Pacific Graphics 2019
C. Theobalt, J. Lee, and G. Wetzstein
(Guest Editors)

Volume 38 (2019), Number 7

Pyramid Multi-View Stereo with Local Consistency

Jie Liao[†1] Yanping Fu[1] Qingan Yan[2] Chunxia Xiao[1]

[1]School of Computer Science, Wuhan University, China
[2]JD.com

Abstract

In this paper, we propose a PatchMatch-based Multi-View Stereo (MVS) algorithm which can efficiently estimate geometry for the textureless area. Conventional PatchMatch-based MVS algorithms estimate depth and normal hypotheses mainly by optimizing photometric consistency metrics between patch in the reference image and its projection on other images. The photometric consistency works well in textured regions but can not discriminate textureless regions, which makes geometry estimation for textureless regions hard work. To address this issue, we introduce the local consistency. Based on the assumption that neighboring pixels with similar colors likely belong to the same surface and share approximate depth-normal values, local consistency guides the depth and normal estimation with geometry from neighboring pixels with similar colors. To fasten the convergence of pixelwise local consistency across the image, we further introduce a pyramid architecture similar to previous work which can also provide coarse estimation at upper levels. We validate the effectiveness of our method on the ETH3D benchmark and Tanks and Temples benchmark. Results show that our method outperforms the state-of-the-art.

CCS Concepts

• ***Computing methodologies*** → ***Computer graphics; Point-based models;***

1. Introduction

Given a set of images that are manually calibrated or calibrated by Structure-From-Motion algorithms [SSS06, AFS*09, SF16, YYL*16, YYZX17], Multi-View Stereo (MVS) recovers a dense 3D representation of the target scene. The reconstruction results are the key ingredients of automatic geometry, scene classification, image-based modeling and robot navigation. Thanks to the publishing of 3D reconstruction benchmark [SvHV*08, SSG*17, KPZK17], the reconstruction results of MVS algorithms can be quantitatively and effectively evaluated. This facilitates the design of MVS algorithms and boosts vigorous progress in the field.

Among them, the PatchMatch-based algorithms are currently the top-performing approaches for robust and accurate 3D reconstruction. Although [YLL*18, HMK*18, YLL*18, YLL*19] have exploited a new direction of MVS algorithms based on deep learning, these algorithms rely on the training datasets. If no similar scenes have appeared in the training datasets, the target scenes can not be entirely and accurately reconstructed by machine learning MVS algorithms.

The core and most challenging procedure of PatchMatch-based MVS is the estimation of depth maps, which are subsequently fused into a point cloud and thus make great sense in the quality of reconstruction results. To estimate accurate and dense depth maps, PatchMatch-based MVS employs PatchMatch to perform pixelwise dense matching. PatchMatch was first proposed by Barnes *et al.* [BSFG09] to match pixels between two images. As a single-pixel contains hardly enough information to support robust matching, a window centered at the pixel is applied to collect neighboring information which is known as the patch in PatchMatch algorithms. The core idea of PatchMatch is to randomly initialize the matching relationship like translation between patches across two images and iteratively propagate good matches to surrounding areas. Given calibrated images, Bleyer *et al.* [BR11] extended the 2D matching relationship like translation and scale to 3D via epipolar geometry and homography projection. Schönberger *et al.* [SZPF16] furthered this work and published COLMAP, which could jointly estimate the depth and normal and perform pixel-wise view selection.

One fundamental building block of PatchMatch is the similarity measurement between patches, which is utilized to judge the matching across different images. The most commonly used measures are the Sum of Squared Difference (SSD), Normalized Cross Correlation (NCC), and bilaterally weighted NCC. As SSD is sensitive to capturing conditions like color balance, it is hardly adopted by MVS algorithms. NCC and bilaterally weighted NCC measure the structural similarity between patches and are hence robust to some capturing variations. However, they can not measure the sim-

[†] This work was partly supported by The National Key Research and Development Program of China (2017YFB1002600), the NSFC (No. 61672390), Wuhan Science and Technology Plan Project (No. 2017010201010109), and Key Technological Innovation Projects of Hubei Province (2018AAA062). Chunxia Xiao is the corresponding author of this paper.

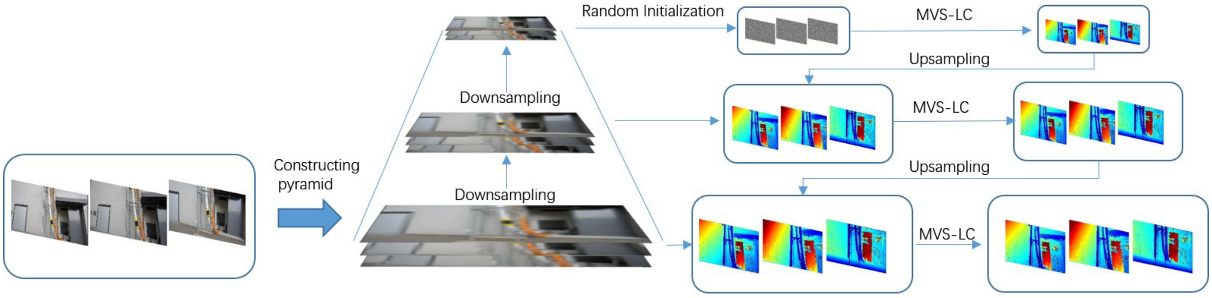


Figure 1: *The overview of our algorithm. MVS-LC is our algorithm integrating local consistency to COLMAP. At the top level of the pyramid architecture, depth maps are first initialized with random values. Then they are optimized using MVS-LC. Depth maps estimated from the upper level are upsampled and passed to current level as the initial depth maps. To avoid depth hypotheses from upper level stuck by geometric consistency, we assign patches that are inconsistent in photometry with projections in other visible images with random depth values as described in Section 4.3.*

ilarity when one patch is textureless as NCC will be meaningless according to the definition. Therefore, the depth of the textureless regions is hard to be defined with enough confidence since the photometric measurement alone hardly discerns neighboring regions. To overcome this problem, Romanoni *et al.* [RM19] assumed that textureless areas were often piecewise flat and fit one plane for each textureless superpixel of the input images thus to iteratively estimate the depth and normal of pixels inside the superpixel. The drawback of their method is that the assumption does not suit for textureless curve surface. Xu *et al.* [XT19] utilized downsampled images and median filter to estimate the coarse depth values and employed geometric consistency to guide the propagation of depth hypotheses to the higher resolution. However, some details may be smoothed.

In this paper, we propose a hierarchical MVS algorithm which can effectively estimate the depth and normal in textureless regions while keeping the detailed structure. The key insight is that neighboring pixels with similar colors likely belong to the same surface and share approximate depth-normal values. The main contributions of our paper are summarized as follows:

- We modified the photometric consistency measurement to make it suitable for patches with homogeneous texture. COLMAP utilizes bilaterally weighted NCC to measure the photometric consistency between two patches, and when at least one of the patches to be measured is color homogeneous the photometric consistency metric is set to -1, which is intuitively unreasonable. According to the number of textureless patches, we adopt different assignment strategies for photometric consistency metric. Compared to COLMAP, photometric consistency values computed by our method is more reliable, which is of significance for the view selection procedure.
- We introduced the local consistency to guide geometry estimation for textureless regions. By applying local consistency in the depth and normal optimization procedure, depth and normal values belonging to neighboring pixels are taken into consideration and are weighted by their current photometric-geometric consistency and their color similarity with current pixel. Besides, we replace the constant transition probability of view selection leveraged in [SZPF16] and [ZDJF14] with a variable based on the color similarity between neighboring pixels. The new transition probability will favor coincident view selection for neighboring pixels with similar colors.

We demonstrate our method on the ETH3D benchmark and Tanks and Temples benchmark and name it as PLC. Results show that our method outperforms the state-of-the-art methods.

2. Related Work

In the last decade, we have seen vigorous progress in MVS algorithms, and it can be observed that PatchMatch-based MVS algorithms are currently the top-performing approaches according to benchmarks [SSG*17,JDV*14,SvHV*08,KPZK17]. Since our approach is also PatchMatch-based, we limit our discussion henceforth to PatchMatch-based MVS in this section.

The PatchMatch seminal paper by Barnes *et al.* [BSFG09] proposed a randomized framework to quickly find dense approximate nearest neighbor matches between images via random initialization and propagation of good matches to surrounding areas. Hereafter Barnes *et al.* [BSGF10] generalized the original PatchMatch in matching measurements, searching domains and number of nearest neighbors. HaCohen *et al.* [HSGL11] interleaved [BSGF10] with fitting a global non-linear parametric color model and aggregating consistent matching regions using locally adaptive constraints to address dense matching under different lightning and non-rigid transformations.

