Hybrid-MVS: Robust Multi-View Reconstruction with Hybrid Optimization of Visual and Depth Cues

Chenchen Li, Liyang Zhou, Hanqing Jiang, Zhuang Zhang, Xiaojun Xiang, Han Sun, Qing Luan, Hujun Bao *Member, IEEE*, Guofeng Zhang *Member, IEEE*

***Abstract*—Consumer-level RGB-D cameras have been widely used for dense 3D reconstruction of scenes. Especially for textureless or non-lambertian surfaces, consumer RGB-D cameras can ensure completeness of the reconstructed models at a low cost. However, the reconstruction quality relies heavily on the accuracy of the depth sensors. Digital cameras are also used popularly for capturing high-resolution pictures to achieve high-quality dense reconstruction of the scenes, but cannot handle textureless or non-lambertian regions well due to the visual ambiguity problem. To ensure both completeness and accuracy of the reconstructed 3D models, we propose a hybrid multi-view reconstruction pipeline named Hybrid-MVS, which combines the high-resolution images taken by a digital camera and the low-resolution RGB-D frames captured by a consumer RGB-D camera for robust reconstruction of complicated scenes with challenging textureless and non-lambertian surfaces. Unlike most existing multi-sensor systems which require explicit hardware calibration and synchronization of various sensors, the calibration and synchronization problems between the digital camera and RGB-D camera are implicitly solved for compositing reliable depth prior of the digital images in our pipeline. Especially, we propose a hybrid MVS framework for robust PatchMatch stereo and Delaunay meshing, which tightly couples both visual cues given by the digital images and depth cues from the RGB-D frames to maximize the complementary advantages. The experiments with quantitative and qualitative evaluations demonstrate the effectiveness of the proposed Hybrid-MVS framework, which can successfully achieve high-quality 3D reconstruction of complicated natural scenes with robustness to weakly textured and non-lambertian areas.**

***Index Terms*—multi-view stereo, visual cues, depth cues, hybrid MVS.**

I. Introduction

3D reconstruction of large-scale scenes has attracted more and more attentions due to its usefulness in providing 3D digital content creation in a wide range of applications such as scene understanding, 3D navigation, virtual reality (VR) and augmented reality (AR). Commercial 3D scanners have been invented to ensure high-quality 3D reconstruction, but most commercial 3D scanners are too expensive to be popular to non-professional users because of their high costs. Recently, hand-held scanners are more widely spread with consumer-level RGB-D cameras equipped, but the reconstruction quality and scale degrades with the low-cost depth sensors. Meanwhile, with the popularization of digital cameras, it is becoming convenient to capture large numbers of high-resolution photos of the real scenes. To achieve high-quality 3D reconstruction with accurate geometric details, a more feasible way is to perform multi-view stereo (MVS) [6] on the captured high-resolution images, which has long been a research topic of vital importance in computer vision and photogrammetry. However, complete reconstruction with high accuracy relies heavily on sufficient features extracted from the images. Even with high-resolution images, scenes with textureless or non-lambertian surfaces still bring great challenges for MVS methods to achieve both completeness and accuracy due to the visual ambiguity problem.

To ensure high-quality reconstruction of large-scale scenes with both completeness and accuracy, we propose a hybrid multi-view reconstruction system named Hybrid-MVS. Our system combines a digital camera and a consumer RGB-D camera to achieve robust reconstruction of challenging scenes with textureless or non-lambertian surfaces. A PatchMatch-based depth estimation method is applied to estimate high-quality depth maps for the high-resolution images taken by the digital camera, while the low-resolution RGB-D frames captured by the consumer RGB-D camera are used to composite depth prior for the PatchMatch stereo. The combination of the digital camera and the consumer RGB-D camera is much easier to implement than the currently existing multi-camera arrays [7], [8] or multiple in-place rotated unsynchronized RGB-D cameras mounted on a robot platform [9]. However, for the purpose of making full use of the depth measurements for MVS, explicit calibration and synchronization of the two devices are still inevitable. The calibration problem can be solved using Camera Calibration Toolbox [1], but requires the two devices to be rigidly fixed, which is not so friendly to non-professional users. The hardware synchronization of the two devices is a more difficult task even for professional and technical users. To avoid the problems above, our system provides a hybrid RGB-D alignment module which allows users to individually capture multi-view digital images and RGB-D frames of the scene, and automatically aligns them together using an incremental Structure-from-Motion (SfM) strategy. Moreover, unlike existing RGB-D reconstruction approaches based on geometry refinement such as [10], depth

Chenchen Li, Liyang Zhou, Hanqing Jiang, Zhuang Zhang, Xiaojun Xiang, Han Sun, and Qing Luan are with SenseTime Research. E-mails:{lichenchen,zhouliyang,jianghanqing,zhangzhuang,xiangxiaojun,sunhan,luanqing}@sensetime.com.

Hujun Bao and Guofeng Zhang are with the State Key Lab of CAD&CG, Zhejiang University. E-mails: {baohujun,zhangguofeng}@zju.edu.cn.

Corresponding Author: Guofeng Zhang.

Chenchen Li, Liyang Zhou, and Hanqing Jiang assert equal contribution and joint first authorship.

[1]http://www.vision.caltech.edu/bouguetj/calib_doc/

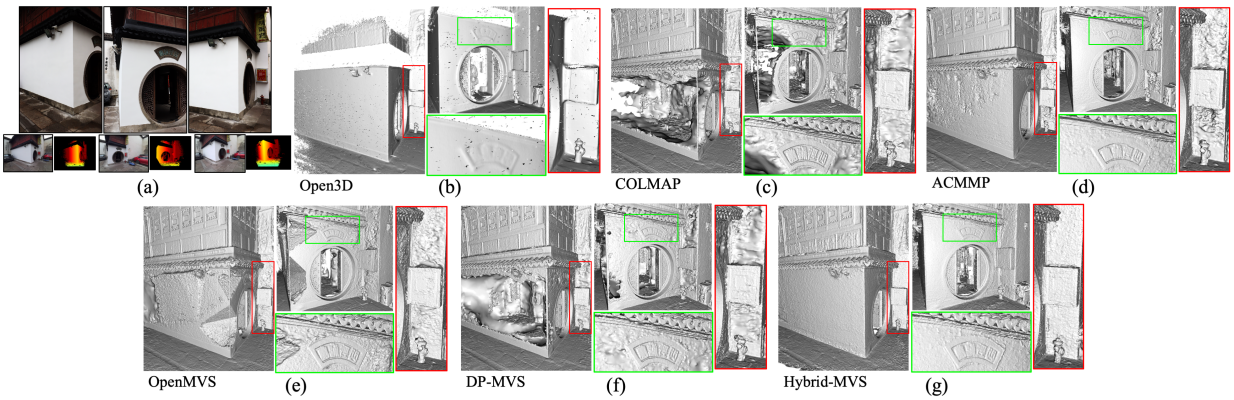


Fig. 1. 3D reconstruction of the famous "Hu Qingyutang" Pharmacy since 1874, captured by a digital camera and an RGB-D camera. (a) Three representative high-resolution images from the digital camera and RGB-D frames from the RGB-D camera. (b) The 3D model reconstructed by fusing depth measurements from the RGB-D camera with Open3D [1], which lacks geometric details. (c) (d) (e) and (f) are the 3D models reconstructed by COLMAP [2], ACMMP [3], OpenMVS [4] and DP-MVS [5] respectively on digital images and color channel of RGB-D frames, which preserve details but lose completeness on textureless walls. (g) The 3D model reconstructed by our Hybrid-MVS taking both the digital images and RGB-D frames as input, which performs the best in accurate geometric details and reconstruction completeness.

enhancement by RGB-D fusion like [11], or 3D panorama construction by [9], we propose hybrid PatchMatch stereo and Delaunay triangulation, which tightly couples both visual cues and depth cues to maximize their complementary advantages. A reconstruction example of the famous cultural relics "Hu Qingyutang" Pharmacy is shown in Fig. 1 which demonstrate the better performance of our hybrid MVS framework in both reconstruction completeness in textureless regions and accurate geometric detail preserving in textured regions, compared to the state-of-the-art (SOTA) MVS methods like Open3D [1], COLMAP [2], ACMMP [3], OpenMVS [4] and DP-MVS [5].

To summarize, our Hybrid-MVS system makes the following main contributions:

- We propose a hybrid multi-view reconstruction pipeline, which combines a digital camera and a consumer RGB-D camera to ensure complete and accurate reconstruction robust to challenging scenes with texturelessness and non-lambertian reflection.
- Unlike other RGB-D fusion methods, the proposed hybrid MVS approach does not require explicit synchronization and manual calibration of the two devices. It innovatively allows users to individually capture high-resolution digital images and RGB-D frames, and aligns the two kinds of data together automatically to avoid human intervention.
- We propose a hybrid PatchMatch stereo approach, which tightly couples both visual cues and depth cues for more robust depth estimation with better completeness and accuracy on textureless and non-lambertian surfaces.
- A novel Delaunay triangulation approach that combines the fused point clouds from digital images and the depth measurements captured by the RGB-D camera to generate a complete and accurate surface mesh for textureless and non-lambertian regions.

This paper is organized as follows. Section II briefly presents related work. Section III gives an overview of the proposed Hybrid-MVS system. The hybrid RGB-D alignment and depth prior composition modules are described in section IV and section V respectively. Section VI describes the hybrid PatchMatch stereo framework, and section VII gives the hybrid Delaunay meshing module. Finally, we evaluate the proposed Hybrid-MVS pipeline in section VIII.

II. Related Work

Most existing MVS approaches reconstruct 3D models from multi-view images or video. According to the taxonomy given in [12], we can divide these visual based MVS methods into four categories: voxel based, surface evolution based, feature growing based, and depth map merging based methods. The **voxel based methods** extract a photo-consistent surface by voxel coloring framework [13], graph-cut optimization [14], or adaptive subdivision of multi-resolution 3D volume by Sinha et al. [15]. These methods are limited by the voxel grid resolution that vitally affects the reconstruction accuracy and space, and they usually assume the compact objects have a tight enclosing bounding box. The **surface evolution based methods** such as [16]–[18] first make an initial surface guess which is then evolved iteratively by minimizing the photo-consistency measurement with variational mesh refinement. A common drawback of the surface evolution based methods is the difficulty to find a reliable initial surface, especially for complicated scenes. Also, the surface evolution is easily mislead by the ambiguous or mistaken photo-consistency caused by textureless or non-lambertian areas. The **feature growing based methods** like [19]–[21] firstly reconstruct 3D feature points from regions with textures, and then expand these feature points to textureless areas, by a region growing process for features matched by per-view and per-pixel image selection. However, these feature point growing methods still suffer from the complicated computational complexity for large-scale scenes. Additionally, the feature expansion is limited to textured regions, leading to the difficulty in handling textureless areas. The **depth map merging based methods** such as [2], [22]–[28] estimate a depth map for each view, after which all the depth maps are merged together to generate one single model, with visibility taken into account. However,

these methods usually face the problem of computational efficiency for high-quality estimation. Zhang et al. [29] proposed a real-time design for accelerating a binocular local matching algorithm by parallel computing. Recently, some works such as [30]–[32] leverage the time complexity of multi-view stereo problem by adopting semi-global matching (SGM) with hardware-efficient parallelism or pipelined architecture. Some learning-based MVS reconstruction approaches such as [33]–[39] are still limited in robustness and generalization for natural scenes, although having made significant achievements on a variety of benchmarks. Generally, scenes with textureless or non-lambertian surfaces still remains to be great challenges for these visual based MVS methods to achieve both completeness and accuracy.

