

Planar Prior Assisted PatchMatch Multi-View Stereo

Qingshan Xu and Wenbing Tao*

National Key Laboratory of Science and Technology on Multispectral Information Processing
 School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China
 {qingshanxu, wenbingtao}@hust.edu.cn

Abstract

The completeness of 3D models is still a challenging problem in multi-view stereo (MVS) due to the unreliable photometric consistency in low-textured areas. Since low-textured areas usually exhibit strong planarity, planar models are advantageous to the depth estimation of low-textured areas. On the other hand, PatchMatch multi-view stereo is very efficient for its sampling and propagation scheme. By taking advantage of planar models and PatchMatch multi-view stereo, we propose a planar prior assisted PatchMatch multi-view stereo framework in this paper. In detail, we utilize a probabilistic graphical model to embed planar models into PatchMatch multi-view stereo and contribute a novel multi-view aggregated matching cost. This novel cost takes both photometric consistency and planar compatibility into consideration, making it suited for the depth estimation of both non-planar and planar regions. Experimental results demonstrate that our method can efficiently recover the depth information of extremely low-textured areas, thus obtaining high complete 3D models and achieving state-of-the-art performance.

Introduction

Multi-view stereo (MVS) aims to estimate the dense 3D model of the scene from a given set of calibrated images. Due to its wide applications in virtual/augmented reality and 3D printing and so on, much progress has been made in this domain (Furukawa and Ponce 2010; Strecha, Fransens, and Van Gool 2006; Merrell et al. 2007; Goesele et al. 2007; Liu et al. 2009; Schönberger et al. 2016) in the last few years. However, recovering a dense and realist 3D model is still a challenging problem since the depth estimation in low-textured areas always fails.

The failure of depth estimation in low-textured areas mainly comes from the unreliable photometric consistency measure in these areas. As low-textured areas always appear in smooth homogeneous surfaces (Figure 1), many methods (Woodford et al. 2009; Gallup, Frahm, and Pollefeys 2010) assume that these surfaces are piecewise planar. Then, they formulate this prior as a regularization term in a global

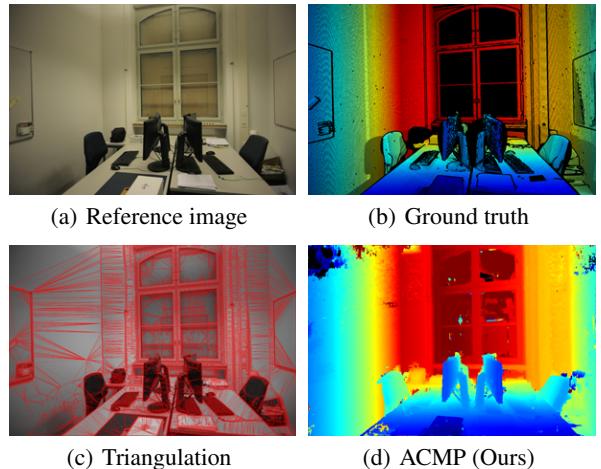


Figure 1: Illustration of discrimination and depth maps obtained by our method. The piecewise planar priors (c) through triangulation can adaptively gain discrimination for different low-textured areas. This helps to obtain better depth estimation for large low-textured areas (d).

energy framework to recover the depth estimation in low-textured areas. Due to the difficulty in solving such optimization problems, the efficiency of these methods is low and they are easy to be trapped in local optima. Recently, PatchMatch multi-view stereo methods (Zheng et al. 2014; Galliani, Lasinger, and Schindler 2015; Schönberger et al. 2016; Xu and Tao 2019) become popular as their used PatchMatch-based optimization (Barnes et al. 2009) makes depth map estimation efficient and accurate. As these methods do not explicitly model the planar priors, these methods still encounter the failure in low-textured areas. Based on the individual advantages of planar prior models and PatchMatch multi-view stereo, we expect to construct a planar prior assisted PatchMatch multi-view stereo framework to efficiently recover the depth estimation in low-textured areas.

To embed the planar prior models into PatchMatch multi-view stereo, in this work we rethink the right way to build the

*Corresponding author

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

multi-view aggregated matching cost and propose a planar prior assisted PatchMatch multi-view stereo framework to help the depth estimation in low-textured areas. In conventional PatchMatch multi-view stereo methods, sparse credible correspondences can be distinguished in discriminative regions, such as edges and corners. As these correspondences always coincide with the vertices in the mesh representation of 3D models, this means the sparse credible correspondences almost constitute the skeleton of a 3D model. Therefore, we first triangulate these correspondences to produce planar models. Note that, the planar priors are also suited for non-planar regions as credible correspondences are very dense in these regions and can adaptively form triangular primitives of different sizes. To derive the planar prior assisted multi-view aggregated matching cost, we leverage a probabilistic graphical model to simultaneously model photometric consistency and planar compatibility. The planar compatibility constrains predicted depth estimates to fall within an appropriate depth range while the photometric consistency can better reflect the depth changes in well-textured areas. At last, to alleviate the influence of unreliable planar priors, multi-view geometric consistency is enforced to rectify erroneous depth estimates.