Although the works above-mentioned have achieved robust, dense and pixel-wise correspondence estimation between images, their results can not be applied directly to the matching procedure of MVS. Since their works build dense correspondence fields only in perspective of two dimensions, the mapping relationship is limited to similarity transformation which leads to the accuracy of estimated correspondence hardly reaching the requirement for 3D reconstruction. Notable attempt for applying the idea of PatchMatch to stereo matching is [BR11] proposed by Bleyer *et al.*, which alternated the fixed-sized square window with slanted support window onto which the support region was projected. Several variants of this algorithm have been proposed like PMBP by Besse *et al.* [BRFK14] and PM-Huber by Heise *et al.* [HKJK13], which introduced explicit regularization based on [BR11] and achieved smoother depth estimation while preserving edge discontinuities. Yan *et al.* [YYZX17].

Previous works successfully integrate the idea of PatchMatch into pairwise stereo matching. The first PatchMatch-based Multi-View Stereo was proposed by Shen [She13]. By applying a simplified method of Bleyer *et al.* [BR11] to a subset of image pairs which are chosen according to shared points computed by Structure from Motion and mutual parallax angle, their method estimates a set of depth maps. Subsequently, these depth maps are refined according to geometric consistency across multiple views and fused into a point cloud. Galliani *et al.* [GLS15] modified the propagation scheme of PatchMatch so that it can be massively parallelized on GPU. Differently from Shen [She13], for each reference image Galliani *et al.* [GLS15] selected a subset of source images according to geometric priors for depth estimation. The drawback of these two works is that their view selection is decoupled from geometry estimation and is performed for the whole reference image but not for each pixel.

Zheng *et al.* [ZDJF14] jointly performed depth estimation and pixel-wise view selection by formulating them into a Hidden Markov Chain. They applied a generalized Expectation-Maximization method to alternatively update depth estimation and view selection while keeping the other fixed. Schönberger *et al.* [SZPF16] modified [ZDJF14] by jointly estimating depths and normals which enable hypotheses for slanted surface and utilizing geometric priors to view selection for higher accuracy. Although Zheng *et al.* [ZDJF14] and Schönberger *et al.* [SZPF16] have made a great contribution to MVS, there are still limitations in their works. NCC or bilaterally weighted NCC which they utilized to measure photometric consistency between patches from different images can not discriminate textureless regions, leading to poor reconstruction results in textureless regions. To address this issue, Romanoni *et al.* [RM19] assumed that textureless regions are piece-wise flat and fitted a plane for each color-homogeneous superpixel segmented from the reference image. This method effectively estimates geometric hypotheses for planar-like surfaces but is not suitable for curved surfaces.

3. Review of the COLMAP Framework

In this section we review the state-of-the-art MVS framework proposed by Schönberger *et al.* [SZPF16] to introduce notations and context for our contributions. Since the framework sweeps every single line independently in four directions for parallel computational tractability, without loss of generality, we only focus on one swept sequence and denote the coordinate of the pixel as value l. Given the reference image $\mathbf{X}^{ref}$ and source images $\mathbf{X}^{src} = \{X^m | m = 1 \dots M\}$, the framework models the sequential depth θ_l and normal $\mathbf{n}_l$ as a Markov process where the unobserved states correspond to binary indicator variables $Z_l^m \in \{0, 1\}$, which indicates whether pixel l is visible in source image m. Then the inference is formulated as a Maximum-A Posterior (MAP) estimation where the posterior probability is:

$$P(\mathbf{Z}, \theta, \mathbf{N} | \mathbf{X}) = \frac{P(\mathbf{Z}, \theta, \mathbf{N}, \mathbf{X})}{P(\mathbf{X})}$$
$$= \frac{1}{P(\mathbf{X})} \prod_{l=1}^{L} \prod_{m=1}^{M} [P(Z_{l,t}^m | Z_{l-1,t}^m, Z_{l,t-1}^m) \tag{1}$$
$$P(\theta_l, \mathbf{n}_l | \theta_l^m, \mathbf{n}_l^m) P(X_l^m | \theta_l, \mathbf{n}_l, Z_l^m)].$$

L is the number of pixels in considered line sweep, $\mathbf{X} = \{\mathbf{X}^{ref}, \mathbf{X}^{src}\}$, $\theta = \{\theta_l | l = 1 \dots L\}$ and $\mathbf{N} = \{\mathbf{n}_l | l = 1 \dots L\}$. The likelihood term

$$P(X_l^m | \theta_l, \mathbf{n}_l, Z_l^m) = \begin{cases} \frac{1}{NA} exp(-\frac{(1-\rho_l^m(\theta_l, \mathbf{n}_l))^2}{2\sigma_\rho^2}) & \text{if } Z_l^m = 1 \\ \frac{1}{N} \mathcal{U} & \text{if } Z_l^m = 0 \end{cases}, \tag{2}$$

represents the occurrence probability of photometric consistency between patch X_l^{ref} centered at pixel l in X^{ref} and the corresponding projection of the patch X_l^m on non-occluded source image X^m. The photometric consistency ρ is computed as bilaterally weighted NCC based on color and planar Euclidean distance, $A = \int_{-1}^{1} exp(-\frac{(1-\rho)^2}{2\sigma_\rho^2}) d\rho$ where σ_ρ is a constant, and N is a constant cancelling out the optimization. $\mathcal{U}$ is the uniform distribution in range $[-1, 1]$ with probability density 0.5. The geometric consistency term $P(\theta_l, \mathbf{n}_l | \theta_l^m, \mathbf{n}_l^m)$ enforces multi-view consistent depth and normal estimates. The spatial and temporal smoothness term $P(Z_{l,t}^m | Z_{l-1,t}^m, Z_{l,t-1}^m)$ enforces spatially smooth occlusion maps with reduced temporal oscillation during the optimization.

Zheng *et al.* [ZDJF14] proposed to use variational inference to solve the computational infeasible Equation 1 and Schönberger *et al.* [SZPF16] modified this work and approximated the posterior probability with a function $q(\mathbf{Z}, \theta, \mathbf{N})$ which minimizes the KL-Divergence with Equation 1. The function $q(\mathbf{Z}, \theta, \mathbf{N})$ is assumed to be factorizable into $q(\mathbf{Z})q(\theta, \mathbf{N})$. To estimate the approximation, they proposed a variation of the Generalized Expectation-Maximization algorithm [NH98]. For tractability, the function $q(\theta, \mathbf{N})$ is constrained to the family of Kronecker delta functions $q(\theta, \mathbf{N}) = \delta(\theta = \theta^*, \mathbf{N} = \mathbf{N}^*)$ where θ^* and $\mathbf{N}^*$ are the depth and normal values to be estimated. In the E step, the functions $q(\theta, \mathbf{n})$ are kept fixed and $q(\mathbf{Z})$ is calculated via forward-backward algorithm through the Hidden Markov Model. In the M step, $q(\mathbf{Z})$ is fixed and $(\theta_l, \mathbf{n}_l)$ is optimized as:

$$\left(\hat{\theta}_l^{\text{opt}}, \hat{n}_l^{\text{opt}}\right) = \underset{\theta_i^*, n_i^*}{\operatorname{argmin}} \frac{1}{|S|} \sum_{m \in S} \xi_l^m \left(\theta_l^*, n_l^*\right), \tag{3}$$

$$\xi_l^m \left(\theta_l^*, n_l^*\right) = (1 - \rho_l^m(\theta_l^*, \mathbf{n}_l^*)) + \eta \min\left(\psi_l^m, \psi_{\max}\right), \tag{4}$$

where S is the subset of the source images which are selected according to the probability $P(Z_l^m = 1)$ that favors pixel l is visible in image X^m and coherent with three geometric priors which encourages wide baseline, similar resolutions and front-facing patch. $\xi_l^m\left(\theta_l^*, n_l^*\right)$ represents the combination of photometric and geometric discrepancy. η here is set to 0.5 as a constant regularizer. $\psi_l^m = \|x_l - H_l^m H_l x_l\|$ is the forward-backward reprojection error, where H_l and H_l^m denotes relatively the transformation from the reference image to the source image and from the source to the reference image. $\psi^{max} = 3px$ is the maximum reprojection error.

4. Algorithm

Given camera parameters, 3D reconstruction boils down to a matching problem under the constraints of epipolar geometry. PatchMatch-based MVS employs a patch as the proxy of a pixel for 3D matching. By measuring the photometric consistency between the patch in the reference image and its corresponding projection on the other source image, PatchMatch-based MVS justify which depth and normal values are better hypotheses. Mainstream PatchMatch-based MVS algorithms adopt NCC or bilaterally weighted NCC as the measurement, which can effectively measure the structural similarity between patches and contribute a lot to the hypothesis estimation in textured regions. However, those photometric measurements are unreliable in textureless regions. Particularly, NCC and bilaterally weighted NCC can not distinguish patches with homogeneous colors. Besides, NCC and bilaterally weighted NCC are meaningless according to the definitions when one of the patches to be measured is color homogeneous.

