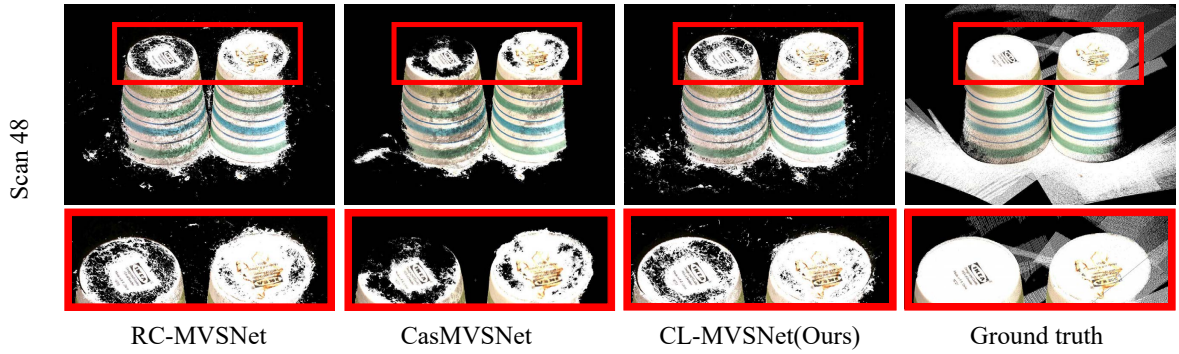


Figure 9. **Comparison of reconstructed results on scan48 of DTU benchmark [1].** CL-MVSNet reconstructs more complete results than the SOTA unsupervised method [5] and the supervised counterpart [13].

Table 3. **Quantitative results of F-score on Tanks&Temples benchmark.**

		Intermediate									Advance						
	Method	Mean↑	Fam.↑	Fra.↑	Hor.↑	Lig.↑	M60↑	Pan.↑	Pla.↑	Tra.↑	Mean↑	Aud.↑	Bal.↑	Cou.↑	Mus.↑	Pal.↑	Tem.↑
Traditional	COLMAP [31, 32]	42.14	50.41	22.25	25.63	56.43	44.83	46.97	48.53	42.04	27.24	16.02	25.23	34.70	41.51	18.05	27.94
	ACMM [44]	57.27	69.24	51.45	46.97	63.20	55.07	57.64	**60.08**	54.48	34.02	23.41	32.91	41.17	48.13	23.87	34.60
	ACMP [46]	58.41	70.30	54.06	**54.11**	61.65	54.16	57.60	58.12	57.25	37.44	**30.12**	34.68	**44.58**	50.64	27.20	37.43
	ACMMP [43]	**59.38**	**70.93**	**55.39**	51.80	**63.83**	**55.94**	**59.47**	59.51	**58.20**	**37.84**	30.05	**35.36**	44.51	**50.95**	**27.43**	**38.73**
Supervised	MVSNet [50]	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69	-	-	-	-	-	-	-
	PatchmatchNet [35]	53.15	66.99	52.64	43.24	54.87	52.87	49.54	54.21	**50.81**	**32.31**	**23.69**	37.73	**30.04**	41.80	**28.31**	**32.29**
	CVP-MVSNet [49]	54.03	**76.50**	47.74	36.34	55.12	**57.28**	**54.28**	57.43	47.54	-	-	-	-	-	-	-
	UCS-Net [7]	54.83	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89	-	-	-	-	-	-	-
	CasMVSNet [13]	**56.42**	76.36	**58.45**	**46.20**	**55.53**	56.11	54.02	**58.17**	46.56	31.12	19.81	**38.46**	29.10	**43.87**	27.36	28.11
Multi-stage Self-supervised	Self-sup CVP [47]	46.71	64.95	38.79	24.98	49.73	52.57	51.53	50.66	40.45	-	-	-	-	-	-	-
	U-MVSNet [42]	57.15	76.49	60.04	49.20	55.52	55.33	51.22	56.77	52.63	-	-	-	-	-	-	-
	KD-MVS [10]	**64.14**	**80.42**	**67.42**	**54.02**	**64.52**	**64.18**	**61.60**	**62.37**	**58.59**	**37.96**	**27.24**	**44.10**	**35.47**	**49.16**	**34.68**	**37.11**
End-to-end Unsupervised	MVS2 [8]	37.21	47.74	21.55	19.50	44.54	44.86	46.32	43.38	29.72	-	-	-	-	-	-	-
	M3VSNet [15]	37.67	47.74	24.38	18.74	44.42	43.45	44.95	47.39	30.31	-	-	-	-	-	-	-
	JDACS-MS [41]	45.48	66.62	38.25	36.11	46.12	46.66	45.25	47.69	37.16	-	-	-	-	-	-	-
	DS-MVSNet [21]	54.76	74.99	59.78	42.15	53.66	53.52	52.57	55.38	46.03	-	-	-	-	-	-	-
	RC-MVSNet [5]	55.04	75.26	53.50	45.52	53.49	54.85	52.30	56.06	49.37	30.82	21.72	37.22	28.62	37.37	27.88	**32.09**
	CL-MVSNet(Ours)	**59.39**	**76.35**	**62.37**	**49.93**	**60.02**	**57.44**	**59.97**	**56.74**	**52.28**	**37.03**	**28.07**	**43.55**	**37.47**	**50.86**	**31.45**	30.78