In a nutshell, our contributions are as follows: **1)** We propose a novel planar prior assisted PatchMatch multi-view stereo framework for multi-view depth map estimation. This framework not only inherits the high efficiency of PatchMatch multi-view stereo but also leverages planar priors to help the depth estimation in low-textured areas. **2)** We adopt a probabilistic graphical model to induce a novel multi-view aggregated matching cost. This novel cost function takes both photometric consistency and planar compatibility into consideration. We demonstrate the effectiveness of our method by yielding state-of-the-art dense 3D reconstructions on ETH3D benchmark (Schöps et al. 2017). Our code will be available at <https://github.com/GhiXu/ACMP>.

Related Work

Our work is relevant to both PatchMatch multi-view stereo and planar priors, therefore we will review relevant literature in these areas.

PatchMatch Multi-View Stereo PatchMatch multi-view stereo methods exploit the core idea of PatchMatch (Barnes et al. 2009), sampling and propagation , to effectively estimate depth maps for each image. Focusing on different problems, many PatchMatch multi-view stereo methods have been proposed. (Zheng et al. 2014; Schönberger et al. 2016) jointly estimate depth maps and pixelwise view selection by a probabilistic graphical model. (Galliani, Lasinger, and Schindler 2015) utilizes a diffusion-like propagation scheme to make better use of the parallelization of GPUs. By inheriting the checkerboard pattern of (Galliani, Lasinger, and Schindler 2015), ACMH (Xu and Tao 2019) designs an adaptive checkerboard sampling strategy to propagate more reliable hypotheses. Moreover, ACMH further exploits these hypotheses to infer pixelwise view selection. However, as the photometric consistency on which these methods depend cannot get reliable discrimination in low-textured areas, the depth estimation of these methods always fails in

these areas. To get reliable discrimination from low-textured areas, (Wei, Resch, and Lensch 2014) leverages multi-scale scheme to achieve this at low resolution images. Then, it propagates the discrimination to the original resolution images by considering the relative depth difference from all neighboring views. On the multi-scale scheme, ACMM (Xu and Tao 2019) further considers the influence of view selection and leverages multi-scale geometric consistency to propagate the discrimination. Additionally, (Romanoni and Matteucci 2019) extracts superpixels at two scales and constrains the hypotheses in low-textured areas by the planes fitted for each superpixel. However, the discrimination obtained by the multi-scale scheme sometimes is limited by the predefined scales especially for large low-textured areas. In contrast, we leverage piecewise planar priors built from triangulation to adaptively acquire the discrimination for different low-textured areas.

Planar Priors Planar prior models are popular in 3D reconstruction as many scenes can be represented by a variety of plane primitives, especially for man-made environments. (Gennert 1988; Woodford et al. 2009) formulate planar prior models as second-order smoothness priors in a global energy function framework. This leads to a triple clique representation in the global energy function, making the optimization very difficult. To distinguish planar regions and non-planar objects in urban scenes, (Gallup, Frahm, and Pollefeys 2010) train a planar classifier to obtain the raw segmentation results and combine these segments with multi-view photometric consistency to define a global energy function to refine the predictions. Besides, it is worth noting that there exist some methods (Geiger, Roser, and Urtasun 2010; Zhang et al. 2015) employing triangulation to construct planar priors for disparity estimation in stereo matching. (Geiger, Roser, and Urtasun 2010) forms a triangulation on robustly matched correspondences to build a prior over the disparity space. This can not only reduce the disparity search space, but also recover low-textured surfaces. To simultaneously output a disparity map and a 3D triangulation mesh, (Zhang et al. 2015) first partitions stereo images into 2D triangles with shared vertices. Then, it formulates a two-layer Markov random field to jointly model disparity maps and vertex splitting probabilities. This assumes that the scene structure is piecewise planar and imposes regularization in the region-based stereo. Similar to (Geiger, Roser, and Urtasun 2010), our method also leverages the triangulation to build piecewise planar priors. Differently, embedding the priors into multi-view stereo is nontrivial as it needs to consider the visibility and geometry constraints of different views. To this aim, we embrace planar priors with PatchMatch MVS to consider the visibility and geometric constraints of different views.

Planar Prior Assisted PatchMatch MVS

Suppose we have a set of input images $\mathcal{I} = \{I_m | m = 1 \dots N\}$ with their corresponding camera parameters $\mathcal{P} = \{P_m | m = 1 \dots N\}$. Each image will be sequentially taken as a reference image I_{ref} while the other images are source images $\mathcal{I}_{src} = \{I_j | I_j \in \mathcal{I} \wedge I_j \neq I_{ref}\}$. Our work focuses on estimating the depth map of I_{ref} in turn.

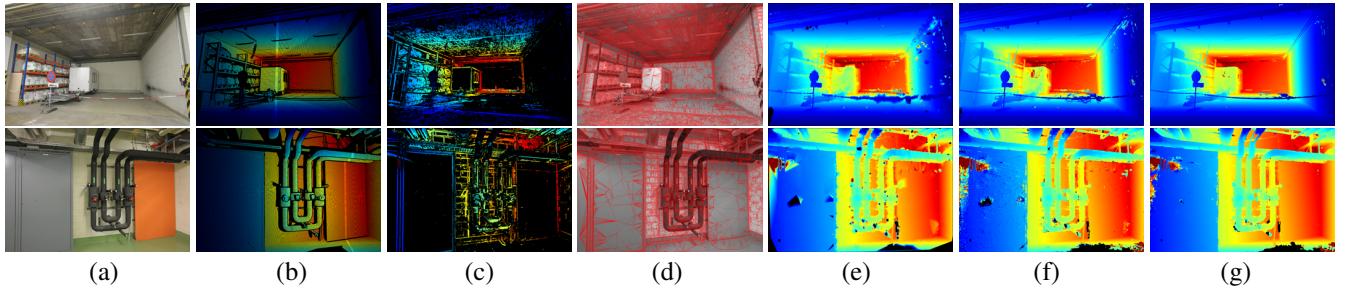


Figure 2: (a) images; (b) ground truth; (c) sparse correspondences; (d) triangulation; (e) planar model directly calculated from (d); (f) planar prior assisted PatchMatch MVS; (g) geometric consistency.