With the development of consumer-level RGB-D cameras such as Microsoft Kinect and Intel RealSense, some methods incorporate depths with intensities to improve the reconstruction robustness and accuracy. These methods can also be divided into three categories: depth fusion based methods, depth enhancement based methods, and geometric refinement based methods. The **depth fusion based methods** such as KinectFusion [40], BundleFusion [41] and InfiniTAM [42] use Iterative Closest Point (ICP) [43] to simultaneously track poses of the input depth maps, and fuse all the tracked depths into a global TSDF model. For example, Xiang et al. [44] present a novel online 3D scanning system on a mobile device, which utilizes adaptive voxel resized TSDF fusion for large object reconstruction. Nevertheless, the reconstruction accuracy of these methods depends heavily on the depth quality of the RGB-D camera used as input. The **depth enhancement based methods** like [11], [45]–[49] further improve the depth resolution and quality of consumer RGB-D camera. Some works [50]–[54] combine depth maps and images, which is more similar to our hybrid PatchMatch stereo approach. However, these methods perform well based on a basic assumption that depth discontinuities relate strongly to color discontinuities, which is usually contradicted for challenging textureless and non-lambertian cases. The **geometric refinement based methods** like [10], [55]–[57] optimize the existing surface model to refine the reconstruction directly. Most of these methods refine the geometry according to multi-view photo-consistency measurement. Therefore, their optimization results are also easily affected by texturelessness and non-lambertian surfaces. Li et al. [58] extend MVS pipeline by utilizing Lidar data to help depth map estimation as well as mesh extraction, which is more similar to our work. However, it simply uses Lidar prior for PatchMatch initialization which cannot avoid degradation of weakly textured or non-lambertian depths.

To better tackle challenging problems of textureless or non-lambertian regions, some recent efforts have been made to improve robustness and generalization by combining traditional MVS approaches with learning methods. For example, Yang et al. [32] tackle noisy depth estimates produced by SGM in textureless regions, using a lightweight depth refinement network. Yang and Jiang [59] achieve practical 3D reconstruction for textureless scenes, by using deep learning algorithms combined with traditional methods for feature extraction and matching from light pattern augmented images. Stathopoulou et al. [60] leverage semantic priors to PatchMatch-based MVS, which significantly improves the depth map and normal map quality in challenging weakly textured regions. However, even with the efforts above, it still remains to be a challenge for 3D reconstruction of large textureless indoor offices or non-lambertian buildings which commonly occur in urban scenarios, with merely visual information available.

Accurate depth map estimation is a vitally important and challenging problem for a high-quality MVS reconstruction. Recently, PatchMatch stereo methods [2], [61]–[69] have been proved powerful in depth map estimation, which is the key concern of a high-quality MVS reconstruction. Unlike Zheng et al. [62] and COLMAP [2] which use the traditional sequential propagation scheme, Galliani et al. [63] and Xu and Tao [70] utilize checkerboard propagation that significantly improves computational parallelism to practically reduce time complexity. As an extension to the work of [70], ACMM [64] improves the depth quality in textureless areas by incorporating a coarse-to-fine scheme. ACMP [67] extends ACMM [64] by introducing the planar compatibility to the matching cost aggregation in addition to photometric consistency. ACMMP [3] further extends both ACMM [64] and ACMP [67] by taking both multi-scale geometric consistency and planar prior into consideration, and improves the discrimination of visual ambiguity greatly. TAPA-MVS [65] proposed novel hypotheses that expand reliable depths from textured regions to neighboring textureless ones. Furthermore, Schönberger et al. [2], Xu and Tao [64] additionally incorporate forward/backward reprojection errors for PatchMatch. Besides, MARMVS [71] select the optimal patch scale for each pixel in addition to reduce matching ambiguities. UniMVSNet [39] propose a novel representation for more generalized depth estimation, which unifies the advantages of both regression and classification networks in a coarse-to-fine framework. However, these methods focus on time efficiency improvement or handling textureless reconstruction with merely visual information, but seldom have any strategy for joint optimization of visual and depth cues for more robust PatchMatch, which is exactly the main focus of our method.

III. System overview

The user first takes M multi-view high-resolution digital images of a scenes with a digital camera, which we denote as $\mathcal{I} = \{I_1, I_2, \cdots, I_M\}$. Then, a consumer-level RGB-D camera is used to take a multi-view RGB-D video stream of the scene. The video stream is decompressed to N low-resolution RGB-D frames denoted as $\mathcal{D} = \{D_1, D_2, \cdots, D_N\}$. The video is required to cover enough scene content of the digital images. In order to ensure this requirement, the acquisition route and orientation of the RGB-D video stream should be as consistent as possible with the digital camera, so that more than 70% of the RGB-D frames are guaranteed to overlap with at least one digital image which shares more than 30% common view with viewing angle no more than $45°$. Here, we assume that the consumer-level RGB-D camera contains the factory intrinsic parameters and relative extrinsic parameters

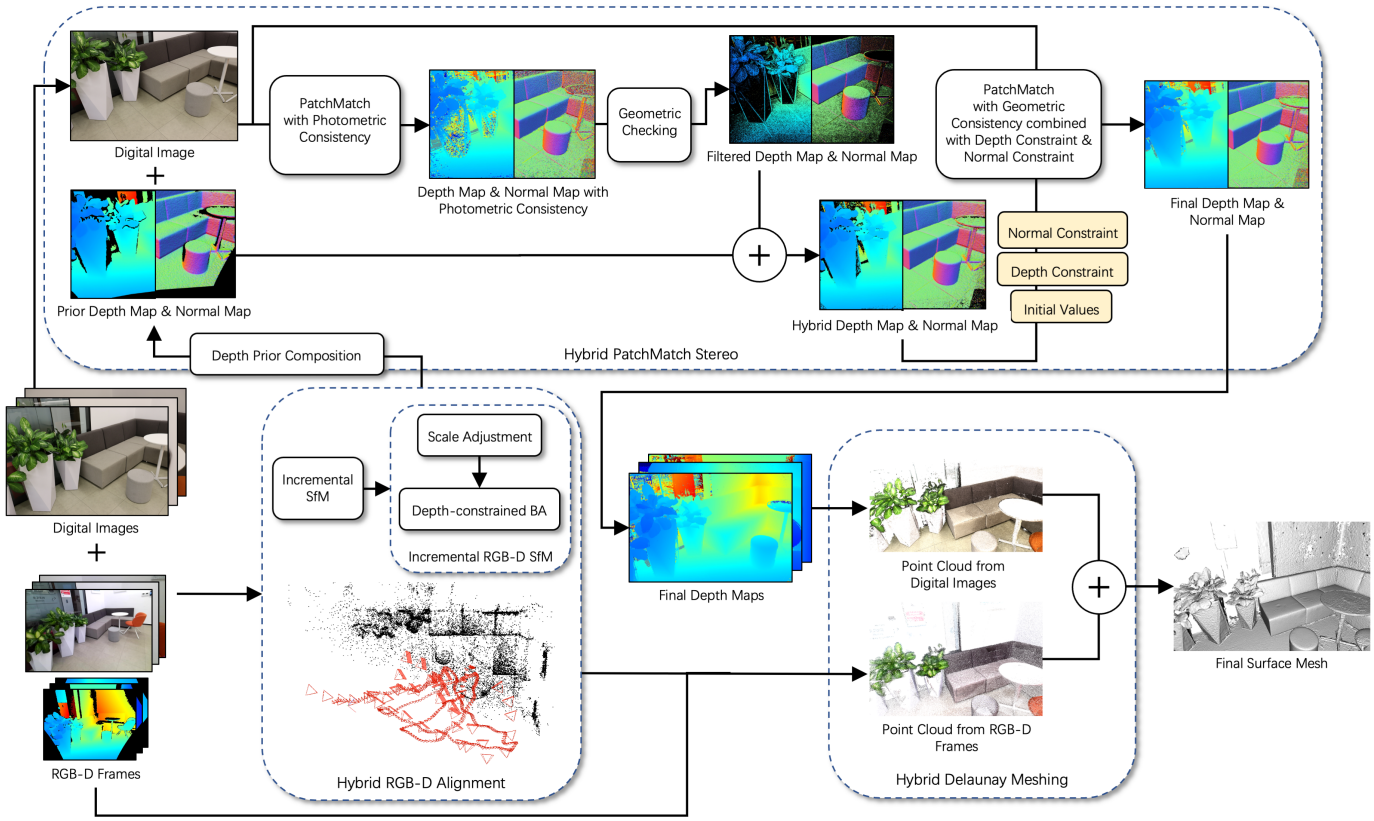


Fig. 2. System overview, which consists of a hybrid RGB-D alignment module for registering digital images and RGB-D frames together, a hybrid PatchMatch stereo module which estimates high-quality depth maps and normal maps by tightly coupling photometric cues and depth cues, and a hybrid Delaunay meshing module applied to separately fused point clouds from digital images and RGB-D frames to generate the final surface mesh.

of the embedded color and depth sensors, based on which the depth measurements are already aligned with the colors. Our Hybrid-MVS system is applied for the input multi-view digital images and RGB-D frames to robustly reconstruct an accurate surface model of the captured scene. The steps of the proposed Hybrid-MVS framework is outlined in Fig. 2. A sparse map is reconstructed first to jointly register the input digital images and RGB-D frames in a unified 3D space with true scale using a depth-constrained incremental SfM with scale recovery. Then, for each digital image, an accurate high-resolution depth map is estimated by a hybrid PatchMatch stereo approach, which combine the prior depth map composited from the low-resolution RGB-D frames with visual cues. After that, all the depth maps of digital images and all the RGB-D frames are fused separately to two point clouds, and a hybrid Delaunay meshing is applied to the two separate point clouds to generate the final surface mesh represented as $\mathbf{S}$. The main steps of our framework will be described in detail in the following subsections.

IV. Hybrid RGB-D Alignment

Our hybrid RGB-D alignment registers the digital images and RGB-D frames together into a unified coordinate system to perform a joint calibration for the intrinsic and extrinsic parameters of the digital images $\mathcal{I}$ and the extrinsic parameters of the RGB-D frames $\mathcal{D}$. The hybrid RGB-D alignment module consists of two major stages. The first stage is a traditional incremental SfM proposed in [72] for digital images to derive their intrinsic and extrinsic parameters. In the second stage, the RGB-D frames are further registered into the SfM result of the first stage, by applying a novel depth-constrained incremental SfM scheme while recovering the true scene scale based on the depth measurements.

Here we give a brief introduction to the ordinary SfM pipeline in the first stage. SfM extracts a set of 3D map points $\mathcal{X} \in \mathbb{R}^3$ of the captured scene with consistently rich features among the multiple input images, while solving the intrinsic and the extrinsic parameters of the input images. For each digital image $I_i \in \mathcal{I}$, we denote its intrinsic and extrinsic parameters as $\mathbf{K}_i$ and $\mathbf{M}_i = [\mathbf{R}_i|\mathbf{t}_i]$ respectively, with $\mathbf{R}_i$ and $\mathbf{t}_i$ the rotation and translation parts of $\mathbf{M}_i$. The intrinsic parameters and extrinsic poses of the registered images are optimized together with the 3D map points by bundle adjustment (BA) [73] that minimizes the reprojection errors defined as:

$$E_{BA} = \sum_{\mathbf{X}_K \in \mathcal{X}} \sum_{\mathbf{x}_i^k} \left\| \pi(\mathbf{M}_i \mathbf{X}_K) - \mathbf{x}_i^k \right\|^2 \quad (1)$$

with $\mathbf{M}_i$ being the the global-to-local transformation of image I_i, and $\mathbf{x}_i^k$ being a 2D feature in I_i corresponding to the 3D map point $\mathbf{X}_K$. $\pi(x, y, z) = (\frac{x}{z} f_u + c_u, \frac{y}{z} f_v + c_v)$ is the projection position function, with (f_u, f_v) the focal lengths in u and v directions, and (c_u, c_v) the optical center. Feature correspondences with large errors are filtered out as outliers

in our system. Fig. 3(c) show incremental SfM result of the first stage, with all the digital images registered successfully.