As the view selection procedure is tightly related to the photometric consistency metrics, we first modify the photometric consistency measurement utilized in COLMAP to make patches with homogeneous colors visible in the correct source images. Intuitively, pixels in textureless regions likely belong to the same surface. Based on this observation, we proposed the local consistency to favor approximate depth and normal hypotheses for neighboring pixels with similar colors. The depth and normal optimization for pixels in textureless regions will be pixel-wisely guided by surrounding geometry according to local consistency with the geometry of textured regions as the boundary conditions. This procedure is demonstrated in Figure 2. To facilitate the convergence of local consistency, we further introduce the pyramid architecture. In the remainder of this section, we describe in turn the modification of photometric consistency measurement, local consistency and the pyramid architecture.

4.1. Photometric Consistency Measurement Modification

In the COLMAP framework, bilaterally weighted NCC is always set to -1 once the reference patch is color homogeneous. This will result in the photometric consistency metrics for textureless patches to be assigned with the lowest value and can not discriminate with different depth and normal hypotheses. Although our proposed local consistency can contribute to picking up the optimal depth and normal hypotheses for a textureless patch, the corresponding view selection probability $P(Z_l^m = 1)$ is computed according to photometric consistency metric. Unconditional minimum photometric consistency metrics for patches with homogeneous colors are unreasonable and will disturb the selection of source images S in Equation 8. The consequence is that incorrect source image, for example, images where pixel l in reference image is occluded may be selected for depth and normal estimation of pixel l. To avoid this situation, we modified the photometric consistency measurement. Denoting h as the number of textureless patches in X_l^{ref} and X_l^m. The photometric consistency is computed as:

$$\rho_l^m = \begin{cases} g & \text{if } h = 0 \\ -1 & \text{if } h = 1 \\ -1 & \text{if } h = 2 \text{ and } |c_l^m - c_l| > 3\sigma_c \\ \eta + 0.1 & \text{if } h = 2 \text{ and } |c_l^m - c_l| \le 3\sigma_c \end{cases}, \tag{5}$$

where g is calculated using bilaterally weighted NCC. c_l and c_l^m are the color values of patch l and its corresponding projection on image m if they are both color homogeneous. σ_c is a constant usually set as 0.05. η is calculated according to Equation 2 which satisfies

$$\frac{1}{NA} exp(-\frac{(1-\eta)}{2\sigma_\rho^2}) = \frac{1}{2N}, \tag{6}$$

$$\eta = 1 - \sigma_\rho \sqrt{-2\ln(\frac{A}{2})}. \tag{7}$$

For the case that both patches to be matched are textured, we still use bilaterally weighted NCC to calculate the photometric consistency. Intuitively, the probability that a textured patch in one image is captured in the other image as a textureless patch is very low, so our strategy sets $\rho = -1$ which is the lower boundary of NCC in this situation. When both the patch l and its projection on image m is color homogeneous, ideally the more their colors are similar, the higher the probability that they are relative is. However, this assumption is unreliable as variable capturing conditions may make the same surface show different colors in images. So we only make a rough assumption that when the color difference between reference patch and projection patch surpasses a certain threshold, the two patches are unlikely to be related, and vice versa. η is the value of ρ in Equation 2 that makes $P(X_l^m|\theta_l, \mathbf{n}_l, Z_l^m)$ is equal in cases that $Z_l^m = 0$ and $Z_l^m = 1$. In this way, if the colors of X_l^{ref} and X_l^m are similar, the photometric consistency metric $\eta + 0.1$ favors a little more that X_l^{ref} is visible in source image m.

4.2. Local Consistency

To estimate the optimal depth and normal values, COLMAP framework takes both the photometric and geometric consistency into consideration. According to photometric consistency, depth and normal values are optimized so that the appearances of the patch to be matched and its projection on other visible images are similar, which is intuitive and contributes to precise estimation for regions with textures. Geometric consistency measures the distance between the patch center and its forward-backward re-projection with the other visible source image. By optimizing the geometric consistency, COLMAP ensures depth estimation from different views is consistent with each other, which will further refine the hypotheses estimated according to photometric consistency. However, it is still a hard work to estimate depth and normal values for textureless regions only considering photometric and geometric consistency, since estimation according to photometry fails in textureless regions and hence can not afford effective view selection and hypotheses information for geometric consistency. To better address this issue, we propose the local consistency. Similar to [RM19], we also think neighboring pixels with similar colors belong to the same surface and there exists a geometric relationship

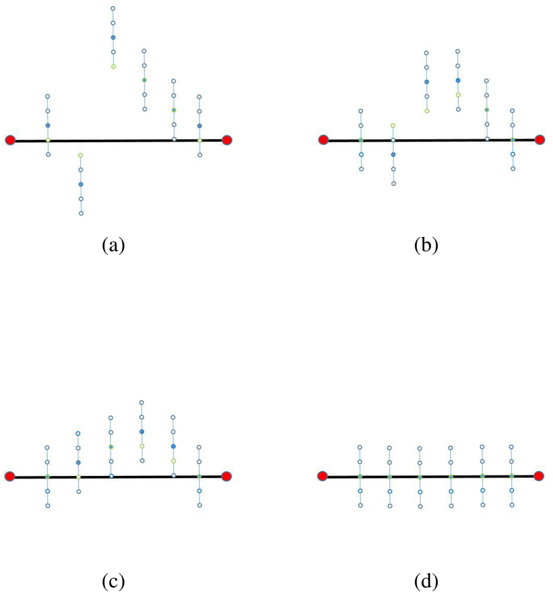


Figure 2: *The depth optimization procedure with local consistency. The black line represents the target scene surface. Two red circles represent the reliable hypotheses with high photometric and geometric consistency. Each circle sequence represents the depth candidates with the solid circle as the current depth and circle with green outline as the optimal depth coincide with local consistency. (a), (b) and (c) are sequential states of depth estimation and (d) is the final stable state.*

for pixels inside textureless regions. The difference is that [RM19] constrains textureless regions to be piecewise flat while we only assume adjacent pixels with similar colors have approximate depth and normal values.

In the COLMAP framework, photometric and geometric consistency are mainly considered in the M step of the variational inference, which chooses the most suitable depth and normal values given fixed view selection probabilities. This procedure is formulated as Equation 3. We further integrate our local consistency in this procedure and formulate it as:

$$\left(\hat{\theta}_l^{\text{opt}}, \hat{n}_l^{\text{opt}}\right) = \underset{\theta_i^*, n_i^*}{\text{argmin}} \frac{1}{|S|} \sum_{m \in S} \zeta_l^m \left(\theta_l^*, n_l^*\right), \tag{8}$$

$$\begin{aligned} \zeta_l^m \left(\theta_l^*, n_l^*\right) =& \lambda \xi_l^m \left(\theta_l^*, n_l^*\right) + \\ & \frac{1-\lambda}{|I|} \sum_{i \in I} -\xi_i^m(\theta_i, \mathbf{n}_i) \varphi_{il} (d_{il}^{\theta} + d_{il}^{\mathbf{n}} + d_{il}^{g}), \end{aligned} \tag{9}$$

where λ is a constant regularizer balancing local consistency with geometric and photometric consistency, I is the subset of adjacent pixels around pixel l. $d_{il}^{\mathbf{n}} = \|\mathbf{n_i} - \mathbf{n_l}^*\|$ measures the difference between normalized normal vectors. $d_{il}^{\theta} = min(|\theta_i - \theta_l|/\theta_{max}, 1)$ where θ_{max} is a constant measures the relative depth difference between pixel i and pixel l. d_{il}^{g} favors the normal vectors calculated according to photometry and local geometry to be consistent, which is formulated as:

$$d_{il}^{g} = \|\mathbf{n_l}^* (\theta_i K^{-1} x_i - \theta_l K^{-1} x_l)\|, \tag{10}$$

where K is the calibration of the reference image and x_i and x_l denote the coordinates of pixel i and l in the reference image. $\varphi_{il} = exp(-\frac{(c_l - c_i)^2}{\sigma_c^2})$ where σ_c is the same variable adopted in Equation 5. φ_{il} assigns high weights to neighboring pixels whose color is similar to pixel l. And $-\xi_i^m(\theta_i, \mathbf{n}_i)$ assigns high weights to neighboring pixels whose geometric and photometric consistency is high.