Currently, there exist two popular PatchMatch multi-view stereo frameworks, including sequential propagation pattern (Bailer, Finckh, and Lensch 2012; Schönberger et al. 2016) and checkerboard propagation pattern (Galliani, Lasinger, and Schindler 2015; Xu and Tao 2019). As pointed out in (Xu and Tao 2019), the latter one is more efficient and effective than the former one, thus we will build our algorithm on the checkerboard propagation pattern.

Different from the conventional PatchMatch multi-view stereo methods, our method also takes as input the sparse credible correspondences of the reference image. To estimate the depth information of low-textured areas, our method consists of two stage. In the first stage, sparse correspondences are generated by conventional PatchMatch MVS methods and thresholding. Then, we triangulate these correspondences to produce planar models. In the second stage, we jointly consider the previous obtained planar models and photometric consistency by constructing a probabilistic graphical model. This derives a novel multi-view aggregated matching cost. By embedding this novel cost to the pipeline of PatchMatch MVS, we can obtain good depth estimation for low-textured areas.

Planar Model Construction

To start our algorithm, we implement the method of (Xu and Tao 2019) to obtain sparse credible correspondences. The method follows the pipeline of PatchMatch MVS, iteratively performing adaptive checkerboard sampling and propagation, hypothesis updating via multi-view aggregated photometric consistency cost and refinement. A depth estimate will be considered as a credible correspondence if its final cost is lower than 0.1 (Figure 2c).

Given the sparse credible correspondences of I_{ref} , they always characterize the structure of a scene. Although the depth estimation for low-textured areas is lost, people can imagine the whole 3D model of a scene according to these correspondences. Based on this observation, we first triangulate these sparse credible correspondences to adaptively generate triangular primitives of different sizes. As can be seen from Figure 2d, the triangular primitives in well-textured areas are relatively small so that the structures of non-planar regions can be kept. On the other hand, the triangular primitives in low-textured areas are as large as possible to incorporate the information of credible correspondences.

For each triangular primitive, we use its corresponding three vertices to calculate its plane parameters in the coordinate of the reference camera, including depth information and normal information. The pixels inside the same triangular primitive share the same plane parameters. Figure 2e shows two examples of planar models. It can be observed that the priori plane parameters can almost coincide with the optimal estimates for low-textured areas. It is worth noting that the structures of thin objects are also described by the generated triangular primitives.

Planar Prior Assistance

With the priori plane hypotheses, the depth estimate for low-textured areas can be better approximated. However, these priori plane hypotheses also lead to many blocking artifacts in well-textured areas, especially in boundaries, whose depth information should be estimated well by photometric consistency. To take both photometric consistency and piecewise planar priors into consideration, we employ a probabilistic model to achieve this.

Random Initialization As a first step, we randomly generate a plane hypothesis $\theta_l = [d_l, \mathbf{n}_l]$ for each pixel l in the reference image, where d_l is distance from a 3D plane to the origin and \mathbf{n}_l is normal vector. We first calculate a matching cost for each source image via the plane hypothesis induced homography (Hartley and Zisserman 2004). Then, the initial multi-view aggregated matching cost for each hypothesis is computed by averaging the top- K smallest matching costs.

Hypothesis Sampling and Propagation Following (Galliani, Lasinger, and Schindler 2015; Xu and Tao 2019), we divide all pixels in the reference image into a Red-Black pattern. This allows to use the hypotheses of red pixels to update those of black pixels and vice versa. Then, we use the adaptive checkerboard sampling of (Xu and Tao 2019) to propagate eight neighboring good hypotheses to the current pixel to be estimated. These propagated hypotheses together with the current hypothesis θ_0 of the pixel constitute the current candidate hypothesis set, $\Theta = \{\theta_i | i = 0 \dots 8\}$.

Hypothesis Updating In conventional PatchMatch multi-view stereo methods (Zheng et al. 2014; Schönberger et al. 2016; Xu and Tao 2019), in order to determine the best hypothesis from the candidate hypothesis set, the following multi-view aggregated matching cost is defined by photo-

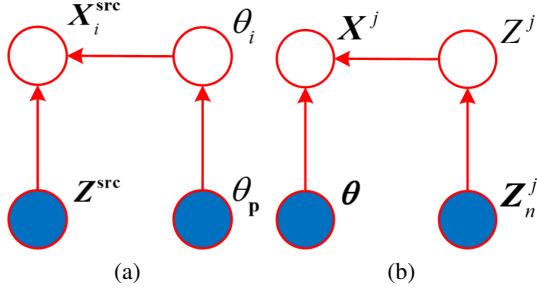


Figure 3: (a) Graphical model of planar prior assistance. Given priori plane hypothesis θ_p , the observation X_i^{src} on source images and the visibility information Z^{src} , the optimal hypothesis θ^* is inferred. (b) Graphical model of view selection. At each iteration, given candidate hypotheses θ , the observation X^j corresponding to the hypotheses on source image I_j , and the visibility of neighboring pixels on source image I_j is Z_n^j , the visibility of pixel l on source image I_j , Z_l^j , is inferred.

metric consistency to measure the multi-view similarity,

$$c_{\text{photo}}(\theta_i) = \frac{\sum_j w_j \cdot m_{i,j}}{\sum_j w_j}, \quad (1)$$

where $m_{i,j}$ is the matching cost between the reference patch and its corresponding source patch observed on source image I_j via θ_i and w_j is the view selection weight of I_j . As the photometric consistency is unreliable in low-textured areas, these methods always fail in these areas.