In the second stage, the RGB-D frames are further registered incrementally into the scene graph reconstructed in the first stage. In our system, we use the vocabulary tree engine [74] to find the most similar digital images for the new RGB-D frame to match features. While registering a new RGB-D frame, the intrinsic parameters and extrinsic poses of the digital images are fixed as constants to ensure that the accuracy of SfM in the first stage won't be affected by the incoming consumer-level RGB-D frames which usually contains color images with much lower quality than the digital images.

To solve the scale recovery problem, we propose a novel incremental SfM scheme for RGB-D frame registration, which not only recover the true scale information, but also improve the registration robustness by incorporating a depth-constrained BA with scale adjustment. Fig. 3(e) shows the result of our depth-constrained incremental SfM after the second stage, which successfully registers the RGB-D frames to the digital images, with true scale also recovered in the meantime. We will give a detailed description of this RGB-D incremental SfM approach in the following two subsections.

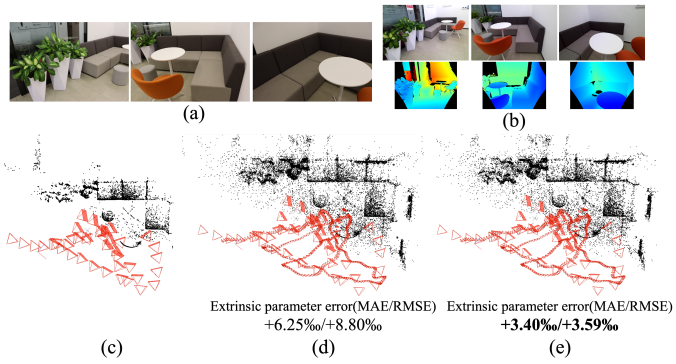


Fig. 3. (a) Three representative input high-resolution images of the "Sofa" case from the digital camera. (b) Three representative low-resolution RGB-D frames of the "Sofa" case from the RGB-D camera. (c) SfM result with only the digital images registered, which is unable to recover the true scale. (d) Hybrid SfM result with digital images (visualized with larger camera focal length) and RGB-D frames (with smaller focal length) registered together by scale adjustment and ordinary BA. (e) Hybrid SfM result with digital images and RGB-D frames registered by the combination of scale adjustment and depth-constrained BA, which optimizes the final intrinsic and extrinsic parameters with better accuracy.

A. Scale adjustment

To refine the scale information, we collect a set of scale candidates by computing the ratios between all the possible depth measurements and their corresponding projection depths on the RGB-D frames according to the registered intrinsic and extrinsic parameters. The scale candidates are defined as:

$$\mathcal{S} = \left\{ \frac{D_i(\pi(\mathbf{M}_i\mathbf{X}_K))}{d(\mathbf{M}_i\mathbf{X}_K)} \mid \mathbf{X}_K \in \mathcal{X}, D_i \in \mathcal{D} \right\}, \tag{2}$$

where $D_i(\pi(\mathbf{M}_i\mathbf{X}_K))$ represents the depth measurement at the projection position of 3D map point $\mathbf{X}_K$ on RGB-D frame D_i, while $d(\mathbf{M}_i\mathbf{X}_K)$ is the projection depth of $\mathbf{X}_K$ on D_i. We ignore the candidate scales for invalid depth measurements , and select the average of the remaining candidate scales as the final scale factor $\int$. The scale adjustment is carried out by scaling the positions of the map points $\mathcal{X}$ and the translation part of the extrinsic poses of all the registered images by $\int$, and applied before each time of BA, only if there are at least 100 scale candidates in our experiments. As can be seen in Fig. 3(c) and (d), the false scale of the ordinary SfM reconstruction in the first stage is successfully recovered to the true size after our scale adjustment strategy.

B. Depth-constrained Bundle Adjustment

As can be seen in Eq. (1), the ordinary BA [73] only considers 2D reprojection errors. Since RGB-D frames $\mathcal{D}$ have an additional depth measurement for each pixel, we extend Eq. (1) to a more robust format by incorporating additional depth projection errors as constraints in the following form:

$$\begin{aligned} E'_{BA} = \sum_{\mathbf{X}_K \in \mathcal{X}} \sum_{\mathbf{x}_i^k} \Big(& \left\| \pi(\mathbf{M}_i\mathbf{X}_K) - \mathbf{x}_i^k \right\|^2 + \\ & \alpha \left\| d(\mathbf{M}_i\mathbf{X}_K) - D_i(\pi(\mathbf{M}_i\mathbf{X}_K)) \right\|^2, \end{aligned} \tag{3}$$

where $\mathbf{M}_i$ is the transformation matrix of RGB-D frame D_i, $D_i(\pi(\mathbf{M}_i\mathbf{X}_K))$ is the depth measurement at the projection position of $\mathbf{X}_K$ on D_i, $d(\mathbf{M}_i\mathbf{X}_K)$ represents the projection depth of $\mathbf{X}_K$ on D_i transformed by $\mathbf{M}_i$, and α is the weight of the depth constraints, which we empirically set to 100 in the experiments. Again, we ignore the depth constraint for invalid depth measurements. Since depth projection errors are used as additional geometric constraints in BA, it is helpful to improving the registration robustness when visual features are insufficient or non-lambertian. Fig. 4 shows the results of the depth-constrained BA compared with ordinary BA for an indoor scene "Corridor" with both textureless and non-lambertian regions. The ordinary BA causes severe pose drift due to insufficient visual features as shown in Fig. 4(a), while the depth-constrained BA leads to a continuous pose trajectory without drift as can be seen in Fig. 4(b). Moreover, we also evaluate the extrinsic parameter accuracies of ordinary BA and depth-constrained BA by comparing the relative errors between all depths estimated by our PatchMatch stereo using the estimated extrinsic poses as in section VI and prior depths composed from reference ToF measurements in section V as ground truth (GT). As can be seen in Fig. 3(d) and (e), our depth-constrained BA is helpful for a more accurate scale adjustment with smaller extrinsic parameter error in Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), since it solves extrinsic parameters for the RGB-D frames with higher depth consistency by minimizing the reprojection errors defined in Eq. (3), so as to successfully suppress depth errors of the scale candidates for a more accurate scale estimation.

After the two stages, our hybrid RGB-D alignment outputs a sparse reconstruction of $\mathcal{I}$ and $\mathcal{D}$, including intrinsic and extrinsic parameters of all the digital images and RGB-D frames with true scale.

V. Depth Prior Composition

After hybrid RGB-D alignment finishes, we composite a novel depth map for each high-resolution digital image $I_i \in \mathcal{I}$

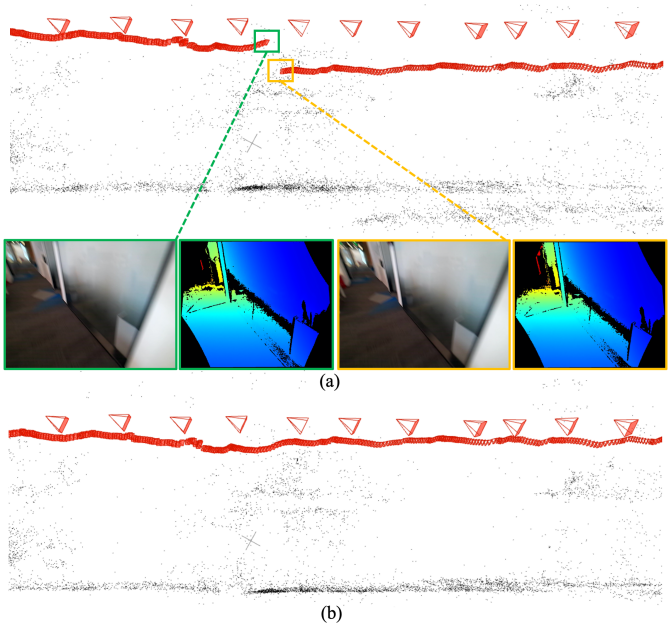


Fig. 4. (a) Incremental SfM result with digital images (with larger camera focal length) and RGB-D frames (with smaller focal length) by ordinary BA for case "Corridor", where two successive RGB-D frames containing textureless and non-lambertian surfaces are shown with severe drifting poses. (b) Our hybrid SfM result by depth-constrained BA without pose drift.

as its depth prior. The novel depth maps are composited by reprojecting the raw depth measurements of $\mathcal{D}$ to the views of the digital images $\mathcal{I}$ according to the registered parameters of the digital images and RGB-D frames. For each digital image I_i, we select a number of reference RGB-D frames from $\mathcal{D}$ for depth reprojection. The selected reference RGB-D frames which are denoted as $\mathcal{R}_d(I_i)$ should have the most similar field-of-view to I_i in order to provide more depth measurements for depth composition. We measure the view similarity between the digital images and the RGB-D frames by computing a score between I_i and each $D_j \in \mathcal{D}$ as:

$$S(I_i, D_j) = \frac{\tau_a - \min\left(\phi(I_i, D_j), \tau_a\right)}{\tau_a} \cdot \frac{\tau_d - \min\left(\delta(I_i, D_j), \tau_d\right)}{\tau_d}, \tag{4}$$

where $\phi(I_i, D_j)$ measures the angle difference between the viewing directions of I_i and D_j, and τ_a is an angle truncation threshold which we set to $60°$. $\delta(I_i, D_j)$ measures the distance between the camera positions of I_i and D_j, and τ_d is the corresponding distance truncation threshold which is assigned as the 10% smallest distance of $\{\delta(I_i, D_k) | D_k \in \mathcal{D}\}$ in our experiments. The largest scored three RGB-D frames are chosen as the reference frame set $\mathcal{R}_d(I_i)$.

Once the reference RGB-D frames are selected, the raw depth measurements in $\mathcal{R}_d(I_i)$ are reprojected to the target image I_i to composite the novel depth map $\hat{D}_i$. Since the resolutions, intrinsic and extrinsic parameters of I_i and $D_j \in \mathcal{R}_d(I_i)$ may differ widely, a mesh rendering scheme is adopted for depth reprojection to ensure completeness of the composite depth map. For each depth $d \in D_j$ at pixel $\mathbf{x} = (u, v)$, we project it back to 3D space get a global 3D point by $\mathbf{P} = {\mathbf{M}_j}^{-1} \rho(u, v, d)$, where $\mathbf{M}_j$ is the global-to-local transformation of frame D_j, and $\rho(u, v, d) = (\frac{u - c_u}{f_u} d, \frac{v - c_v}{f_v} d, d)$ is the back projection function. The projected 3D space points are triangulated by the image pixel grid connection to build a 3D surface mesh. We disconnect neighboring pixel depths across depth boundaries to better handle occlusions caused by depth discontinuity, with neighboring depth difference over 3% of the absolute depth value as depth boundary. The 3D surface mesh is rendered by the intrinsic parameters and extrinsic pose of I_i to get a depth buffer in the target view with rendered per-pixel depths as candidate depths for I_i. Since there are more than one reference frames in $\mathcal{R}_d(I_i)$ reprojected, each pixel in the digital image I_i might have a stack of several candidate depths, which are averaged to get a final depth. All the final depths composite a novel high-resolution depth map for I_i, which is denoted as $\hat{D}_i$. Meanwhile, we compute a corresponding normal map $\hat{N}_i$ from the prior depth map $\hat{D}_i$. Fig. 6(b) shows the composited depth maps and normal maps of representative digital images from cases "Sofa" and "Horse Head". The novel high-resolution depth maps and normal maps will be used as prior for our hybrid PatchMatch stereo module in the following section.