In addition, we also apply our local consistency assumption to the estimation of view selection probability $P(Z_l^m)$ where Z_l^m is deemed as the hidden variable in Hidden Markov Chain. In [SZPF16] and [ZDJF14], the transition probability of Z is formulated as

$$P\left(Z_l^m | Z_{l-1}^m\right) = \begin{pmatrix} \gamma & 1-\gamma \\ 1-\gamma & \gamma \end{pmatrix}, \tag{11}$$

where γ is a preset constant (usually set to 0.999), which means that the transition probability of view selection between all adjacent pixels is the same despite the fact that they may belong to different objects. Here we define γ as a variable which favors view selection smoothness between pixels with similar colors and formulate it as

$$\gamma = \mu \varphi_{l(l-1)}, \tag{12}$$

where μ is a constant experimentally set as 0.999. We introduce μ here to avoid $\gamma = 1$ which will lead to over smoothness of view selection only based on color similarity but regardless of photometric consistency.

4.3. Pyramid Architecture for MVS

By applying local consistency in the M step of variational inference, different depth and normal hypotheses of the textureless patch can be discriminated according to neighboring conditions. However, for high-resolution images which are very ubiquitously used now, to estimate the accurate depth and normal hypotheses for textureless regions which occupy a large number of pixels via only local consistency in the PatchMatch framework is hard work as it needs vast iterations to make the local consistency convergence. Therefore we introduce the pyramid architecture similar to the previous work [XT19] here to utilize the multi-scale texture information and facilitate the procedure. The overview of our pyramid architecture for MVS is illustrated in Figure 1.

We first construct a κ-scale pyramid for all input images with a downsampling factor ι (commonly set as 0.5). For the first level of the architecture, we first randomly initialize depth and normal hypotheses for each pixel which is the same with [SZPF16] and [ZDJF14]. Then we perform depth and normal estimation through PatchMatch framework with only local and photometric consistency taken into consideration, which is followed by estimation with photometric, geometric and local consistency. The reason why we ignore geometric consistency in the beginning is that for randomized depth values the forward-backward re-projection errors easily beyond the threshold ψ^{max} and hence beyond the geometric consistency constraint.

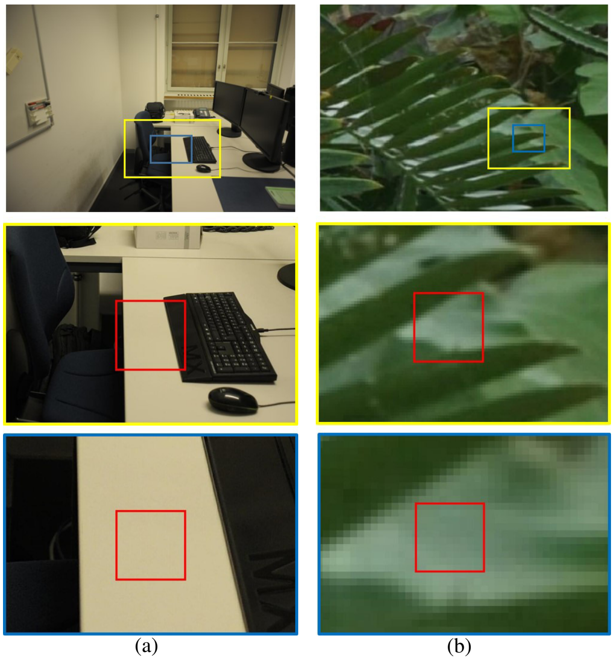


Figure 3: *With increasing resolution, some patches centered at the same position start to lose texture as their capturing area decreases. The first row demonstrates the source images. The colored rectangles in source images show the corresponding regions for images outlined in the same color in the second and third rows. The second and third rows demonstrate the areas (in red rectangles) captured by patches centered at the same position in images with different resolutions. The images in the second row are with higher resolution than the images in the third row and are in the upper level of the pyramid architecture. For case (a) depth hypothesis propagated from the upper level is reliable as the primary geometry in the patch is nearly the same. For case (b), depth hypothesis from the upper level is unreliable for the current level as the geometry changes.*

Subsequently, the following step is performed repetitively level by level until the bottom of the pyramid architecture. Hypotheses estimated by the upper level is passed down to the current level as the initialization. For increased resolution, normal maps are upsampled through joint bilateral upsampler [KCLU07] while depth values are upsampled by intersecting the back projection ray with the adjacent patch in 3D space [WYL*19]. As the proportion of a patch in the image in the current level is reduced compared with that in the upper level, some patches start to acquire no texture as illustrated in Figure 3. Among these patches, some still belong to the same surface that corresponding patches from upper level approximate as shown in Figure 3(a). In this situation, the propagated hypotheses are reliable. Some may not belong to the same surface with neighboring pixels but are deemed as the texture of neighboring surface in the upper level as shown in Figure 3(b) and the corresponding propagated depth and normal values are incorrect. To distinguish these two cases, we compare the appearance of the patch X_l^{ref} with the appearance of its projections on selected source images S and count the number of images n_l^c on which the appearance of the projection is in correspondence with that of the patch from the reference image. For image $m \in S$, we define that if X_l^m is also color homogeneous and satisfies $\|c_l^{ref} - c_l^m\| < c^{max}$ where c_l^{ref} and c_l^m are the color values of X_l^{ref} and X_l^m and c^{max} is the preset threshold, then X_l^m is coincident with X_l^{ref} and n_l^c is incremented by one. If finally $n_l^c \leq 5$ then patch l is thought to be propagated with wrong hypotheses and we assign θ_l and $\mathbf{n}_l$ with random values to avoid θ_l and $\mathbf{n}_l$ are stuck by local and geometric consistency.

5. Experiments

In this section, we first evaluate and compare our method with the state-of-the-art MVS methods [SZPF16, KHSM17, YLL*18, YLL*19, XT19, RM19] on the ETH3D benchmark and Tanks and Temples benchmark. We then perform the ablation study based on the *office* dataset provided the ETH3D benchmark. Finally, we analyze the limitations of our method and demonstrate the failure works. All the experiments in this section are conducted on a single machine with an Intel Xeon(R) CPU E5-2630, 64G RAM and GeForce RTX 2080 Ti.

5.1. ETH3D Benchmark

For evaluation of MVS algorithms, the ETH3D benchmark provides both high resolution and low-resolution datasets, which are further classified into training dataset and test dataset. Ground-truth (GT) is only provided for training dataset, which allows parameter tuning. Reconstruction results submitted to the ETH3D benchmark will be evaluated in three aspects as completeness, accuracy and F_1 score. The completeness is computed as the percentage of points from GT which are within a certain distance τ from the model. The accuracy is computed as the percentage of points from the model which are within a distance τ from the GT. The F_1 score is the harmonic average of completeness and accuracy. For a detailed description of the ETH3D benchmark, it is suggested to refer to [SSG*17].

We set the level of pyramid architecture $\kappa = 5$, the constant regularizer $\lambda = 0.97$ and $\sigma_c = 0.05$. θ_{max} is assigned with the max depth value acquired from the sparse reconstruction. All other parameters are the same as the default of COLMAP. We use the fusion method provided by COLMAP to fuse depth maps into point clouds.

Table 5 shows the F_1 scores of our proposed method against published state-of-the-art MVS algorithms on test datasets with thresholds of $5cm$ and $10cm$. It can be observed that PLC ranks first for several scenarios. For the threshold of $5cm$, our method ranks the second over all test datasets with only 0.09 points off the first. For the threshold of $10cm$, our method ranks the first over high-resolution test datasets and over all test datasets. It should be noticed that PLC significantly improve the performance of base framework COLMAP. The improvement is attributed to the modification of photometric consistency on textureless regions and local consistency constraint which can effectively discriminate depth and normal hypotheses for textureless regions according to geometry and consistency cost of neighboring pixels.