In contrast, given the planar priors as described before, we leverage a probabilistic graphical model to derive our novel multi-view aggregated matching cost. To construct the graphical model, we define the patch on pixel l of I_{ref} as X^{ref} . Also, the patches observed on all source images via θ_i are X_i^{src} , the visibility information of all source images is assumed to be Z^{src} and the planar prior at pixel l is $\theta_p = [d_p, n_p]$. Then, the graphical model of our approach is depicted in Figure 3a. The joint probability is

$$P(\theta_i, X_i^{\text{src}}, Z^{\text{src}}, \theta_p) \propto P(X_i^{\text{src}}|\theta_i, Z^{\text{src}})P(\theta_i|\theta_p). \quad (2)$$

In this way, the maximum a-posteriori estimate of the plane hypothesis θ^* is given by

$$\theta^* = \arg \max P(\theta_i|X_i^{\text{src}}, Z^{\text{src}}, \theta_p). \quad (3)$$

The above posterior can be factorized as

$$P(\theta_i|X_i^{\text{src}}, Z^{\text{src}}, \theta_p) \propto P(X_i^{\text{src}}|\theta_i, Z^{\text{src}})P(\theta_i|\theta_p). \quad (4)$$

Next, we define the likelihood function as follows,

$$P(X_i^{\text{src}}|\theta_i, Z^{\text{src}}) = e^{-\frac{c_{\text{photo}}(\theta_i)^2}{\alpha}}. \quad (5)$$

This function encodes the photometric consistency, making the low multi-view aggregated photometric consistency cost have high probability. It encourages our whole algorithm to choose the hypothesis with lower multi-view aggregated photometric consistency cost, which is consistent with the

hypothesis update criteria in the conventional PatchMatch multi-view stereo methods. However, due to the unreliability of photometric consistency in low-textured areas, this likelihood function will not reflect hypothesis changes in these areas. In this case, it is important to leverage planar priors to reflect these changes. Thus, we define the planar prior as

$$P(\theta_i|\theta_p) = \gamma + e^{-\frac{(d_i - d_p)^2}{2\lambda_d}} \cdot e^{-\frac{\arccos^2 n_i^\top n_p}{2\lambda_n}}, \quad (6)$$

where λ_d is the bandwidth of depth difference and λ_n is the bandwidth of normal difference. The planar prior encourages the propagated hypotheses to be close to the planar model at pixel l . We substitute Equation (4)-(6) into Equation (3) and take the negative logarithm algorithm to get the following planar prior assisted multi-view aggregated matching cost

$$c_{\text{p-photo}}(\theta_i) = \frac{c_{\text{photo}}(\theta_i)^2}{\alpha} - \log[\gamma + e^{-\frac{(d_i - d_p)^2}{2\lambda_d}} \cdot e^{-\frac{\arccos^2 n_i^\top n_p}{2\lambda_n}}]. \quad (7)$$

Note that, the first term that encodes photometric consistency is the main component in the above equation. This means that the photometric consistency will change more obviously than the planar prior in well-textured areas. Moreover, when the photometric consistency cannot reflect hypothesis changes in low-textured areas, the planar prior will play a major role in the hypothesis updating.

As mentioned above, the photometric consistency is important to determine the depth information in well-textured areas. This can rectify the erroneous depth estimates induced by planar models in no-planar areas. According to Equation (1), the reliability of multi-view aggregated photometric consistency depends upon the view selection weights. To calculate these weights, we design another probabilistic graphical model to make full use of the photometric consistency of different source images and the view selection information of neighboring pixels.

Specifically, we denote that the visibility of pixel l on source image I_j is Z_l^j , the visibility of neighboring pixels of pixel l on source image I_j is Z_n^j , the candidate hypotheses are θ , the patch on pixel l in the reference image is X^{ref} , and its corresponding patches observed on source image I_j via θ are $X^j = \{X_i^j | i = 0 \dots 8\}$. The graphical model is depicted in Figure 3b. According to the states of neighboring pixels, the joint probability is

$$P(X^j, Z^j, \theta, Z_n^j) \propto P(X^j|Z^j, \theta)P(Z_l^j|Z_n^j), \quad (8)$$

where $P(X^j|Z^j, \theta) = \sum_{i=0}^8 P(X_i^j|Z^j, \theta_i)$ independently models the possible source image subset for each hypothesis while $P(Z_l^j|Z_n^j) = \sum_{l' \in \mathcal{N}(l)} P(Z_l^j|Z_{l'}^j)$ models the smoothness of the view selection of neighboring pixels. Note that, $\mathcal{N}(l)$ stands for the four neighboring pixels of pixel l . Specifically, we define $P(X_i^j|Z^j, \theta_i)$ as

$$P(X_i^j|Z^j, \theta_i) = e^{-\frac{m_{i,j}^2}{2\sigma^2}}, \quad (9)$$

where σ is a constant. And, $P(Z_l^j|Z_{l'}^j)$ is defined as

$$P(Z_l^j|Z_{l'}^j) = \begin{cases} \eta, & \text{if } Z_l^j = Z_{l'}^j; \\ 1 - \eta, & \text{else.} \end{cases} \quad (10)$$