4.5. Ablation Study

We perform an ablation study to confirm the effectiveness of each part in our model under different configurations as shown in Tab. 4. Applying the image-level contrastive consistency loss to the model leads the model to be more context-aware for indistinguishable regions, which enhances the robustness and generalizability. Besides, the model combined with the scene-level contrastive consistency loss tends to perform better in areas of view-dependent effects, thus gaining more accurate and complete

Figure 10. **Qualitative results on Tanks&Temples [19].**

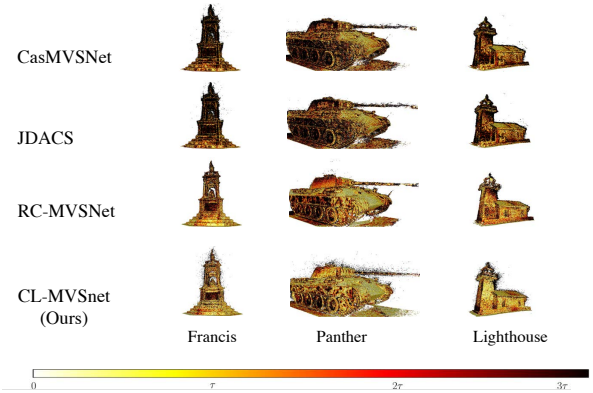


Figure 11. **Qualitative comparison of reconstruction quality on Tanks&Temples benchmark [19].** Darker regions contain larger errors. CL-MVSNet yields better performance than the SOTA unsupervised methods [5, 41] and the supervised counterpart [13].

reconstruction results. Moreover, $\mathcal{L}0.5$ photometric consistency loss further advances the CL-MVSNet, which guides to model to focus on accurate points, resulting in an improvement in the accuracy of the model. More ablation studies are included in Supplementary Materials.

4.6. Efficiency Comparison to SOTA Multi-stage Self-supervised Method

As aforementioned, multi-stage self-supervised methods cannot be trained end-to-end. Specifically, KD-MVS [10] takes several training rounds for distillation, needs complex pre-training and pre-processing before each round, uses additional dataset [52] for pseudo label generation, and requires extra storage space to save the pseudo labels. Without the above limitations, our method adopts the same backbone as theirs, but converges and achieves competitive results in just 16 epochs, with a training time of only 10 hours per epoch on one single NVIDIA Tesla V100.

5. Conclusion and Limitation

Conclusion. We have presented an effective unsupervised approach for Multi-View Stereo, termed as CL-MVSNet, which leverages dual-level contrastive learning to handle the issues of indistinguishable regions and view-dependent effects. For indistinguishable regions, we propose an image-level contrastive branch to encourage the model to take more contextual information into account. For view-dependent effects, a scene-level contrastive branch is adopted to boost the robustness. Besides, we explore an $\mathcal{L}0.5$ photometric consistency loss to emphasize the penalty of accurate points, resulting in more accurate reconstruction. We experimentally demonstrate that CL-MVSNet outperforms all SOTA end-to-end unsupervised MVS methods and the supervised counterpart on the DTU [1] and Tanks&Temples [19] benchmarks.

Table 4. **Ablation study of different components of our proposed CL-MVSNet on DTU [1].**

L_{PC}	L_{DA} [41]	L_{ICC}	L_{SCC}	$L_{0.5PC}$	Acc.↓	Comp.↓	Overall↓
✓	✓				0.422	0.334	0.378
✓		✓			0.403	0.315	0.359
✓		✓	✓		0.392	0.286	0.339
		✓	✓	✓	**0.375**	**0.283**	**0.329**

Limitation. Our model has addressed the limitations of indistinguishable regions and view-dependent effects, but the accurate depth estimation in object edge areas remains a challenge. It is worth noting that this is a common problem in unsupervised MVS methods. To mitigate this issue, we adopt an edge-aware depth smoothness loss proposed in [17], which is based on the assumption that the gradient maps of the input reference image and the inferred depth map should be similar. However, this simple assumption may be invalid in many cases. For instance, there may be significant color gradient changes within the same object.

Acknowledgements. This work is financially supported by National Natural Science Foundation of China U21B2012 and 62072013, Shenzhen Science and Technology Program-Shenzhen Cultivation of Excellent Scientific and Technological Innovation Talents project(Grant No. RCJC20200714114435057), Shenzhen Science and Technology Program-Shenzhen Hong Kong joint funding project (Grant No. SGDX20211123144400001), this work is also financially supported for Outstanding Talents Training Fund in Shenzhen.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120:153–168, 2016.

[2] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9902–9912, 2022.

[3] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Multi-view depth estimation by fusing single-view depth probability with multi-view geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2842–2851, 2022.

[4] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009.

[5] Di Chang, Aljaž Božič, Tong Zhang, Qingsong Yan, Yingcong Chen, Sabine Süsstrunk, and Matthias Nießner. Rc-mvsnet: unsupervised multi-view stereo with neural rendering. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 665–680. Springer, 2022.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020.