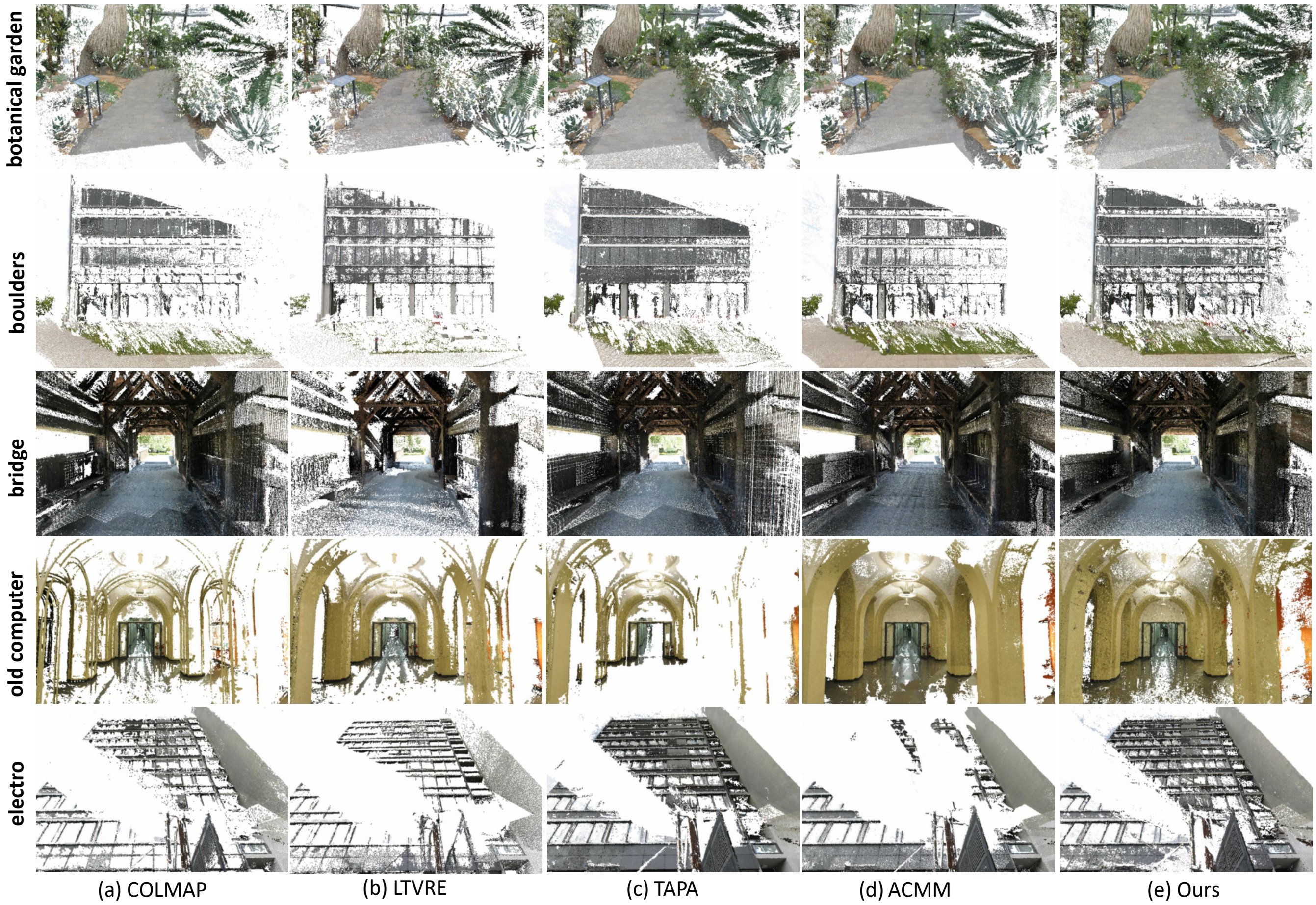


Figure 4: *Qualitative point cloud comparisons between different algorithms on some high-resolution multi-view test datasets of ETH3D benchmark.*

Figure 5 shows the completeness evaluation on high resolution training dataset. It can be observed that the overall visual completeness of our results for these scenarios is better than others.

Figure 4 exhibits qualitative point cloud comparisons. Compared to TAPA-MVS, our method and ACMM perform better for the reconstruction of textureless curve regions as shown in the *old computer* dataset of Figure 4, where the assumption made by [RM19] that textureless regions are piece-wise planar fails.

Figure 6 shows the qualitative comparison of the reconstructed details between ACMM, TAPA-MVS and our method. PLC reconstructed more details without distorting the structures. The arch of the window and the edge of the roof in *observatory* dataset and door in *old computer* dataset were distorted by ACMM while delicately reconstructed by PLC. We infer that the distortion is mainly caused by median filter adopted by ACMM which is intended to filter out outlier depth estimates but will also smooth and distort inlier depth estimates.

5.2. Tanks and Temples Benchmark

The Tanks and Temples benchmark presents both training data and testing data. The testing datasets are organized into two groups: intermediate and advanced. The intermediate group contains sculptures, large vehicles, and house-scale buildings with outside-looking-in camera trajectories. The advanced group contains large indoor scenes imaged from within and large outdoor scenes with complex geometric layouts and camera trajectories. Reconstruction results submitted to the Tanks and Temples benchmark are evaluated on recall, precision and F-score. The definitions of recall, precision and F-score are the same as that of completeness, accuracy and F_1 score in the ETH3D benchmark as described in Section 5.1. For a detailed description of the Tanks and Temples benchmark, it is suggested to refer to [KPZK17].

We set the level of pyramid architecture $\kappa = 4$, the constant regularizer $\lambda = 0.99$ and $\sigma_c = 0.05$. θ_{max} is assigned with the max depth value acquired from the sparse reconstruction. We use the fusion method provided by COLMAP to fuse depth maps into point clouds and set the maximum re-projection error as 0.5 in this procedure. All other parameters are the same as the default of COLMAP.

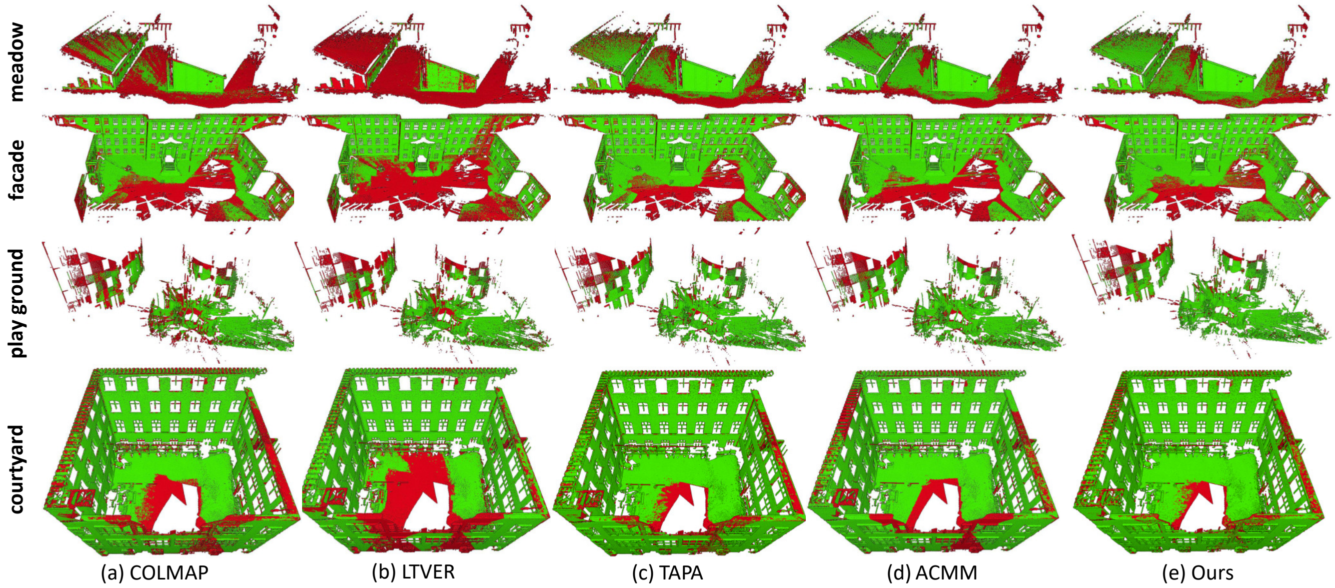


Figure 5: *Completeness comparison on the ETH3D benchmark. The models demonstrated are the ground truth provided by ETH3D benchmark. We select the threshold for completeness evaluation as 5cm. The green parts of the models are areas where there exist points reconstructed by MVS algorithms within a distance of 5cm. The red parts of the models are areas where there are no reconstructed points within 5cm.*

Dataset	Mean	Family	Francis	Horse	LightHouse	M60	Panther	Playground	Train
τ(mm)	╲	3	5	3	5	5	5	10	5
COLMAP	42.14	50.41	22.25	25.63	56.43	44.83	45.97	48.53	42.04
MVSNet	43.48	55.99	28.55	25.07	50.79	53.96	50.886	47.9	34.69
R-MVSNet Dense	50.55	**73.01**	**54.46**	43.42	43.88	46.8	46.69	50.88	45.25
ACMM	**57.27**	69.24	51.45	**46.97**	**63.2**	**55.07**	**57.64**	**60.08**	**54.48**
PLC	54.56	70.04	50.3	41.44	58.86	49.19	55.53	56.41	54.13