VI. Hybrid PatchMatch stereo

Although we already have a complete depth map for each digital image, the composite depth maps turn out to be poor in depth details and contain some noise due to the low-resolution RGB-D frames from the consumer-level RGB-D camera, as demonstrated in Fig. 5(b). To achieve a high-quality depth map estimation with more accurate geometric details for each digital image, a diffusion-like propagation based PatchMatch scheme is applied, with the composite depth map as prior. Existing sequential propagation strategies like COLMAP [2] and OpenMVS [4] usually can produce depth maps with better geometric details than the prior depth measurements from the consumer RGB-D camera, but lose depth completeness in textureless or non-lambertian areas, as can be seen in Fig. 5(c) and (e). ACMMP [3] applies a coarse-to-fine checkerboard-based propagation strategy to better handle textureless areas, but still cannot ensure depth accuracy and completeness for textureless or non-lambertian regions, as shown in Fig. 5(d). To thoroughly settle the challenging problems of textureless or non-lambertian regions for MVS, we propose a robust hybrid PatchMatch stereo approach, which fully combines both the visual cues and the prior depths of the high-resolution digital images to achieve better reconstruction completeness and geometric accuracy for textured, textureless and non-lambertian surfaces, as can be seen from our final depth estimation results in Fig. 5(f). To better visualize the detailed improvement of depth results, the multi-view depth maps by different methods are fused to 3D surface models using our Delaunay meshing method described in section VII. We quatitatively evaluate the geometric completeness and accuracy on MAE and RMSE (in millimeters) of the fused models for "Horse Head" by comparing them to the GT model. As the evaluation shown in Fig. 5, our hybrid PatchMatch stereo performs the best in both accuracy and completeness.

A detailed pipeline of our hybrid PatchMatch method can be seen in the "Hybrid PatchMatch Stereo" module of Fig.

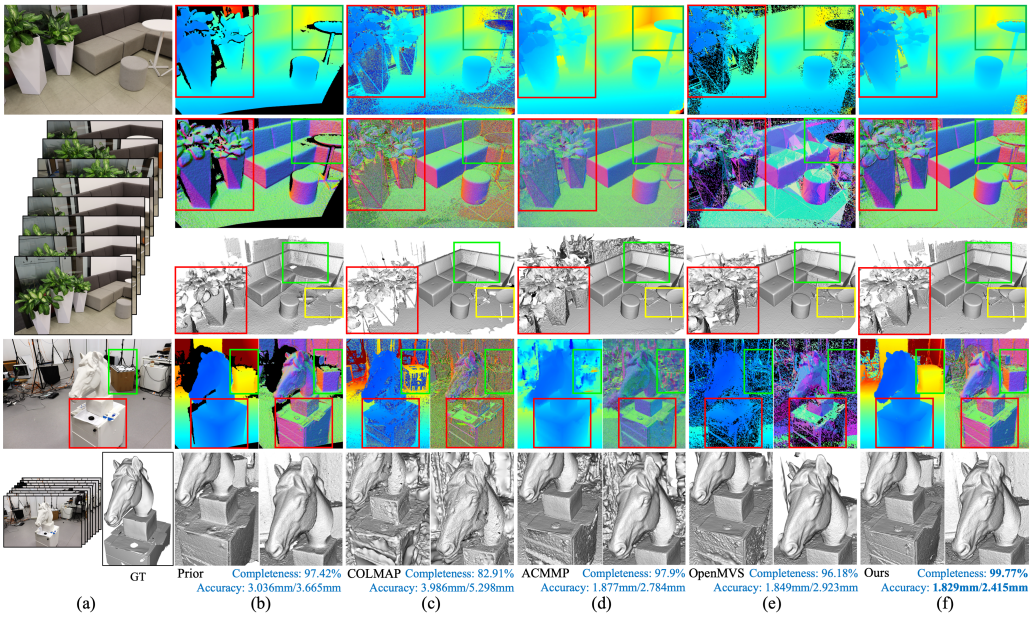


Fig. 5. (a) A representative digital image and its reference images for each case of "Sofa" and "Horse Head", with the GT model of "Horse Head" given. (b) The prior depth maps and normal maps of (a). (c) The depth maps and normal maps of (a) estimated by COLMAP [2]. (d) The depth maps and normal maps by ACMMP [3]. (e) The depth maps and normal maps by OpenMVS [4]. (f) The depth and normal results by our hybrid PatchMatch stereo. The depth maps by different methods are fused to 3D surface models using our Delaunay meshing method. We further evaluate the geometric completeness and accuracy on MAE/RMSE of the fused models for "Horse Head" by comparing them to the GT model, to show the effectiveness of our hybrid PatchMatch in better reconstruction accuracy and completeness.

2, which takes each digital image with its low-quality prior depth map and a set of reference digital images as input, and produces a high-quality depth map and normal map for the current image, via a diffusion-like hybrid PatchMatch scheme which tightly couples the photometric cues and depth priors. Specifically, photometric consistency depth propagation proposed by DP-MVS [5] is performed firstly to generate the initial depth maps of digital images, whose erroneous initial depths and normals are then filtered with geometric consistency check. After that, each filtered depth map is filled up with the prior depths to generate a hybrid depth map. Similarly, the filtered normal map is also combined with the prior normals to get a hybrid normal map. Finally, geometric consistency guided PatchMatch is applied to acquire the final high-quality depth map and normal map, which fully uses the hybrid depth map and normal map as geometric constraints for the propagation process, to ensure both depth completeness and geometric accuracy. The entire PatchMatch stereo process is summarized in Algorithm 1. More details will be given in the following subsections to demonstrate the effectiveness of this hybrid PatchMatch stereo approach.

Algorithm 1 Hybrid PatchMatch Stereo

Input: Digital image I_i, prior depth map $\hat{D}_i$ and normal map $\hat{N}_i$
Output: Optimized depth map D^* and normal map N^*

1: Photometric consistency depth propagation proposed by DP-MVS [5] to obtain initial depth map and normal map for I_i
2: Geometric consistency check to filter erroneous initial depths and normals for I_i
3: Fill up the filtered depth map with prior $\hat{D}_i$ to get hybrid D_i^h
4: Fill up the filtered normal map with prior $\hat{N}_i$ to get hybrid N_i^h
5: Geometric consistency guided PatchMatch with depth and normal constraints by D_i^h and N_i^h with Eq. (5) to acquire the optimized depth map D_i^* and normal map N_i^* for I_i

A. Hybrid depth map and normal map initialization

The prior depth map and normal map is helpful to constrain the depth propagation. However the prior depth map might contain some noisy depths or lose details due to the low-quality consumer RGB-D camera, which might to degrade the depth quality. Considering the depth map generated with photometric consistency measurement is more accurate and more geometric details especially in the textured regions, we combine the estimated depth map with the prior depth map to generate a complete depth map, which we call hybrid depth map and denote as D_i^h. The hybrid depth map is favorable to maximize their complementary advantages by tightly coupling the visual cues from the digital camera and depth cues from the RGB-D camera. We use multi-view geometric consistency check to filter out the errors of the initial depth map caused by texturelessness and non-lambertian reflection. Specifically, for each pixel $\mathbf{x} = (u, v)$ in I_i with its initial depth d estimated by PatchMatch, we check its geometric consistency in a reference image $I_j \in \mathcal{R}(I_i)$, by projecting it to I_j to get its projection position by $\mathbf{x}_j = \pi(\mathbf{M}_j\mathbf{M}_i^1\rho(u, v, d))$ with $\rho(\cdot)$ the back projection function and $\pi(\cdot)$ the projection position function, and projecting $\mathbf{x}_j$ back to I_i to get a reprojection position $\hat{\mathbf{x}}_j$ and its reprojection error $\|\hat{\mathbf{x}}_j - \mathbf{x}\|$. If the average reprojection error over all the reference images $\mathcal{R}(I_i)$ exceeds 3 pixels, we consider the pixel $\mathbf{x}$ geometrically inconsistent and filter its depth and normal out. Since the estimated depth map and the prior depth map are in consistent 3D scale, we use the

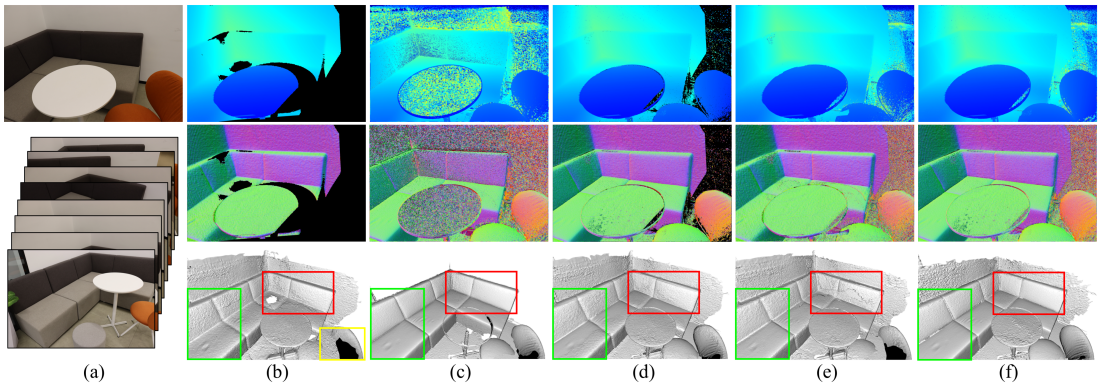


Fig. 6. Examplar effect of hybrid depth map and normal map initialization. (a) A representative digital image and its reference images for case "Sofa". (b) The prior depth map and normal map of (a). (c) The depth map and normal map by DP-MVS PatchMatch [5] with only photometric consistency costs. (d) The hybrid depth map and normal map of (a). (e) The depth map and normal map estimated by our hybrid PatchMatch with (b) as initial constraint. (f) The final depth map and normal map by our hybrid PatchMatch with (d) as constraint. The depth map results are fused to 3D surface models using our Delaunay meshing method to show the effectiveness of our hybrid combination strategy in mesh details.

prior depths from $\hat{D}_i$ to fill up the filtered depths to ensure a complete depth map for I_i. The same completion strategy is applied to the filtered normals to get a complete hybrid normal map denote as N_i^h. An exemplar hybrid depth map and normal map are shown in Fig. 6(d), which turn out to have better geometric details preserved than the prior depth map in Fig. 6(b), and fewer depth and normal errors than the PatchMatch result in Fig. 6(c). We also verify the usefulness of this hybrid combination strategy by comparing the final hybrid PatchMatch result using the hybrid depth map and normal map as constraints for geometric consistency guided propagation in Fig. 6(f) and the corresponding result with only prior depths and normals as constraints in Fig. 6(e), from which we can see that our hybrid combination strategy is helpful to a better recovery of the accurate geometric details in the final depth map if there are geometric details lost or noisy depths in the low-quality prior depth maps.