According to Bayesian rule, the view selection probability of pixel l on source image I_j is

$$P(Z^j | \mathbf{X}^j, \boldsymbol{\theta}, \mathbf{Z}_n^j) \propto P(\mathbf{X}^j | Z^j, \boldsymbol{\theta}) P(Z^j | \mathbf{Z}_n^j). \quad (11)$$

Based on the above view selection probabilities, we employ the Monte-Carlo sampling (Bishop 2006) to define the weight for each source image I_j as w_j .

Refinement After each hypothesis updating, we refine the current selected hypothesis $\theta_c = [d_c, \mathbf{n}_c]$ by generating extra candidate hypothesis set. Following (Schönberger et al. 2016; Xu and Tao 2019), this candidate hypothesis set is defined as

$$\{[d_p, \mathbf{n}_c], [d_r, \mathbf{n}_c], [d_c, \mathbf{n}_p], [d_c, \mathbf{n}_r], [d_r, \mathbf{n}_r], [d_p, \mathbf{n}_p]\}, \quad (12)$$

where d_p and \mathbf{n}_p are the randomly perturbed depth and normal with respect to θ_c , d_r and \mathbf{n}_r are randomly generated depth and normal. If the cost of a new hypothesis is less than that of the current hypothesis, we will set it as the current hypothesis. The above hypothesis sampling and propagation, hypothesis updating and refinement are performed iteratively to produce the depth map for the reference image.

Geometric Consistency

In the previous Section, we consider both piecewise planar priors and photometric consistency to get better depth estimation in low-textured and well-textured areas. However, there still exist some errors. This attributes to some unreliable planar priors caused by some intractable erroneous sparse correspondences. To tackle these errors, we resort to multi-view geometric consistency (Schönberger et al. 2016; Xu and Tao 2019), which is defined as

$$c_{\text{geo}}(\theta_i) = \frac{\sum_j w_j \cdot (m_{i,j} + \lambda_{\text{geo}} \cdot \min(\Delta e_j(\theta_i), \tau_{\text{geo}}))}{\sum_j w_j}, \quad (13)$$

where λ_{geo} is a geometric consistency regularizer, $\Delta e_j(\theta_i)$ is the reprojection error between I_{ref} and I_j induced by θ_i , and τ_{geo} is a truncation threshold to robustify the reprojection error against occlusions.

The Algorithm

The overall pipeline of our algorithm is summarized in Algorithm 1. From step 1 to step 9, we generate initial depth maps via conventional multi-view aggregated photometric consistency cost. Then, in step 10 and step 11, we select credible correspondences and triangulate them to generate planar models. From step 12 to step 20, we generate plane-awareness depth maps by our proposed planar prior assisted multi-view aggregated matching cost. Then, these depth maps are used as additional input in step 21. We further optimize these depth maps via geometric consistency from step 22 to step 30. To make each PatchMatch MVS process converge, T_{photo} , $T_{\text{p-photo}}$ and T_{geo} are set to 3, 3, 2, respectively.

Fusion

The depth maps estimated for individual images always contain noise and outliers. We follow the conventional PatchMatch pipeline (Schönberger et al. 2016; Xu and Tao 2019)

Algorithm 1 Planar Prior Assisted PatchMatch MVS

Input: multi-view images with their camera parameters

Output: hypothesis maps

```

1: for each image do
2:   set reference image and source images
3:   randomly initialize a hypothesis map
4:   for iteration  $i = 1$  to  $T_{\text{photo}}$  do
5:     hypothesis sampling and propagation
6:     update the hypothesis map via Equation (1)
7:     refinement via Equation (12)
8:   end for
9: end for
10: sparse credible correspondences selection
11: triangulation and generate planar models
12: for each image do
13:   set reference image and source images
14:   randomly initialize a hypothesis map
15:   for iteration  $i = 1$  to  $T_{\text{p-photo}}$  do
16:     hypothesis sampling and propagation
17:     update the hypothesis map via Equation (7)
18:     refinement via Equation (12)
19:   end for
20: end for
21: use the hypothesis maps obtained above as extra input
22: for each image do
23:   set reference image and source images
24:   use the previous obtained hypothesis map for the reference image as initialization
25:   for iteration  $i = 1$  to  $T_{\text{geo}}$  do
26:     hypothesis sampling and propagation
27:     update the hypothesis map via Equation (13)
28:     refinement via Equation (12)
29:   end for
30: end for

```

to use a fusion step to produce the final point cloud. Each image is treated as the reference image in turn and its depth estimates are unprojected to the world coordinate to obtain 3D points. These 3D points are further projected to neighboring images to calculate their projected depths, normals and image coordinates. According to the estimated depths and normals in projected image coordinates, a consistent estimate is determined if its relative depth difference is lower than 0.01, normal difference is lower than 10° , and reprojection error is less than 2 pixels. The estimate that has two consistent neighboring estimates are kept and their unprojected 3D points are averaged to produce the final 3D point.