Table 1: *Quantitative F-score comparisons based on the intermediate dataset group of Tanks and Temples benchmark. τ is the default threshold for evaluation of each dataset provided by the benchmark.*

Dataset	Mean	Auditorium	Ballroom	Courtroom	Museum	Palace	Temple
τ(mm)	╲	10	10	10	10	30	15
COLMAP	27.24	16.02	25.23	34.7	41.51	18.05	27.94
R-MVSNet	29.55	19.49	31.45	29.99	42.31	22.94	31.1
ACMM	34.02	**23.41**	**32.91**	41.17	48.13	23.87	**34.60**
PLC	**34.44**	23.02	30.95	**42.50**	**49.61**	**24.09**	34.46

Table 2: *Quantitative F-score comparisons based on the advanced dataset group of Tanks and Temples benchmark. τ is the default threshold for evaluation of each dataset provided by the benchmark.*

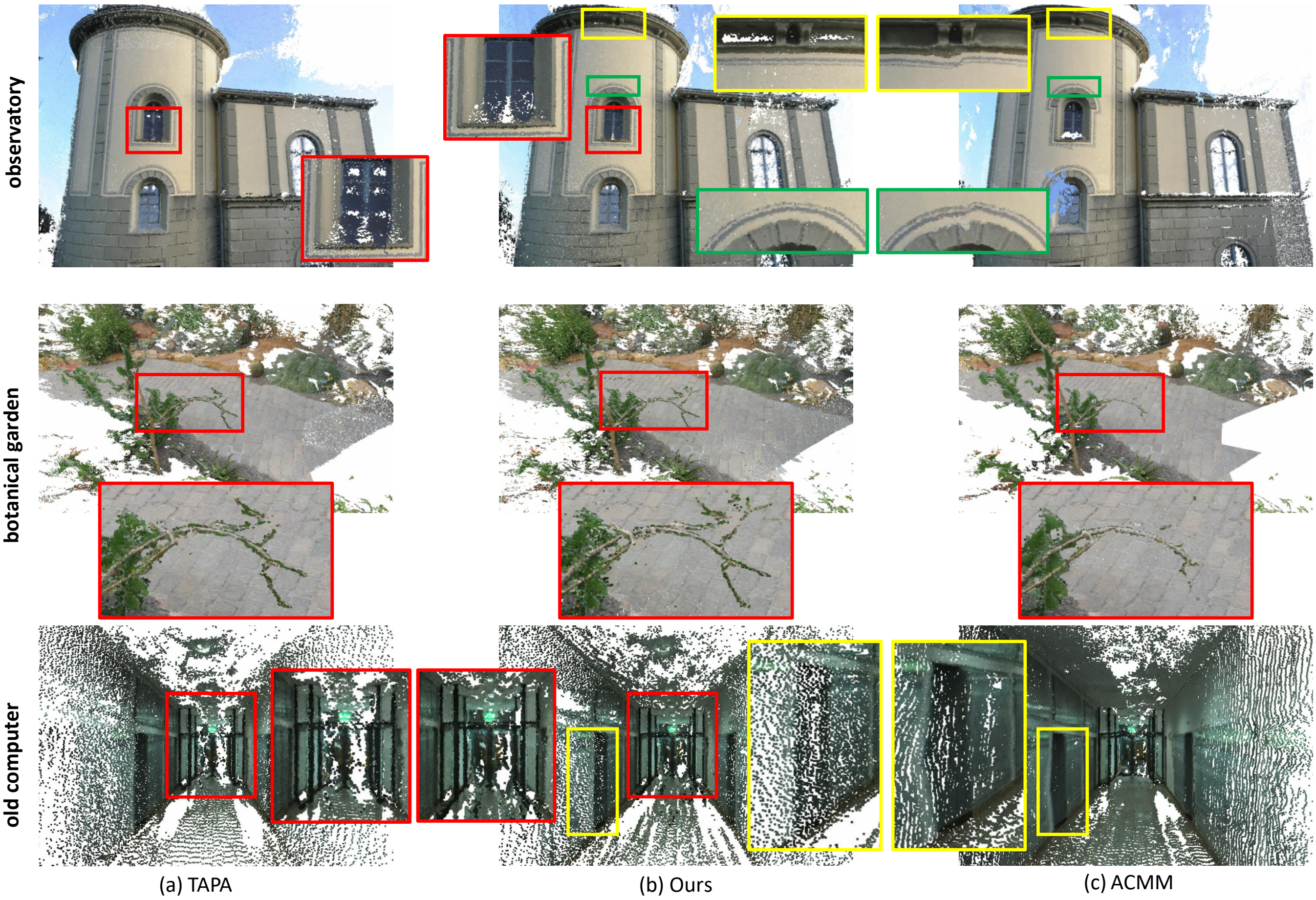


Figure 6: *Detail comparison on ETH3D benchmark. The small images with colored outlines are the enlarged views for areas in the nearest big images with the same color of edges.*

Table 1 and Table 2 demonstrate the F-scores of our proposed method against published state-of-the-art MVS algorithms on the intermediate and advanced groups. For the intermediate group, the mean F-score of models generated by our method ranks the second. For the advanced group, our method ranks the first on three of the six datasets and gets the highest mean F-score.

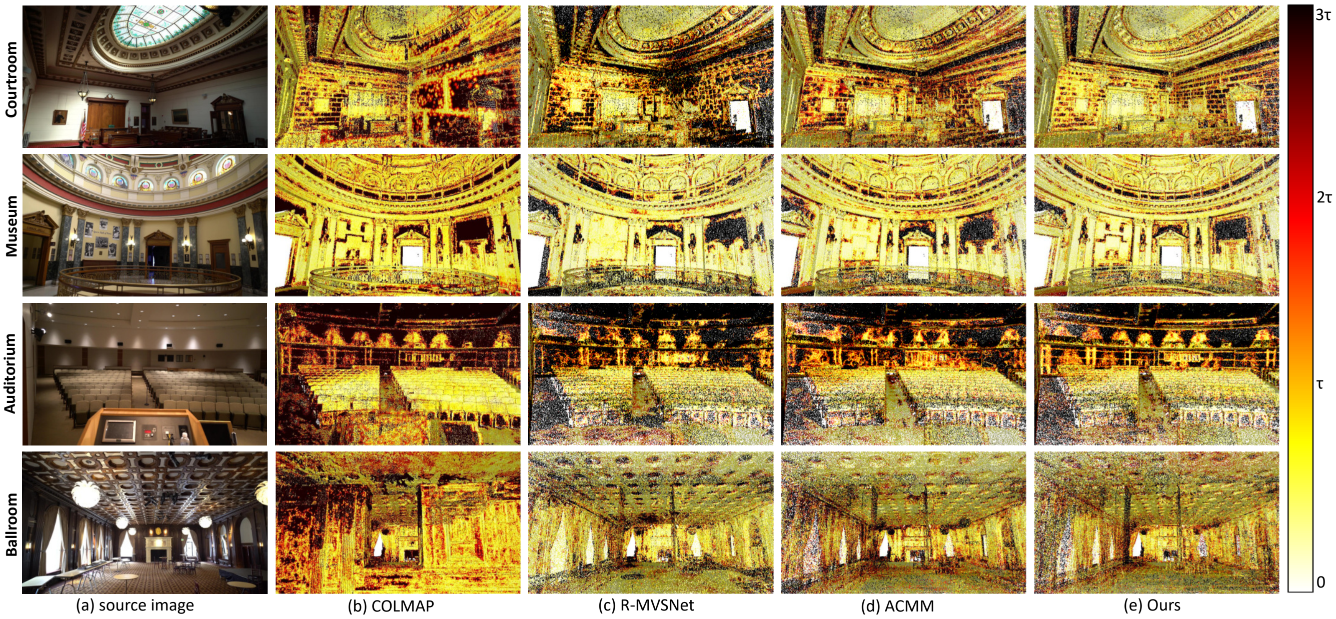
Figure 7 demonstrates the qualitative recall map comparisons on indoor datasets of the advanced group which contains textureless regions. It can be observed that the completeness of the models reconstructed by our method and ACMM outperforms the others, while our method reconstructs the detailed structures better than ACMM (e.g., the wall of the *Auditorium*)

5.3. Ablation Study

We assess the effectiveness of modified photometric consistency, local consistency and pyramid architecture on the *office* dataset from high-resolution training datasets provided by ETH3D benchmark. Table 3 represents the F_1 score of PLC-MVS without modifying photometric consistency measurement (PC), without pyramid architecture but using the same number of iterations (PA) and with decreasing λ coefficient (local consistency is out of service while $\lambda = 1$). Besides, we add an evaluation of COLMAP to Table 3 as the baseline. Figure 8 demonstrates the corresponding depth maps.