B. PatchMatch with hybrid depth and normal constraints

Now we present our hybrid PatchMatch propagation for I_i with geometric consistency as guidance, which incorporates the hybrid depth map D_i^h and the hybrid normal map N_i^h as constraints. A simple way is to use the hybrid depth map and hybrid normal map as reliable initial values for PatchMatch propagation. We illustrate the result of this simple strategy in Fig. 8(c), which shows its favorable improvement in depth and normal accuracy of the textureless regions compared with the PatchMatch result of DP-MVS [5] in Fig. 8(b), as highlighted in the rectangles. However, this strategy is not robust enough to avoid degeneration caused by textureless or non-lambertian surfaces during PatchMatch propagation. Therefore, we propose to further strengthen the guiding effect of the hybrid depth map D_i^h and the hybrid normal map N_i^h during the propagation process. Specifically, for each hypothesis $\mathbf{x}_p$ in I_i, we add a depth weight and a normal weight to its accumulated multi-view cost as follows:

$$\begin{aligned} w_d(\mathbf{x}_p) &= 1 - \lambda_d exp\left(-\frac{(D_i(\mathbf{x}_p) - D_i^h(\mathbf{x}_p))^2}{\sigma^2}\right) \\ w_n(\mathbf{x}_p) &= 1 - \lambda_n exp\left(-\frac{(1 - \langle N_i(\mathbf{x}_p) \cdot N_i^h(\mathbf{x}_p) \rangle)^2}{\sigma^2}\right) \\ C^*(\mathbf{x}_p) &= w_d(\mathbf{x}_p) w_n(\mathbf{x}_p) C(\mathbf{x}_p), \end{aligned} \tag{5}$$

where $C^*(\mathbf{x}_p)$ is the weighted cost of the ordinary multi-view cost $C(\mathbf{x}_p)$. λ_d and λ_n are balance factors for depth weight and normal weight respectively, which is set to 0.4 and 0.53 in the experiments to make a trade-off between the prior and multi-view cost. $\sigma = 1 - \lambda_c exp\left(-|I_i^v(\mathbf{x}_p)|\right)$ is a coefficiency proportional to the color covariance $I_i^v(\mathbf{x}_p)$ computed from the 11×11 image patch centered at $\mathbf{x}_p$, where λ_c is the weighted term of $I_i^v(\mathbf{x}_p)$, and is set to 0.8 in the experiments. Intuitively, the smaller σ is, the larger influence the depth and normal weights impose on the accumulated multi-view cost to guide the optimization towards the prior depth and normal, which is exactly the situation of textureless regions. In contrast, for textured regions, the photometric cost is strong enough, so that there is no special need for guiding weights. Therefore, a larger σ suppresses the weight influence on the multi-view cost. Here $D_i(\mathbf{x}_p)$ and $N_i(\mathbf{x}_p)$ represents the currently updated depth and normal values of $\mathbf{x}_p$ during the propagation. The weighted multi-view costs are used for PatchMatch propagation instead of the ordinary ones, to achieve a high-quality depth map and normal map estimation for each digital image I_i, which we denote as D_i^* and N_i^*. For textureless or non-lambertian regions, the photometric costs for various depths and normals are non-discriminative, resulting in noisy depths and normals. In comparison, this depth and normal weighting cost scheme helps to eliminate the depth and normal ambiguity, leading to noise-free depth maps and normal maps. Specifically, if the depth and normal of a hypothesis are close to the prior depth and normal, the weighted cost decreases to a small value, which encourages the propagation to choose this depth and normal hypothesis. In this way, the depth and normal optimization trends not too far away from the hybrid depths and normals. For textured lambertian regions, the weight influence is suppressed, so that

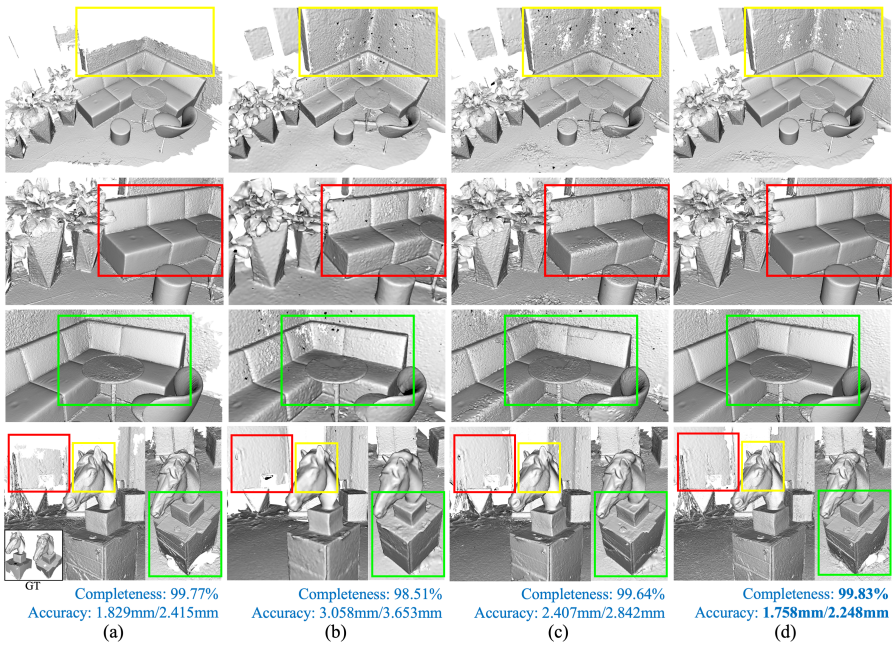


Fig. 7. (a) The surface meshes of cases "Sofa" and "Horse Head" extracted from the fused point cloud of digital images, with GT model of "Horse Head" given. (b) The surface meshes extracted from the fused point cloud of RGB-D frames. (c) The surface meshes extracted from the combination of point clouds from both the digital images and the RGB-D frames. (d) The surface meshes extracted by our hybrid Delaunay meshing of point clouds from both the digital images and the RGB-D frames. We evaluate geometric completeness and accuracy on MAE/RMSE for all the surface meshes of "Horse Head" compared to its GT model to show the effectiveness of our hybrid Delaunay meshing in preserving geometric details and completeness.

the discriminative costs from high-resolution digital images are preserved as much as possible to recover the depth and normal details.

VII. Hybrid Delaunay Meshing

After we have an optimized high-quality depth map D_i^* and normal map N_i^* for each digital image, all the depths are projected back to 3D space to get a set of 3D points with normals. The 3D points of all the digital images are fused together to a point cloud we denote as $\mathbf{P}_c$. Each 3D point inside $\mathbf{P}_c$ contains the set of digital images where it has been triangulated and visible. The depth map fusion is carried out in a similar way to the graph-based framework proposed in [2]. According to the depth and geometry consistency among different image views, the consistent pixels are connected recursively during the fusion process. Meanwhile, the depth maps $\{D_j \in \mathcal{D}\}$ of all the RGB-D frames captured by the RGB-D camera are also fused in the same way to a point cloud denoted as $\mathbf{P}_d$. Both $\mathbf{P}_c$ and $\mathbf{P}_d$ participate in a Delaunay meshing method to reconstruct the final surface mesh of the scene.

A detail-aware 3D Delaunay triangulation similar to DP-MVS [5] is applied to the point clouds to build tetrahedra $\mathbf{T}$, which are then labeled inside or outside the surface through an graph-cuts based energy minimization. Let $\mathbf{L}$ denote the labeling. For each tetrahedron $\tau \in \mathbf{T}$, we have $\mathbf{L}(\tau) \in \{inner, outer\}$. Let $\mathbf{F}$ denote the set of facets shared by two neighboring tetrahedra. We follow the energy function defined in [5] for inside/outside labeling problem of tetrahedra $\mathbf{T}$ as:

$$E(\mathbf{T},\mathbf{F},\mathbf{L}) = \sum_{\tau\in\mathbf{T}} E_d(\tau,\mathbf{L}(\tau)) + \sum_{f\in\mathbf{F}} E_s(f,\mathbf{L}(\tau),\mathbf{L}(\tau')), \tag{6}$$

where E_d is the data term for tetrahedron τ, and E_s is the smooth term for triangular facet f shared by each pair of neighboring tetrahedra (τ,τ'). The facets shared by neighboring tetrahedra with different labels compose the final surface mesh.

For our case, a straightforward way is to simply combine $\mathbf{P}_c$ and $\mathbf{P}_d$ as input for Delaunay triangulation. However, the point cloud $\mathbf{P}_c$ from digital images usually have more accurate structures and more geometric details than $\mathbf{P}_d$ from the consumer RGB-D camera, while $\mathbf{P}_d$ usually has higher completeness of the scene than $\mathbf{P}_c$ due to its higher frame frequency, as can be seen in Fig. 7(a) and (b). Besides, there might exist misalignment between $\mathbf{P}_c$ and $\mathbf{P}_d$. Therefore, straightforward combination of $\mathbf{P}_c$ and $\mathbf{P}_d$ might cause artifacts and more geometric noise in the final surface mesh, as shown in Fig. 7(c). To better preserve the geometric accuracy and the reconstruction completeness of the final surface model, $\mathbf{P}_c$ should have higher priority than $\mathbf{P}_d$ during the Delaunay meshing process. We propose to extend the meshing strategy proposed in [5], by accumulating the data terms and smooth terms for the shooting line of sight of each 3D point $P \in \mathbf{P}_c \cup \mathbf{P}_d$ as follows:

$$\begin{aligned} &E_d(\tau_v, inner) && += \alpha_v(P) \\ &E_d(\tau_p, outer) && += \alpha_v(P) \\ &E_s(f_i, inner, outer) && += \omega_{f_i}\alpha_v(P) \\ &E_s(f_i, outer, inner) && += \omega_{f_i}\alpha_v(P) \\ &\omega_{f_i} = \omega_h(f_i)\omega_d(f_i)\omega_v(f_i)\omega_q(f_i), \end{aligned} \tag{7}$$

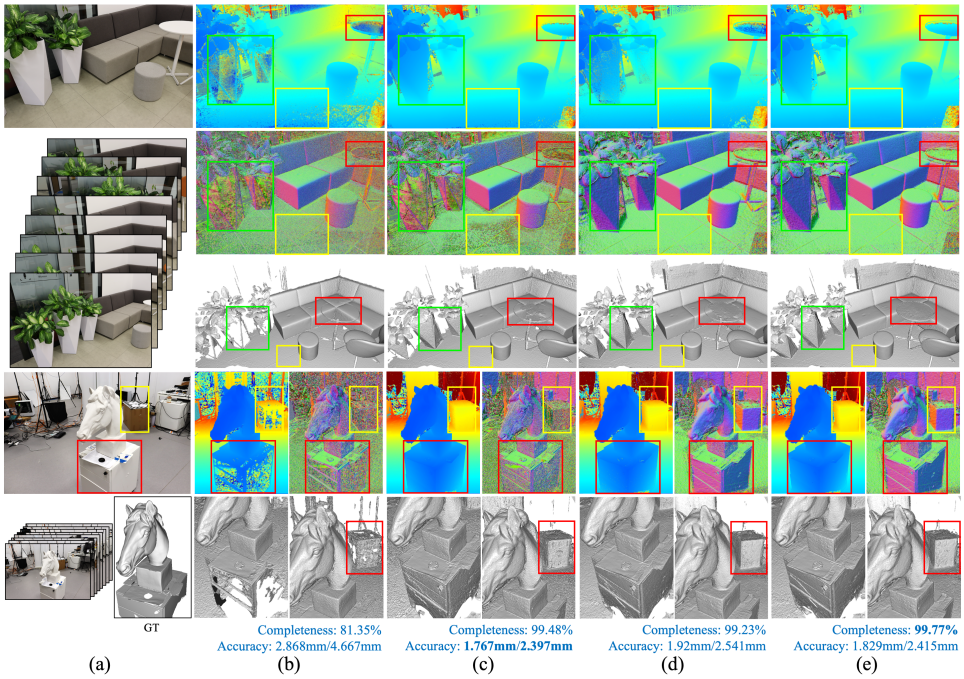


Fig. 8. Examplar effects of depth and normal constraints on hybrid PatchMatch stereo. (a) A representative digital image and its reference images for each case of "Sofa" and "Horse Head", and GT model of "Horse Head". (b) The depth maps and normal maps estimated by DP-MVS [5]. (c) The estimated depth maps and normal maps with only the hybrid depth map as constraint. (d) The depth maps and normal maps with only the hybrid normal map as constraint. (e) The final depth maps and normal maps by our hybrid PatchMatch stereo with both depth and normal constraints. All the depth map results are fused to 3D surface models using our Delaunay meshing method, with the geometric completeness and accuracy on MAE/RMSE of the fused models evaluated for "Horse Head" by comparison to the GT model to show the effectiveness of depth and normal constraints in completeness and mesh details.