Experiments

Datasets We evaluate the effectiveness of our method on high-resolution multi-view stereo dataset of ETH3D benchmark (Schöps et al. 2017). This dataset contains images at a resolution of 6048×4032 with calibration. Following (Schönberger et al. 2016; Xu and Tao 2019), we resize this imagery to no more than 3200 pixels for each dimension while keeping the original aspect ratio. The dataset is further split into training datasets and test datasets. Besides ground

Table 1: Percentage of pixels with absolute errors below $2cm$ and $10cm$ on the high-resolution multi-view training datasets of ETH3D benchmark (in %). ACMP\G means ACMP without geometric consistency. The related values are from (Xu and Tao 2019). The best results are marked in bold.

error	method	Ave.	indoor						outdoor						
			deli.	kick.	offi.	pipes	relief	relief.	terrai.	courty.	elec.	faca.	mead.	playgr.	terrace
$2cm$	COLMAP	65.0	69.7	43.5	26.3	41.1	86.3	85.8	57.6	82.6	71.0	74.2	54.6	70.9	80.8
	openMVS	55.2	63.9	36.9	29.3	31.8	80.1	81.5	57.9	64.3	55.6	55.4	29.8	57.9	73.6
	ACMH	68.5	73.3	42.7	32.3	53.6	89.1	90.3	71.4	79.9	74.8	68.5	57.1	75.3	82.0
	ACMM	80.5	77.7	66.7	51.2	76.5	96.0	95.7	85.4	84.4	86.8	74.5	77.1	84.3	89.7
	ACMP\G	72.9	76.9	47.4	48.6	65.3	91.3	92.6	79.1	78.8	79.2	68.9	59.5	73.7	85.9
	ACMP	81.9	81.9	62.0	65.6	78.6	94.8	95.8	88.4	84.5	88.7	76.6	76.8	80.6	90.7
$10cm$	COLMAP	73.7	80.6	51.4	34.2	47.8	89.6	89.3	63.5	93.4	77.4	90.9	70.1	81.0	89.1
	openMVS	66.5	79.1	45.1	38.2	42.1	84.1	86.0	67.1	79.1	64.8	77.4	48.4	68.7	83.8
	ACMH	79.1	84.2	51.9	41.8	61.7	92.3	94.1	77.8	93.7	83.4	90.8	78.6	86.9	91.5
	ACMM	90.7	93.0	80.0	64.8	83.9	98.2	98.4	90.4	97.3	94.7	93.4	91.7	95.1	98.0
	ACMP\G	85.1	90.7	61.3	66.6	74.8	94.5	96.6	86.1	93.7	88.6	90.7	80.0	86.8	95.3
	ACMP	90.6	95.4	72.4	78.0	84.2	96.8	98.2	92.4	96.1	94.4	93.2	87.5	91.5	97.9

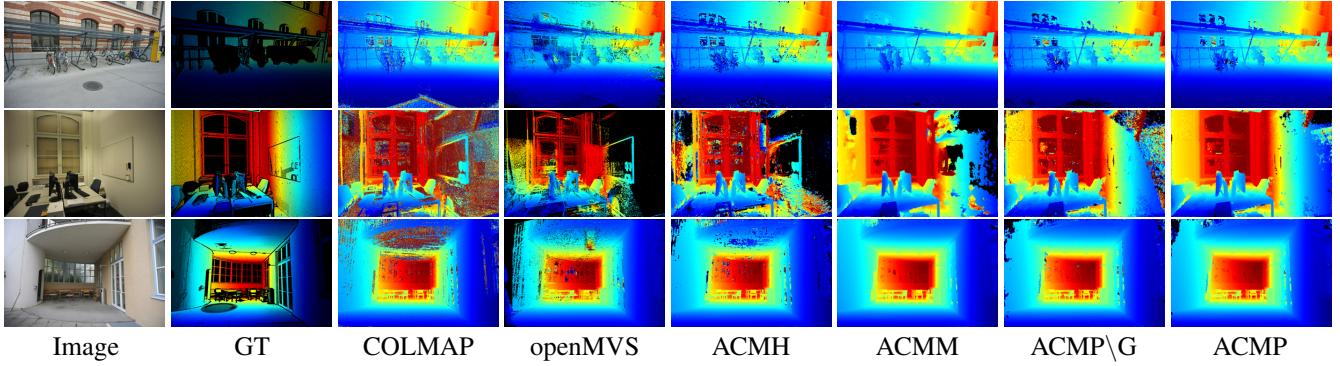


Figure 4: Qualitative depth map map comparisons of different methods on some high-resolution multi-view training datasets (courty., offi. and terrace) of ETH3D benchmark. Black pixels in GT mean no ground truth data.

truth point clouds, ground truth depth maps are also provided for training datasets. Thus, we first evaluate the depth estimation on training datasets. As for test datasets, we submit our reconstructed point clouds to the benchmark’s website (Schöps et al.) to evaluate them.

Evaluation Metrics In depth map evaluation, we calculate the percentage of pixels with an absolute depth error less than $2cm$ and $10cm$ from ground truth. For point cloud evaluation, we assess reconstructed point clouds in terms of accuracy, completeness and F_1 score.