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

[9] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8585–8594, 2022.

[10] Yikang Ding, Qingtian Zhu, Xiangyue Liu, Wentao Yuan, Haotian Zhang, and CHi Zhang. Kd-mvs: Knowledge distillation based self-supervised learning for mvs. *arXiv preprint arXiv:2207.10425*, 2022.

[11] Haonan Dong and Jian Yao. Patchmvsnet: Patch-wise unsupervised multi-view stereo for weakly-textured surface reconstruction. *arXiv preprint arXiv:2203.02156*, 2022.

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[13] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. 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*, pages 2495–2504, 2020.

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[15] Baichuan Huang, Hongwei Yi, Can Huang, Yijia He, Jingbin Liu, and Xiao Liu. M3vsnet: Unsupervised multi-metric multi-view stereo network. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3163–3167. IEEE, 2021.

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

[17] Tejas Khot, Shubham Agrawal, Shubham Tulsiani, Christoph Mertz, Simon Lucey, and Martial Hebert. Learning unsupervised multi-view stereopsis via robust photometric consistency. *arXiv preprint arXiv:1905.02706*, 2019.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[20] Jiacheng Li, Chang Chen, and Zhiwei Xiong. Contextual outpainting with object-level contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11451–11460, 2022.

[21] Jingliang Li, Zhengda Lu, Yiqun Wang, Ying Wang, and Jun Xiao. Ds-mvsnet: Unsupervised multi-view stereo via depth synthesis. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5593–5601, 2022.

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[23] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10452–10461, 2019.

[24] Xinjun Ma, Yue Gong, Qirui Wang, Jingwei Huang, Lei Chen, and Fan Yu. Epp-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5732–5740, 2021.

[25] Zeyu Ma, Zachary Teed, and Jia Deng. Multiview stereo with cascaded epipolar raft. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 734–750. Springer, 2022.

[26] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5667–5675, 2018.

[27] Zhenxing Mi, Chang Di, and Dan Xu. Generalized binary search network for highly-efficient multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12991–13000, 2022.

[28] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717, 2020.

[29] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8645–8654, 2022.

[30] Ke Qiu, Yawen Lai, Shiyi Liu, and Ronggang Wang. Self-supervised multi-view stereo via inter and intra network pseudo depth. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2305–2313, 2022.

[31] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[32] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016.

[33] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, pages 519–528. IEEE, 2006.

[34] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. Itermvs: iterative probability estimation for efficient multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8606–8615, 2022.

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

[36] Likang Wang, Yue Gong, Xinjun Ma, Qirui Wang, Kaixuan Zhou, and Lei Chen. Is-mvsnet: Importance sampling-based mvsnet. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 668–683. Springer, 2022.

[37] Xiaofeng Wang, Zheng Zhu, Guan Huang, Fangbo Qin, Yun Ye, Yijia He, Xu Chi, and Xingang Wang. Mvster: epipolar transformer for efficient multi-view stereo. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 573–591. Springer, 2022.

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[39] Junhua Xi, Yifei Shi, Yijie Wang, Yulan Guo, and Kai Xu. Raymvsnet: Learning ray-based 1d implicit fields for accurate multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8595–8605, 2022.

[40] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021.

[41] Hongbin Xu, Zhipeng Zhou, Yu Qiao, Wenxiong Kang, and Qiuxia Wu. Self-supervised multi-view stereo via effective co-segmentation and data-augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3030–3038, 2021.

[42] Hongbin Xu, Zhipeng Zhou, Yali Wang, Wenxiong Kang, Baigui Sun, Hao Li, and Yu Qiao. Digging into uncertainty in self-supervised multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6078–6087, 2021.

[43] Qingshan Xu, Weihang Kong, Wenbing Tao, and Marc Pollefeys. Multi-scale geometric consistency guided and planar prior assisted multi-view stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4945–4963, 2022.

[44] Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5483–5492, 2019.

[45] Qingshan Xu and Wenbing Tao. Learning inverse depth regression for multi-view stereo with correlation cost volume. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12508–12515, 2020.

[46] Qingshan Xu and Wenbing Tao. Planar prior assisted patchmatch multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12516–12523, 2020.

[47] Jiayu Yang, Jose M Alvarez, and Miaomiao Liu. Self-supervised learning of depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7526–7534, 2021.

[48] Jiayu Yang, Jose M Alvarez, and Miaomiao Liu. Non-parametric depth distribution modelling based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8626–8634, 2022.

[49] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4877–4886, 2020.

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

[51] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5525–5534, 2019.

[52] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020.

[53] Jinzhi Zhang, Ruofan Tang, Zheng Cao, Jing Xiao, Ruqi Huang, and Lu Fang. Elasticmvs: Learning elastic part representation for self-supervised multi-view stereopsis. In *Advances in Neural Information Processing Systems*.

[54] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. *arXiv preprint arXiv:2008.07928*, 2020.

[55] Rui Zhu, Bingchen Zhao, Jingen Liu, Zhenglong Sun, and Chang Wen Chen. Improving contrastive learning by visualizing feature transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10306–10315, 2021.