where τ_v is the tetrahedron that the center of a camera C_i lies in, and τ_p is the tetrahedron that contains P and intersects with the extended line of sight from C_i. $\alpha_v(P)$ is the unity confidence value for each ray from C_i to P, which is set differently for 3D points from $\mathbf{P}_c$ and $\mathbf{P}_d$. For each 3D point $P \in \mathbf{P}_c$, we set $\alpha_v(P)$ to a higher value 1 in the experiments, while for each $P \in \mathbf{P}_d$ we set $\alpha_v(P)$ to a smaller confidence 0.2, so as to keep the 3D points produced by our hybrid PatchMatch stereo dominant in order to preserve the geometric details not to be degraded by the low-quality depth measurements from the consumer RGB-D camera. ω_v, ω_q, ω_d and ω_h denote the visibility weight, quality weight, density weight and hybrid weight respectively for smooth term penalty. Previous methods like [75] only use ω_v and ω_q, which has limitation in preserving geometric details. DP-MVS [5] uses ω_d to enforce the accuracy of surface geometry. In comparison, we propose this novel hybrid weight ω_h to account for the different vertex combinations of digital images and RGB-D frames. We examine the three vertices of each tetrahedron's facet f_i to determine $\omega_h(f_i)$. Facets with three vertices all from $\mathbf{P}_c$ are more preferred with a small hybrid weight $\omega_h(f_i) = \tau_c$, with higher probability to composite the final surface with better geometric details. Facets with three vertices all from $\mathbf{P}_d$ have a larger hybrid weight $\omega_h(f_i) = \tau_d$, with lower probability to become the final surface. Facets with vertices from both $\mathbf{P}_c$ and $\mathbf{P}_d$ are considered not so confident, and are therefore penalized with the largest hybrid weight $\omega_h(f_i) = \tau_h$ with highest probability to disappear to avoid noise interference. According to the quality difference between $\mathbf{P}_c$ and $\mathbf{P}_d$, we empirically set the parameters $\tau_c = 0.4, \tau_d = 0.8, \tau_h = 1.0$ in our experiments. Considering that the point cloud accuracy from consumer RGB-D camera is generally lower than that from digital images, these coefficiency settings are general for common cases. In this way, the 3D points from $\mathbf{P}_c$ and $\mathbf{P}_d$ can be coherently merged to avoid possible reconstruction noise caused by misalignment while keeping the reconstruction accuracy and completeness in the final surface model.

Fig. 7(d) gives the final surface mesh produced by our hybrid Delaunay meshing approach. Compared to the result of the simple combination strategy shown in Fig. 7(c), it can be seen that the proposed hybrid meshing faithfully preserve better geometric accuracy and reconstruction completeness, by taking the complementary advantages of both point clouds from the digital images and the RGB-D frames. For case "Horse Head", we also quantitatively compare the geometric completeness and accuracy on MAE and RMSE of the surface meshes, which demonstrates that our method performs the best in both completeness and accuracy.

VIII. Experiments

In this section, we perform evaluation of our Hybrid-MVS pipeline, whose core algorithms are implemented in C++, on our experimental benchmark including eight cases, each of which is composed of multi-view digital images captured by Canon 850D camera with 6000×4000 image resolution, and RGB-D video stream captured by Azure Kinect with 2048×1536 image resolution and 640×576 depth resolution in 30FPS decompressed in 6FPS. Each case in

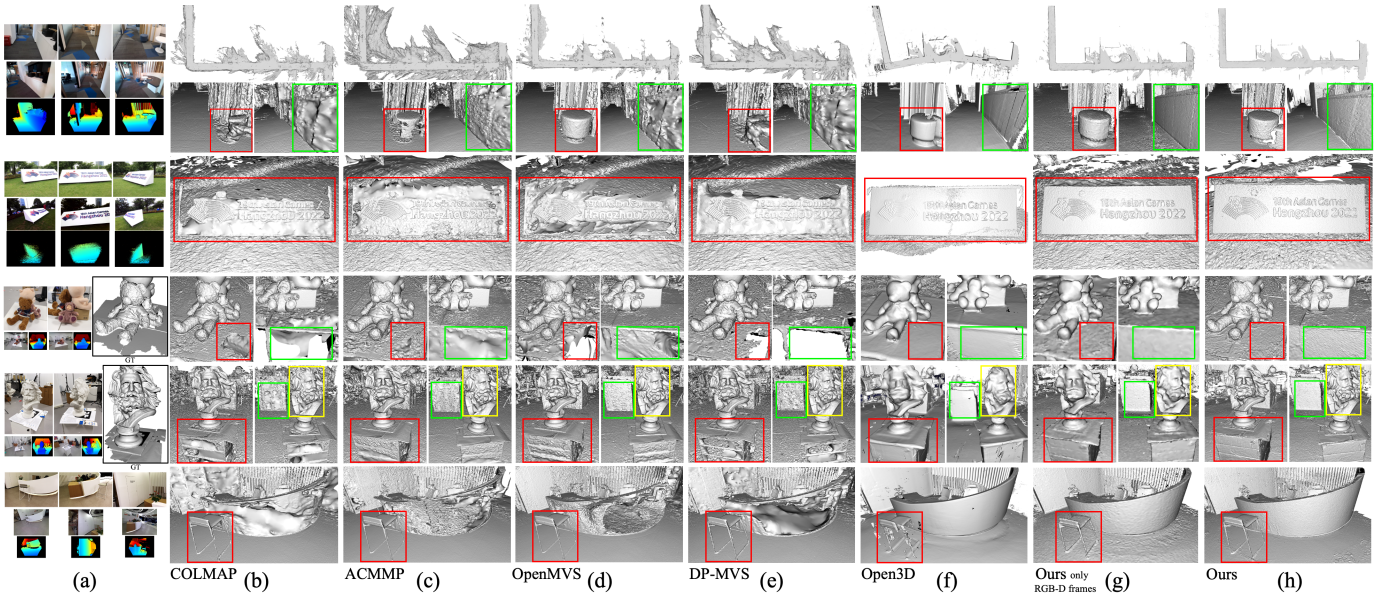


Fig. 9. Qualitative evaluation of our Hybrid-MVS pipeline on the cases "Corridor", "Qianjiang Century Park", "Dolls", "La Marseillaise" and "Front Desk". (a) Some representative source digital images and RGB-D frames for each case, with GT models of "Dolls" and "La Marseillaise" given. (b) The reconstructed surface models by COLMAP [2] with digital images and color channel of RGB-D frames. (c) The reconstructed surface models by ACMMP [3] with the same input as (b). (d) The reconstruction results by OpenMVS [4] with the same input as (b). (e) The reconstructed results by DP-MVS [5] with the same input as (b). (f) The reconstruction results by Open3D [1] with only the RGB-D frames. (g) The 3D surface models reconstructed by our Hybrid-MVS pipeline with only the RGB-D frames, which preserves the detailed structures and completeness better than (f). (h) The reconstructed surface models by Hybrid-MVS with both the digital images and the RGB-D frames, which performs the best in both reconstruction completeness and geometric details. Some detailed structures are highlighted in the rectangles to show the effectiveness of our proposed method.

TABLE I
EVALUATION OF RECONSTRUCTION ACCURACY (MAE AND RMSE IN MILLIMETERS) AND COMPLETENESS FOR COLMAP [2], ACMMP [3], OPENMVS [4], DP-MVS [5] WITH BOTH DIGITAL IMAGES AND COLOR CHANNEL OF RGB-D FRAMES, OPEN3D [1] WITH ONLY RGB-D FRAMES, OUR HYBRID-MVS WITH ONLY RGB-D FRAMES, AND HYBRID-MVS WITH BOTH DIGITAL IMAGES AND RGB-D FRAMES, ON THE CASES "LA MARSEILLAISE", "HORSE HEAD" AND "DOLLS".

Cases	Evaluation	COLMAP with digital images & RGB-D colors	ACMMP with digital images & RGB-D colors	OpenMVS with digital images & RGB-D colors	DP-MVS with digital images & RGB-D colors	Open3D with RGB-D frames	Hybrid-MVS with RGB-D frames	Hybrid-MVS with digital images & RGB-D frames
La Marseillaise	MAE/RMSE(mm)	1.738/3.220	1.340/1.963	1.317/2.410	1.677/3.079	7.269/8.992	2.966/3.603	**1.246/1.729**
	completeness	92.68%	99.48%	97.31%	93.52%	97.49%	98.94%	**99.84%**
Horse Head	MAE/RMSE(mm)	3.986/5.298	1.877/2.784	1.849/2.923	2.837/4.491	18.911/24.336	2.979/3.352	**1.758/2.248**
	completeness	82.91%	97.90%	96.18%	83.62%	94.67%	99.37%	**99.83%**
Dolls	MAE/RMSE(mm)	2.360/3.887	2.035/3.135	1.902/2.688	2.241/3.621	3.651/4.849	3.014/4.168	**1.814/2.568**
	completeness	96.14%	97.85%	99.39%	97.06%	98.34%	98.51%	**99.82%**

TABLE II
DETAILED COMPUTATION TIME OF HYBRID-MVS PIPELINE IN ALL THE STEPS OF CASES "SOFA", "HORSE HEAD" AND "HU QINGYUTANG" IN MINUTES.

Cases	#Digital Images/ #RGB-D Frames	Hybrid RGB-D Alignment	Hybrid PatchMatch	Fusion	Hybrid Delaunay Meshing	Total
Sofa	41/357	14.223	11.985	11.805	17.651	55.664
Horse Head	97/523	29.791	25.549	34.926	30.743	121.009
Hu Qingyutang	476/369	56.273	98.709	76.344	72.069	303.395

the benchmark typically contains some textured regions with geometric details, and some textureless or non-lambertian surfaces. We first exhibit ablation studies on hybrid depth and normal constraints of our PatchMatch stereo. Quantitative and qualitative comparisons of our work to the SOTA methods are then reported to show that our Hybrid-MVS achieves the best reconstruction accuracy and completeness on the detailed structures, textureless regions and lambertian surfaces of the scenes in the benchmark. We also report the time consumption on the stages of SfM, PatchMatch, fusion and meshing of different methods to show the time efficiency of our method.