Parameter Settings Our methods are implemented in C++ with CUDA and executed on a machine with two Intel E5-2630 CPUs and two GTX Titan X GPUs. $\{\epsilon, \alpha, \gamma, \lambda_n, \sigma, \eta, \lambda_{geo}, \tau_{geo}\} = \{0.1, 0.18, 0.5, 5^\circ, 0.3, 0.9, 0.1, 5.0\}$. Besides, λ_d is adaptively set to one sixty-fourth of the depth interval of every reference image. We conduct geometric consistency twice as (Xu and Tao 2019) does.

Depth Map Evaluation We compare our method with some state-of-the-art PatchMatch multi-view stereo methods in depth map evaluation, including COLMAP (Schönberger et al. 2016), openMVS (cDc Seacave), ACMH (Xu and Tao

2019) and ACMM (Xu and Tao 2019). These methods are directly operated on the original resolution images except for ACMM with multi-scale scheme. We denote our method as ACMP because our planar prior assisted model is based on adaptive checkerboard sampling and propagation.

We list comparison results on 13 high resolution multi-view training datasets of ETH3D benchmark in Table 1. In order to validate the effectiveness of our novel multi-view aggregated matching cost, we remove the geometric consistency from our method and denote this method as ACMP\G. We first compare ACMP\G with COLMAP, openMVS and ACMH. As can be seen, ACMP\G is much better than these methods. From the qualitative results in Figure 4, we observe that ACMP\G can estimate depth information of low-textured areas well. Moreover, ACMP\G can also tackle the depth estimation in non-planar regions because our planar prior assisted multi-view aggregated matching cost simultaneously considers the photometric consistency.

As ACMP\G does not consider the geometric consistency, its estimated depth maps contain some noise caused by inappropriate planar models. Therefore, ACMP combines ACMP\G with geometric consistency to handle this

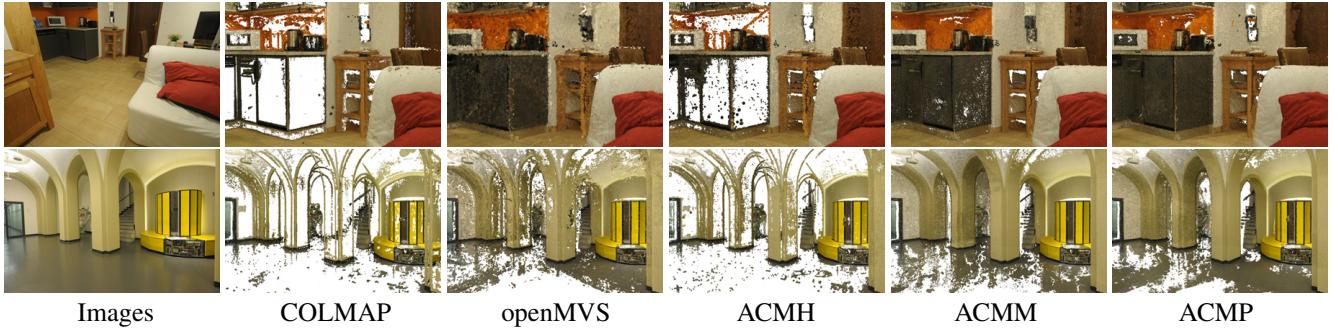


Figure 5: Qualitative point cloud comparisons of different MVS methods on living room and old computer of ETH3D benchmark. These 3D models are reported by the ETH3D benchmark evaluation server (Schöps et al.).

problem. This makes our method competitive with ACMM. Note that, ACMM uses multi-scale geometric consistency to tackle the depth estimation in low-textured areas. However, due to its limited scales and lost image information at its coarsest scale, ACMM sometimes cannot obtain good estimation for low-textured areas at the coarsest scale. This can be reflected by the depth estimation of offi. dataset in Figure 4. Differently, as ACMP adaptively captures the discrimination of different sizes according to triangular primitives, ACMP performs much better than ACMM in offi. dataset.

Table 2: Accuracy, completeness and F_1 score (in %) comparisons of reconstructed point clouds on the high-resolution multi-view test datasets of ETH3D benchmark at evaluation threshold $2cm$. The related values are from (Schöps et al.).

	method	Accuracy	Completeness	F_1
indoor	COLMAP	91.95	59.65	70.41
	openMVS	82.00	75.92	78.33
	ACMH	91.14	64.81	73.93
	ACMM	90.99	72.73	79.84
	ACMP	90.60	74.23	80.57
outdoor	COLMAP	92.04	72.98	80.81
	openMVS	81.93	86.41	84.09
	ACMH	83.96	80.03	81.77
	ACMM	89.63	79.17	83.58
	ACMP	90.35	79.62	84.36
all	COLMAP	91.97	62.98	73.01
	openMVS	81.98	78.54	79.77
	ACMH	89.34	68.62	75.89
	ACMM	90.65	74.34	80.78
	ACMP	90.54	75.58	81.51

Point Cloud Evaluation The quantitative results of reconstructed point clouds are listed in Table 2 and quantitative results are shown in Figure 5. In the case of $2cm$, ACMP achieves the best F_1 score among all methods due to our better estimated depth maps. As for the completeness of 3D models, openMVS produces the best performance as it employs a different fusion scheme with a relaxed number of consistent views for a pixel, which leads to noisy point clouds with low accuracy and high completeness. As illustrated in Figure 5, the 3D models of openMVS are not photo-

Table 3: Running time of different stages of our method for an image of size 3200×2130 pixels on a single GPU.