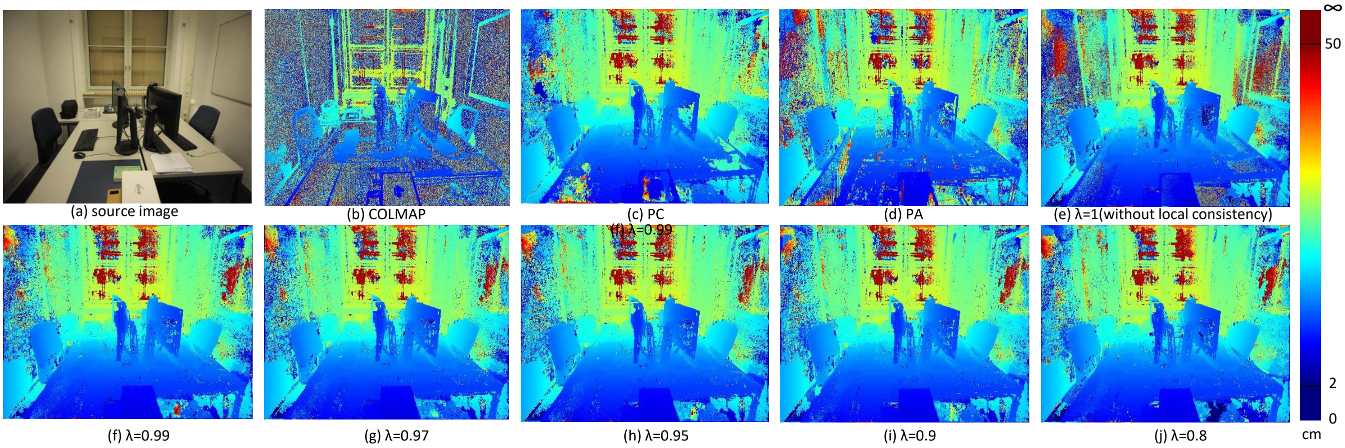
It can be observed that without modifying the photometric consistency measurement (PC), F_1 score and completeness of the desk decreases. The reason is that the view selection procedure is tightly related to the photometric consistency. Lowest photometric consistency value assigned to textureless regions by original measurement will disturb view selection in untextured regions which will in return affect the geometry estimation. Without pyramid architecture (PA), our algorithm can still reconstruct parts of the textureless regions with the guidance of local consistency, but the geometry of textureless regions do not convergence. When $\lambda = 1$, local consistency is out of service and depth hypotheses for textureless desktop are randomized. Applying local consistency will significantly improve the F_1 score as shown in Table 4.

Table 4 demonstrates the accuracy and completeness of results generated by PLC with decreasing λ. It can be observed that accuracy decreases with decreasing λ as some structures may be smoothed by over-weighted local consistency.



(a) source image (b) COLMAP (c) R-MVSNet (d) ACMM (e) Ours

Figure 7: *Qualitative recall map comparisons between different algorithms on some advanced datasets of Tanks and Temples benchmark. The pixel color of the recall map represents the distance from the ground truth to the nearest reconstructed point as shown in the legend on the right. τ is the threshold defined uniquely for each dataset by the benchmark as shown in Table 2.*



(a) source image (b) COLMAP (c) PC (d) PA (e) λ=1(without local consistency)

(f) λ=0.99 (g) λ=0.97 (h) λ=0.95 (i) λ=0.9 (j) λ=0.8

Figure 8: *Ablation Study: without modifying photometric consistency (PC), without pyramid architecture (PA), without local consistency ($\lambda = 1$) and with decreasing local consistency. On the right side is a simplified legend mapping depth values to colors. Depth values between 0 and 2 centimeters are demonstrated with the same color, and it is the same case for depth values above 50 centimeters. Challenging areas are the white desktop, white box, black screen and black floor on the bottom-left of the image.*

τ	COLMAP	PC	PA	$\lambda = 1$	$\lambda = 0.99$	$\lambda = 0.98$	$\lambda = 0.97$	$\lambda = 0.95$	$\lambda = 0.9$	$\lambda = 0.8$
1cm	37.10	37.90	39.94	43.97	46.19	**46.33**	46.19	16.11	44.32	43.64
2cm	47.32	46.67	54.15	55.87	**59.32**	59.19	59.13	58.96	57.26	57.30
5cm	62.27	66.42	75.05	71.13	**75.64**	75.38	75.51	75.58	74.50	74.83
10cm	73.41	79.05	**89.42**	81.79	85.94	85.73	85.90	85.98	85.44	85.54

Table 3: *Ablation study based on F_1 scores for the* office *dataset: without modifying photometric consistency (PC), without pyramid architecture (PA), without local consistency ($\lambda = 1$) and with decreasing local consistency. τ is the threshold for evaluation.*

	Accuracy						Completeness					
λ / τ	0.99	0.98	0.97	0.95	0.9	0.8	0.99	0.98	0.97	0.95	0.9	0.8
1cm	**72.34**	71.51	71.31	70.43	68.93	69.30	33.91	**34.26**	34.15	34.16	32.66	31.86
2cm	**80.10**	79.20	79.01	78.70	76.76	78.00	47.13	**47.25**	47.23	47.14	45.66	45.28
5cm	**88.10**	87.95	87.23	87.00	85.50	86.91	66.27	66.24	**66.55**	64.81	66.01	65.70
10cm	**91.76**	91.28	91.13	90.99	89.77	91.04	80.81	80.82	81.23	**81.54**	81.51	80.66

Table 4: *Ablation study based on F_1 scores for the* office *dataset: with decreasing λ accuracy decreases and completeness first increases and then decreases.*

	5cm					10cm				
	COLMAP	ACMM	TAPA	LTVRE	PLC	COLMAP	ACMM	TAPA	LTVRE	PLC
low-resolution	**81.80**	72.07	76.12	72.22	73.65	81.57	82.20	**84.80**	82.23	83.54
high-resolution	83.96	**89.14**	88.16	86.26	89.06	90.40	92.96	92.30	90.99	**94.11**
indoor	82.04	**88.48**	87.01	84.90	88.03	89.28	92.50	91.22	89.92	**93.43**
outdoor	89.74	91.12	91.62	90.34	**92.15**	93.79	94.35	95.56	94.19	**96.13**
botanical garden	95.28	94.65	96.64	94.44	**97.27**	97.95	96.30	98.33	96.18	**98.58**
boulders	79.75	81.22	80.65	81.50	**82.69**	87.13	88.71	89.60	89.55	**91.15**
bridge	94.25	**95.31**	94.39	92.31	94.29	96.92	**97.58**	96.79	96.31	96.69
door	92.28	96.12	**96.21**	94.89	95.44	95.89	97.36	**97.65**	97.09	97.39
exhibition hall	75.17	**85.31**	76.38	83.03	79.71	82.83	**91.40**	81.83	87.88	87.29
lecture room	78.02	85.92	**86.19**	79.63	84.01	86.88	90.53	**91.52**	85.67	91.37
living room	93.77	94.34	**95.44**	92.67	94.18	97.18	96.65	**97.23**	95.91	96.71
lounge	58.60	70.19	**79.75**	69.69	76.68	73.77	79.77	87.94	78.79	**88.39**
observatory	97.84	98.06	98.08	97.79	**98.75**	99.29	98.75	98.91	98.57	**99.52**
old computer	65.06	**85.65**	62.75	71.20	77.28	78.44	**91.24**	71.31	81.50	87.33
statue	85.92	88.86	**95.29**	86.26	93.41	93.62	91.64	**98.37**	89.98	97.13
terrace	91.63	94.08	**96.13**	91.72	95.02	94.95	95.59	**98.17**	94.44	97.73
lakeside	74.64	75.71	**80.96**	77.23	75.81	83.81	84.94	**88.22**	86.24	85.38
sand box	78.97	79.93	**80.06**	74.60	78.67	87.08	87.68	**88.54**	82.83	87.33
storage room1	57.39	54.53	63.78	**65.16**	61.39	69.46	68.85	75.35	**76.93**	73.45
storage room2	67.89	70.21	**76.37**	67.32	71.89	80.01	81.84	**84.91**	79.30	83.31
tunnel	80.10	79.97	79.45	76.79	**80.49**	87.49	87.67	86.96	85.84	**88.23**
All	80.39	84.12	**84.62**	82.13	84.53	87.81	89.79	90.10	88.41	**91.00**

Table 5: *Quantitative F_1 score comparisons on ETH3D benchmark with thresholds of* 5cm *and* 10cm*.*

5.4. Limitations

For some outdoor scenes, our method falsely estimates the depth values for sky regions, which are sometimes color-homography but inherently unreconstructable for MVS datasets. Due to the employment of geometric constraint for each level of pyramid architecture, some of the erroneous depth values may simultaneously satisfy the photometric and geometric consistency across different views. These incorrect depth hypotheses will not be filtered and will be fused as 3D points. One example that our method falsely reconstructs the sky region is shown in Figure 9. For large area of texture-less regions which may occupy around 1/6 of the reference images (e.g., the white wall shown in Figure 8), several iterations are not enough for the convergence of local consistency.