A. Ablation studies of hybrid PatchMatch stereo

We perform ablation studies to validate the effects of depth and normal constraints on our hybrid PatchMatch stereo by using only one kind of constraint at a time. Fig. 8(e) shows the final depth maps and normal maps of our hybrid PatchMatch with both depth and normal weights as constraints, which performs the best in depth and normal accuracy, completeness and geometric details in the textureless desk and non-lambertian flowerpots, compared to the estimated depth maps and normal maps with only depth weight as constraint in Fig. 8(d) and the results with only normal constraint in Fig. 8(c). To

better visualize the detailed improvement, the multi-view depth maps by different ablation studies are fused to 3D surface models using our Delaunay meshing method of section VII. The geometric completeness and accuracy on MAE and RMSE are also evaluated for the fused models of "Horse Head". From the evaluation shown in Fig. 8, we can see that both depth constraint and normal constraint are helpful to the improvement of depth accuracy, and the combination of the two kinds of constraints can further improve the reconstruction completeness.

B. Qualitative and quantitative evaluations

Fig. 1 and 7 have already demonstrated the reconstructed 3D models of cases "Hu Qingyutang", "Sofa" and "Horse Head" in details. Other five cases "Corridor", "Qianjiang Century Park", "Dolls", "La Marseillaise" and "Front Desk" of our benchmark are shown in Fig. 9, each of which contains some textureless or non-lambertian areas. The case "Corridor" is a relatively large-scale indoor office with textureless and non-lambertian walls. We first give the qualitative comparisons with other SOTA MVS methods including COLMAP [2], ACMMP [3], OpenMVS [4] and DP-MVS [5] in Fig. 9(b), (c), (d) and (e). Since only RGB images are supported by these SOTA works, we use the digital images and color channel of the RGB-D frames for them to reconstruct 3D meshes. From the details of the reconstructed surface meshes, we can see that our Hybrid-MVS can extract more complete surface structures on the low-textured desks and walls of "La Marseillaise" and "Corridor", and the front desk and billboard of "Qianjiang Century Park" which are both textureless and non-lambertian surfaces, while COLMAP and OpenMVS cannot ensure the reconstruction completeness or geometric accuracy on the textureless and non-lambertian regions with their sequential PatchMatch propagation strategy. The 3D models reconstructed by DP-MVS contain more geometric details but cannot ensure reconstruction completeness on the textureless desks and walls and the non-lambertian front desk even with its diffusion-like propagation strategy. ACMMP performs better than COLMAP, OpenMVS and DP-MVS in reconstruction completeness due to its multi-scale framework and planar priors through triangulation, but still cannot handle textureless and non-lambertian surfaces so well as our approach. Since Open3D [1] only support one source RGB-D video stream, we also qualitatively compare the surface mesh reconstructed by Open3D [1] with only the RGB-D frames, the reconstruction result by our Hybrid-MVS with only RGB-D frames, and the result by Hybrid-MVS with both digital images and RGB-D frames as input in Fig. 9(f), (g) and (h). It can be seen from the reconstructed surface mesh details that the results from RGB-D frames by Open3D can ensure reconstruction completeness, but lose detailed structures on the sculpture and chair legs, or lack global structural consistency on the large-scale "Corridor" and "Qianjiang Century Park" due to its local ICP tracking, which are highlighted in the rectangles. Especially, the reconstruction completeness of "Qianjiang Century Park" is severely affected by the invalid depths caused by the outdoor sunlight. In comparison, our Hybrid-MVS with only RGB-D frames can preserve more accurate geometric details without losing completeness, and even complete the missing structures caused by invalid prior depths, while Hybrid-MVS with both digital images and RGB-D frames further ensures both the complete surface structures in textureless or non-lambertian regions and the geometric details of the textured regions.

We further provide quantitative evaluation on the three cases "Horse Head", "Dolls" and "La Marseillaise" which contain GT models scanned by EinScan Pro 2X PLUS digital 3D scanner for evaluation of reconstruction accuracy on both MAE and RMSE, and reconstruction completeness. We use CloudCompare [2] to compare the reconstructed meshes with GT. To evaluate the accuracy of a reconstructed mesh, we first align the mesh with the GT model manually and refine its transformation by ICP fine registration. Then we compute point-to-plane distance between the reconstructed mesh and GT. These routines are achieved with CloudCompare's built-in functions to produce MAE, RMSE and completeness. We evaluate the accuracy and completeness of the reconstructed surface models by COLMAP [2], ACMMP [3], OpenMVS [4], DP-MVS [5] with digital images and RGB-D frames, Open3D [1] with only the RGB-D frames, our Hybrid-MVS with only RGB-D frames, and Hybrid-MVS with both the digital images and the RGB-D frames. From the model accuracy and completeness evaluation in Table I, we can see that our Hybrid-MVS reconstructs the scene models with a millimeter-level accuracy, which turns out to be the best in MAE, RMSE and completeness, compared to all the other approaches. Note that even with only RGB-D frames, Hybrid-MVS can still produce more accurate and complete surface models than Open3D, which verifies the usefulness of our joint optimization of visual and depth cues.

C. Time statistics

Table II gives the time statistics of our pipeline on three typical cases: two indoor examples "Sofa" and "Horse Head" and one outdoor case "Hu Qingyutang". The experiments are conducted on a server platform with two 10-Core Intel Xeon Silver 4114 CPUs @ 2.2GHz, eight GeForce 1080Ti GPUs, and 250GB memory. The hybrid RGB-D alignment, fusion and hybrid Delaunay meshing modules run on CPUs, while the hybrid PatchMatch propagation is speeded up by GPU parrallism. Note that the time consumptions of hybrid RGB-D alignment, fusion and hybrid Delaunay meshing is proportional to the numbers of digital images and RGB-D frames, and the hybrid PatchMatch stereo step is proportional to the number of digital images. Even for the large-scale outdoor scene 'Hu Qingyutang", our Hybrid-MVS can also achieve high-quality 3D reconstruction in a time-efficient way.

IX. Conclusion and Future Work

In this work, we propose a hybrid MVS pipeline, which combines a digital camera and a consumer RGB-D camera to achieve a complete and accurate 3D reconstruction. Both visual cues from the digital camera and depth cues from

[2]http://cloudcompare.org

the RGB-D camera are tightly coupled to maximize their complementary advantages for more robust PatchMatch stereo and Delaunay meshing to ensure complete reconstruction in textureless and non-lambertian regions, while preserving accurate geometric details in textured regions. The effectiveness of the proposed Hybrid-MVS framework for indoor and outdoor scenes is validated in our experiments.

Although the combination of digital camera and RGB-D camera can help improving the reconstruction quality and robustness, it is still not so convenient for a non-professional user to take multi-view digital pictures of a large-scale scene. How to combine more convenient capturing devices like 360 panoramic camera with RGB-D cameras to reduce the data capturing requirements is a problem worth studying in future. Besides, some more complicated scenes with severe repetitive textures or varying illuminations might influence the hybrid RGB-D alignment, thereby affecting the final 3D reconstruction results. Moreover, our method relies on the depth quality of the acquisition device, and therefore benefits little when the acquired depth correctness or completeness cannot be guaranteed. It remains to be a future work for us to jointly take these problems into consideration to develop a more robust and practical multi-view reconstruction system.

ACKNOWLEDGMENTS

The authors wish to thank Jinle Ke, Yi Su, Qi Chen, Fei Jiao, and Chiming Li for their kind help in the development of the Hybrid-MVS system. This work was partially supported by the National Key Research and Development Program of China under Grant 2020YFF0304300, and the National Natural Science Foundation of China (No. 61932003).

REFERENCES

[1] Q. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," *arXiv preprint arXiv:1801.09847*, 2018.

[2] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision*. Springer, 2016, pp. 501–518.

[3] Q. Xu, W. Kong, W. Tao, and M. Pollefeys, "Multi-scale geometric consistency guided and planar prior assisted multi-view stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[4] D. Cernea, "OpenMVS: Multi-view stereo reconstruction library," 2020. [Online]. Available: https://cdcseacave.github.io/openMVS

[5] L. Zhou, Z. Zhang, H. Jiang, H. Sun, H. Bao, and G. Zhang, "DP-MVS: Detail preserving multi-view surface reconstruction of large-scale scenes," *Remote Sensing*, vol. 13, no. 22, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/22/4569

[6] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1, pp. 7–42, 2001.

[7] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor et al., "Fusion4D: Real-time performance capture of challenging scenes," *ACM Transactions on Graphics*, vol. 35, no. 4, pp. 1–13, 2016.

[8] K. Guo, P. Lincoln, P. Davidson, J. Busch, X. Yu, M. Whalen, G. Harvey, S. Orts-Escolano, R. Pandey, J. Dourgarian et al., "The relightables: Volumetric performance capture of humans with realistic relighting," *ACM Transactions on Graphics*, vol. 38, no. 6, pp. 1–19, 2019.

[9] S. Yang, B. Li, Y.-P. Cao, H. Fu, Y.-K. Lai, L. Kobbelt, and S.-M. Hu, "Noise-resilient reconstruction of panoramas and 3d scenes using robotmounted unsynchronized commodity RGB-D cameras," *ACM Transactions on Graphics*, vol. 39, no. 5, pp. 1–15, 2020.

[10] Y. Fu, Q. Yan, J. Liao, and C. Xiao, "Joint texture and geometry optimization for RGB-D reconstruction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5950–5959.

[11] R. Or-El, G. Rosman, A. Wetzler, R. Kimmel, and A. M. Bruckstein, "RGBD-Fusion: Real-time high precision depth recovery," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5407–5416.

[12] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2006, pp. 519–528.

[13] S. M. Seitz and C. R. Dyer, "Photorealistic scene reconstruction by voxel coloring," *International Journal of Computer Vision*, vol. 35, no. 2, pp. 151–173, 1999.

[14] G. Vogiatzis, C. H. Esteban, P. H. Torr, and R. Cipolla, "Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2241–2246, 2007.

[15] S. N. Sinha, P. Mordohai, and M. Pollefeys, "Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh," in *IEEE International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[16] O. Faugeras and R. Keriven, "Variational principles, surface evolution, PDE's, level set methods and the stereo problem," *IEEE Transactions on Image Processing*, vol. 7, no. 3, pp. 336–344, 1998.

[17] V. H. Hiep, R. Keriven, P. Labatut, and J.-P. Pons, "Towards high-resolution large-scale multi-view stereo," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1430–1437.

[18] A. Romanoni and M. Matteucci, "Mesh-based camera pairs selection and occlusion-aware masking for mesh refinement," *Pattern Recognition Letters*, vol. 125, pp. 364–372, 2019.

[19] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz, "Multi-view stereo for community photo collections," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[20] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1362–1376, 2009.

[21] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Towards internet-scale multi-view stereo," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1434–1441.

[22] C. Strecha, R. Fransens, and L. Van Gool, "Combined depth and outlier estimation in multi-view stereo," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2006, pp. 2394–2401.

[23] M. Goesele, B. Curless, and S. M. Seitz, "Multi-view stereo revisited," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2006, pp. 2402–2409.

[24] A. Kuhn, H. Mayer, H. Hirschmüller, and D. Scharstein, "A TV prior for high-quality local multi-view stereo reconstruction," in *International Conference on 3D Vision*, vol. 1. IEEE, 2014, pp. 65–72.

[25] D. Bradley, T. Boubekeur, and W. Heidrich, "Accurate multi-view reconstruction using robust binocular stereo and surface meshing," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[26] Z. Li, W. Zuo, Z. Wang, and L. Zhang, "Confidence-based large-scale dense multi-view stereo," *IEEE Transactions on Image Processing*, vol. 29, pp. 7176–7191, 2020.

[27] S. Shen, "Accurate multiple view 3D reconstruction using patch-based stereo for large-scale scenes," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1901–1914, 2013.