Stage	Time (s)	Ratio (%)
Sparse Correspondences Generation	6.40	30.9
Planar Model Construction	1.08	5.2
Planar Prior Assistance	5.43	26.2
Geometric Consistency	7.79	37.6
Total Time	20.7	-

realistic. Differently, our method can achieve a better trade-off between accuracy and completeness. This makes our estimated 3D models be applicable in the actual environment.

Runtime Analysis We run our method on a single GPU and record the running time of each stage in Table 3. As can be seen, the planar model construction occupies very little runtime. As for sparse correspondences generation, planar prior assistance and geometric consistency, their running time is very close. This is because these stages all employ the same pipeline of PatchMatch multi-view stereo. Therefore, with very little computational cost, our method without geometric consistency can achieve much better reconstruction results than other PatchMatch multi-view stereo methods that are implemented on the original image resolution.

Conclusion

In this work, we propose a planar prior assisted PatchMatch multi-view stereo framework to help the depth estimation in low-textured areas. To tackle depth ambiguities caused by unreliable photometric consistency, we leverage sparse credible correspondences to build planar models. These models can reflect the depth ranges of low-textured areas well but are still biased, especially for non-planar regions. Therefore, we embed these planar models into PatchMatch multi-view stereo by utilizing a probabilistic graphical model. This derives a novel multi-view aggregated matching cost, which jointly consists of photometric consistency and planar priors. This makes our method suited for both planar and non-planar regions. Experiments on ETH3D benchmark demonstrate the effectiveness of our methods by yielding state-of-the-art performance. We note that the performance of

our method is comparable to ACMM. Since ACMM combines ACMH with the multi-scale geometric consistency, we think its performance gain mainly comes from the multi-scale geometric consistency framework. As a separate module, we have demonstrated that our method is much better than ACMH in our experiments. Thus, we may combine our method with the multi-scale geometric consistency to improve the reconstruction performance in the future work.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grants 61772213 and 91748204.

References

- Bailer, C.; Finckh, M.; and Lensch, H. P. A. 2012. Scale robust multi view stereo. In *Proceedings of the European Conference on Computer Vision*, 398–411.
- Barnes, C.; Shechtman, E.; Finkelstein, A.; and Goldman, D. B. 2009. Patchmatch: A randomized correspondence algorithm for structural image editing. In *ACM SIGGRAPH*, 24:1–24:11.
- Bishop, C. M. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag.
- cDc Seacave. openMVS. <https://github.com/cdcseacave/openMVS>.
- Furukawa, Y., and Ponce, J. 2010. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(8):1362–1376.
- Galliani, S.; Lasinger, K.; and Schindler, K. 2015. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, 873–881.
- Gallup, D.; Frahm, J.; and Pollefeys, M. 2010. Piecewise planar and non-planar stereo for urban scene reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1418–1425.
- Geiger, A.; Roser, M.; and Urtasun, R. 2010. Efficient large-scale stereo matching. In *Proceedings of the Asian Conference on Computer Vision*, 25–38.
- Gennert, M. A. 1988. Brightness-based stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, 139–143.
- Goesele, M.; Snavely, N.; Curless, B.; Hoppe, H.; and Seitz, S. M. 2007. Multi-view stereo for community photo collections. In *Proceedings of the IEEE International Conference on Computer Vision*, 1–8.
- Hartley, R., and Zisserman, A. 2004. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition.
- Liu, Y.; Cao, X.; Dai, Q.; and Xu, W. 2009. Continuous depth estimation for multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2121–2128.
- Merrell, P.; Akbarzadeh, A.; Wang, L.; Mordohai, P.; Frahm, J.; Yang, R.; Nister, D.; and Pollefeys, M. 2007. Real-time visibility-based fusion of depth maps. In *Proceedings of the IEEE International Conference on Computer Vision*, 1–8.
- Romanoni, A., and Matteucci, M. 2019. TAPA-MVS: textureless-aware patchmatch multi-view stereo. *CoRR* abs/1903.10929.
- Schönberger, J. L.; Zheng, E.; Frahm, J.-M.; and Pollefeys, M. 2016. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, 501–518.
- Schöps, T.; Schönberger, J. L.; Galliani, S.; Sattler, T.; Schindler, K.; Pollefeys, M.; and Geiger, A. ETH3D Benchmark. <https://www.eth3d.net>.
- Schöps, T.; Schönberger, J. L.; Galliani, S.; Sattler, T.; Schindler, K.; Pollefeys, M.; and Geiger, A. 2017. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2538–2547.
- Strecha, C.; Fransens, R.; and Van Gool, L. 2006. Combined depth and outlier estimation in multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2394–2401.
- Wei, J.; Resch, B.; and Lensch, H. 2014. Multi-view depth map estimation with cross-view consistency. In *Proceedings of the British Machine Vision Conference*.
- Woodford, O.; Torr, P.; Reid, I.; and Fitzgibbon, A. 2009. Global stereo reconstruction under second-order smoothness priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(12):2115–2128.
- Xu, Q., and Tao, W. 2019. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5483–5492.
- Zhang, C.; Li, Z.; Cheng, Y.; Cai, R.; Chao, H.; and Rui, Y. 2015. Meshstereo: A global stereo model with mesh alignment regularization for view interpolation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2057–2065.
- Zheng, E.; Dunn, E.; Jojic, V.; and Frahm, J. M. 2014. Patchmatch based joint view selection and depthmap estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1510–1517.