6. Conclusions and Future Work

We proposed a PatchMatch-based Multi-View Stereo algorithm which can delicately and robustly estimate geometry in textureless regions. We first modify the photometric consistency measurement utilized in COLMAP to make the photometric cost of textureless regions not always maximized and thus assigns its reference value for view selection. By applying local consistency which constrains depth and normal estimates in correspondence to that of neighboring pixels with similar color and low photometric-geometric cost, the optimization procedure can discriminate and choose the optimal hypotheses. The introduction of pyramid architecture similar to [XT19] fastens the convergence of local consistency to form stable and right estimates. Compared to previous works, our method does not rely on the hard assumption that textureless regions are piecewise planar and can handle the geometric estimation pixel-wisely and softly. Experiments on the ETH3D benchmark and Tanks and Temples benchmark show that our method can effectively reconstruct textureless regions even for curve surfaces while keeping the detailed structure.

In the future, we are going to introduce the semantic segmentation for depth estimation of different objects and utilize vary-

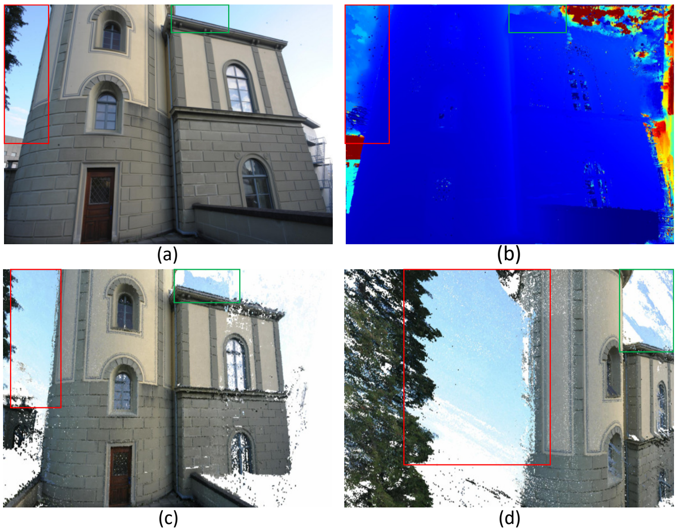


Figure 9: *(a) is the source image. (b) is the corresponding depth map estimated by our method. (c) and (d) are images of the reconstructed point cloud. The sky regions which are textureless and falsely reconstructed are shown in colored boxes.*

ing numbers of iteration for different areas. We are going to study point cloud meshing and combine the texture mapping algorithm proposed by Fu *et al.* [FYY*18] and illumination decomposition method proposed by Zhang *et al.* [ZYL*17] to reconstruct more realistic 3D models.

References

[AFS*09] Agarwal S., Furukawa Y., Snavely N., Curless B., Seitz S., Szeliski R.: Building rome in a day. In *ICCV* (2009). 1

[BR11] Bleyer M., Rhemann Christoph abd Rother C.: Patch-Match stereo-stereo matching with slanted support windows. In *BMVC* (2011). 1, 2, 3

[BRFK14] Besse F., Rother C., Fitzgibbon A., Kautz J.: PMBP: PatchMatch belief propagation for correspondence field estimation. In *ICCV* (2014). 3

[BSFG09] Barnes C., Shechtman E., Finkelstein A., Goldman D. B.: PatchMatch: A randomized correspondence algorithm for structural image editing. *SIGGRAPH* (2009). 1, 2

[BSGF10] Barnes C., Shechtman E., Goldman D. B., Finkelstein A.: The generalized PatchMatch correspondence algorithm. In *ECCV* (2010). 2

[FYY*18] Fu Y., Yan Q., Yang L., Liao J., Xiao C.: Texture mapping for 3D reconstruction with RGB-D sensor. In *CVPR* (2018). 12

[GLS15] Galliani S., Lasinger K., Schindler K.: Massively parallel multiview stereopsis by surface normal diffusion. In *ICCV* (2015). 3

[HKJK13] Heise P., Klose S., Jensen B., Knoll A.: PM-Huber: PatchMatch with huber regularization for stereo matching. In *ICCV* (2013). 3

[HMK*18] Huang P., Matzen K., Kopf J., Ahuja N., Huang J.: DeepMVS: Learning multi-view stereopsis. In *CVPR* (2018). 1

[HSGL11] HaCohen Y., Shechtman E., Goldman D. B., Lischinski D.: Non-rigid dense correspondence with applications for image enhancement. *ACM Transactions on Graphics 30*, 4 (2011), 70. 2

[JDV*14] Jensen R. R., Dahl A. L., Vogiatzis G., Tola E., Aanaes H.: Large scale multi-view stereopsis evaluation. In *CVPR* (2014). 2

[KCLU07] Kopf J., Cohen M. F., Lischinski D., Uyttendaele M.: Joint bilateral upsampling. In *SIGGRAPH* (2007). 6

[KHSM17] Kuhn A., Hirschmuller H., Scharstein D., Mayer H. J.: A TV prior for high-quality scalable multi-view stereo reconstruction. *International Journal of Computer Vision 124*, 1 (2017), 2–17. 6

[KPZK17] Knapitsch A., Park J., Zhou Q.-Y., Koltun V.: Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics 36*, 4 (2017). 1, 2, 7

[NH98] Neal R. M., Hinton G. E.: A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*. Springer, 1998, pp. 355–368. 3

[RM19] Romanoni A., Matteucci M.: TAPA-MVS: Textureless-aware PatchMatch multi-view stereo. In *CVPR* (2019). 2, 3, 4, 5, 6, 7

[SF16] Schönberger J. L., Frahm J.-M.: Structure-from-motion revisited. In *CVPR* (2016). 1

[She13] Shen S.: Accurate multiple view 3D reconstruction using patch-based stereo for large-scale scenes. *IEEE Transactions on Image Processing 22*, 5 (2013), 1901–1914. 3

[SSG*17] Schöps T., Schönberger J. L., Galliani S., Sattler T., Schindler K., Pollefeys M., Geiger A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR* (2017). 1, 2, 6

[SSS06] Snavely N., Seitz S. M., Szeliski R.: Photo tourism: exploring photo collections in 3D. *ACM Transactions on Graphics 25*, 3 (2006), 835–846. 1

[SvHV*08] Strecha C., von Hansen W., Van Gool L., Fua P., Thoennessen U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *CVPR* (2008). 1, 2

[SZPF16] Schönberger J. L., Zheng E., Pollefeys M., Frahm J.-M.: Pixelwise view selection for unstructured multi-view stereo. In *ECCV* (2016). 1, 2, 3, 5, 6

[WYL*19] Wei M., Yan Q., Luo F., Song C., Xiao C.: Joint bilateral propagation upsampling for unstructured multi-view stereo. *The Visual Computer* (2019), 1–13. 6

[XT19] Xu Q., Tao W.: Multi-scale geometric consistency guided multi-view stereo. In *CVPR* (2019). 2, 5, 6, 11

[YLL*18] Yao Y., Luo Z., Li S., Fang T., Quan L.: MVSNet: Depth inference for unstructured multi-view stereo. In *ECCV* (2018). 1, 6

[YLL*19] Yao Y., Luo Z., Li S., Shen T., Fang T., Quan L.: Recurrent MVSNet for high-resolution multi-view stereo depth inference. In *CVPR* (2019). 1, 6

[YYL*16] Yan Q., Yang L., Liang C., Liu H., Hu R., Xiao C.: Geometrically based linear iterative clustering for quantitative feature correspondence. In *PG* (2016). 1

[YYZX17] Yan Q., Yang L., Zhang L., Xiao C.: Distinguishing the indistinguishable: Exploring structural ambiguities via geodesic context. In *CVPR* (2017). 1, 3

[ZDJF14] Zheng E., Dunn E., Jojic V., Frahm J.-M.: PatchMatch based joint view selection and depthmap estimation. In *CVPR* (2014). 2, 3, 5

[ZYL*17] Zhang L., Yan Q., Liu Z., Zou H., Xiao C.: Illumination decomposition for photograph with multiple light sources. *IEEE Transactions on Image Processing 26*, 9 (2017), 4114–4127. 12