[28] E. Tola, C. Strecha, and P. Fua, "Efficient large-scale multi-view stereo for ultra high-resolution image sets," *Machine Vision and Applications*, vol. 23, no. 5, pp. 903–920, 2012.

[29] K. Zhang, J. Lu, Q. Yang, G. Lafruit, R. Lauwereins, and L. Van Gool, "Real-time and accurate stereo: A scalable approach with bitwise fast voting on CUDA," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 867–878, 2011.

[30] Z. Lu, J. Wang, Z. Li, S. Chen, and F. Wu, "A resource-efficient pipelined architecture for real-time semi-global stereo matching," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[31] Y. Lee and H. Kim, "A high-throughput depth estimation processor for accurate semiglobal stereo matching using pipelined inter-pixel aggregation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[32] X. Yang, L. Zhou, H. Jiang, Z. Tang, Y. Wang, H. Bao, and G. Zhang, "Mobile3DRecon: Real-time monocular 3D reconstruction on a mobile phone," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 12, pp. 3446–3456, 2020.

[33] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *European Conference on Computer Vision*, 2018, pp. 767–783.

[34] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent MVSNet for high-resolution multi-view stereo depth inference," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5525–5534.

[35] R. Chen, S. Han, J. Xu, and H. Su, "Point-based multi-view stereo network," in *IEEE International Conference on Computer Vision*, 2019, pp. 1538–1547.

[36] K. Luo, T. Guan, L. Ju, H. Huang, and Y. Luo, "P-MVSNet: Learning patch-wise matching confidence aggregation for multi-view stereo," in *IEEE International Conference on Computer Vision*, 2019, pp. 10452–10461.

[37] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2495–2504.

[38] A. Kuhn, C. Sormann, M. Rossi, O. Erdler, and F. Fraundorfer, "DeepC-MVS: Deep confidence prediction for multi-view stereo reconstruction," in *International Conference on 3D Vision*, 2020, pp. 404–413.

[39] R. Peng, R. Wang, Z. Wang, Y. Lai, and R. Wang, "Rethinking depth estimation for multi-view stereo: A unified representation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 8645–8654.

[40] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, and A. W. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *IEEE International Symposium on Mixed and Augmented Reality*, 2011.

[41] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface re-integration," *ACM Transactions on Graphics*, vol. 36, no. 4, p. 1, 2017.

[42] O. Kahler, V. Prisacariu, C. Ren, X. Sun, P. Torr, and D. Murray, "Very high frame rate volumetric integration of depth images on mobile devices," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 11, pp. 1–1, 2015.

[43] P. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[44] X. Xiang, H. Jiang, G. Zhang, Y. Yu, C. Li, X. Yang, D. Chen, and H. Bao, "Mobile3DScanner: An online 3D scanner for high-quality object reconstruction with a mobile device," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 11, pp. 4245–4255, 2021.

[45] P. Merrell, A. Akbarzadeh, W. Liang, P. Mordohai, J. M. Frahm, R. Yang, D. Nistér, and M. Pollefeys, "Real-time visibility-based fusion of depth maps," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[46] Y. Cui, S. Schuon, D. Chan, S. Thrun, and C. Theobalt, "3D shape scanning with a time-of-flight camera," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1173–1180.

[47] J. Kang, S. Lee, M. Jang, and S. Lee, "Gradient flow evolution for 3D fusion from a single depth sensor," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[48] O. Mac Aodha, N. D. Campbell, A. Nair, and G. J. Brostow, "Patch based synthesis for single depth image super-resolution," in *European Conference on Computer Vision*. Springer, 2012, pp. 71–84.

[49] M. Hornácek, C. Rhemann, M. Gelautz, and C. Rother, "Depth super resolution by rigid body self-similarity in 3D," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1123–1130.

[50] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[51] M.-Y. Liu, O. Tuzel, and Y. Taguchi, "Joint geodesic upsampling of depth images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 169–176.

[52] X. Gao, S. Shen, L. Zhu, T. Shi, Z. Wang, and Z. Hu, "Complete scene reconstruction by merging images and laser scans," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3688–3701, 2019.

[53] X. Gao, H. Cui, L. Zhu, T. Shi, and S. Shen, "Multi-source data-based 3D digital preservation of largescale ancient chinese architecture: A case report," *Virtual Reality and Intelligent Hardware*, vol. 1, no. 5, pp. 525–541, 2019.

[54] S. Lu, X. Ren, and F. Liu, "Depth enhancement via low-rank matrix completion," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3390–3397.

[55] W. Xie, M. Wang, X. Qi, and L. Zhang, "3D surface detail enhancement from a single normal map," in *IEEE International Conference on Computer Vision*, 2017, pp. 2325–2333.

[56] R. Maier, K. Kim, D. Cremers, J. Kautz, and M. Nießner, "Intrinsic3D: High-quality 3D reconstruction by joint appearance and geometry optimization with spatially-varying lighting," in *IEEE International Conference on Computer Vision*, 2017.

[57] G. Choe, J. Park, Y.-W. Tai, and I. S. Kweon, "Refining geometry from depth sensors using IR shading images," *International Journal of Computer Vision*, vol. 122, no. 1, pp. 1–16, 2017.

[58] Z. Li, P. C. Gogia, and M. Kaess, "Dense surface reconstruction from monocular vision and LiDAR," in *IEEE International Conference on Robotics and Automation*. IEEE, 2019, pp. 6905–6911.

[59] X. Yang and G. Jiang, "A practical 3D reconstruction method for weak texture scenes," *Remote Sensing*, vol. 13, no. 16, p. 3103, 2021.

[60] E. K. Stathopoulou, R. Battisti, D. Cernea, F. Remondino, and A. Georgopoulos, "Semantically derived geometric constraints for MVS reconstruction of textureless areas," *Remote Sensing*, vol. 13, no. 6, p. 1053, 2021.

[61] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics*, vol. 28, no. 3, p. 24, 2009.

[62] E. Zheng, E. Dunn, V. Jojic, and J.-M. Frahm, "PatchMatch based joint view selection and depthmap estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1510–1517.

[63] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in *IEEE International Conference on Computer Vision*, 2015, pp. 873–881.

[64] Q. Xu and W. Tao, "Multi-scale geometric consistency guided multi-view stereo," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5483–5492.

[65] A. Romanoni and M. Matteucci, "Tapa-MVS: Textureless-aware PatchMatch multi-view stereo," in *IEEE International Conference on Computer Vision*, 2019, pp. 10 413–10 422.

[66] J. Liao, Y. Fu, Q. Yan, and C. Xiao, "Pyramid multi-view stereo with local consistency," in *Computer Graphics Forum*, vol. 38, no. 7. Wiley Online Library, 2019, pp. 335–346.

[67] Q. Xu and W. Tao, "Planar prior assisted PatchMatch multi-view stereo," in *AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12 516–12 523, 2020.

[68] J. Li, E. Li, Y. Chen, L. Xu, and Y. Zhang, "Bundled depth-map merging for multi-view stereo," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2769–2776.

[69] M. Rossi, M. E. Gheche, A. Kuhn, and P. Frossard, "Joint graph-based depth refinement and normal estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 154–12 163.

[70] Q. Xu and W. Tao, "Multi-view stereo with asymmetric checkerboard propagation and multi-hypothesis joint view selection," *arXiv preprint arXiv:1805.07920*, 2018.

[71] Z. Xu, Y. Liu, X. Shi, Y. Wang, and Y. Zheng, "MARMVS: Matching ambiguity reduced multiple view stereo for efficient large scale scene reconstruction," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5981–5990.

[72] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4104–4113.

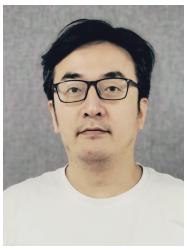
[73] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle Adjustment—A modern synthesis," in *International Workshop on Vision Algorithms: Theory and Practice*. Springer, 2000, pp. 298–372.

[74] J. L. Schönberger, T. Price, T. Sattler, J.-M. Frahm, and M. Pollefeys, "A vote-and-verify strategy for fast spatial verification in image retrieval," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 321–337.

[75] P. Labatut, J.-P. Pons, and R. Keriven, "Robust and efficient surface reconstruction from range data," *Computer Graphics Forum*, vol. 28, no. 8, pp. 2275–2290, 2009.

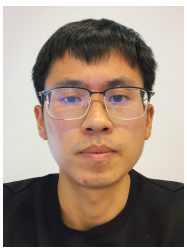
Chenchen Li received his master's degree in Computer Science from Wuhan University, where he majored in Computational Differential Geometry. He is currently a Senior Researcher in SenseTime Group Ltd, China. He received the best journal paper nominee of ISMAR 2021. His research interests focus on computer vision, multi-view stereo, 3D reconstruction, neural radiance fields, and augmented reality.

Liyang Zhou is currently a Senior Algorithm Engineer in SenseTime Group Ltd, China. He received his master degree from Zhejiang University in 2016, and the best paper award of ISMAR 2020. His research interests include structure from motion, multi-view stereo, and 3D reconstruction.



Hanqing Jiang received his Ph.D. degree from Zhejiang University, after which he was a postdoctoral researcher in the State Key Lab of CAD&CG, Zhejiang University. He is currently a Research Director in SenseTime Group Ltd, China. He received the best paper award of ISMAR 2020 and the best journal paper nominee of ISMAR 2021. His research interests focus on computer vision, including video enhancement, multi-view stereo, 3D reconstruction, neural radiance fields, and augmented reality.

Zhuang Zhang received his Ph.D. degree in Measurement Technology and Instruments from the University of Chinese Academy of Sciences. He is currently a Senior Algorithm Engineer in SenseTime Group Ltd, China. His research interests focus on computer vision, include 3D reconstruction, multi-sensor fusion, and augmented reality.



Xiaojun Xiang received his master degree from the State Key Lab of CAD&CG, Zhejiang University. He is currently a Senior Researcher in SenseTime Group Ltd, China. He received the best journal paper nominee of ISMAR 2021. His research interests lie in the field of 3D computer vision, including multi-view stereo, vectorized modeling, and implicit neural reconstruction.

Han Sun is currently a researcher in SenseTime Group Ltd, China. He received his master degree from University of Science and Technology of China in 2018. His research interests include structure from motion, multi-view stereo, and 3D reconstruction.

Qing Luan obtained her Ph.D. degree in Computer Vision from the University of Science and Technology of China, after which she worked in Microsoft's product team in Redmond, WA for 6 years. She is currently the General Manager of Digital Entertainment and Culture Business, Digital World Group at SenseTime Group Ltd. Her main research and product focus is on augment reality and AI content generation, especially on entertainment video AR effect and video content generation.

Hujun Bao is currently a professor in the Computer Science Department of Zhejiang University, and the former director of the State Key Lab of CAD&CG. His research interests include computer graphics, computer vision and mixed reality. He leads the mixed reality group in the lab to make a wide range of research on 3D reconstruction and modeling, real-time rendering and virtual reality, realtime 3D fusion and augmented reality. Many of these algorithms have been successfully integrated into the mixed reality system SenseMARS.

Guofeng Zhang is now a professor at State Key Lab of CAD&CG, Zhejiang University. He received his BS and Ph.D. degrees in Computer Science from Zhejiang University, in 2003 and 2009, respectively. He received the National Excellent Doctoral Dissertation Award, the Excellent Doctoral Dissertation Award of China Computer Federation and the best paper award of ISMAR 2020. His research interests include structure-from-motion, SLAM, 3D reconstruction, augmented reality, video segmentation, and editing.