IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 20, 2019, accepted October 6, 2019, date of publication October 9, 2019, date of current version April 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2946456

Weakly Supported Plane Surface Reconstruction via Plane Segmentation Guided Point Cloud Enhancement

SHEN YAN, YANG PENG, GUANGYUE WANG, SHIMING LAI, AND MAOJUN ZHANG
College of Systems Engineering, National University of Defense Technology, Changsha 410073, China

Corresponding author: Shiming Lai (shiming413@nudt.edu.cn)

This work was supported by the National Natural Science Foundation of China, Grant No. 61703415, "Research on Video Stitching for Dynamic Scenes", Principal Participator, January 2018–November 2020.

ABSTRACT Most of the widely used multi-view 3D reconstruction algorithms assume that object appearance is predominantly diffuse and full of good texture. For the objects that violate this restriction, the surface can hardly be reconstructed because such area lacks sufficient support from dense point clouds. To tackle this problem, we introduce a novel two-stage prior-guided method based on point clouds enhancement to enable the application of multi-view reconstruction approaches in such scenes. In the first stage, we optimize the original PlaneNet plane segmentation priors by taking advantage of the estimated depth map and confidence map from multi-view stereo. In the second stage, we correct and supply 3D point clouds for the weakly supported plane surface on the basis of the upgraded priors. Furthermore, we utilize a slight disturbance of the enhanced point clouds to facilitate the subsequent mesh reconstruction. The proposed point cloud enhancement approach is evaluated on the large-scale *DTU* dataset. Our method significantly outperforms previous multi-view stereo state-of-the-arts. We also demonstrate weakly supported plane surface reconstruction results from real-world photos that are unachievable with either the methods aiming at preserving weakly supported surfaces or the traditional state-of-the-art 3D reconstruction systems.

INDEX TERMS 3D reconstruction, weakly supported surface, real image reconstruction.

I. INTRODUCTION

3D reconstruction refers to the rebuilding of certain real 3D objects or 3D scenes, making them easily accessible for human perception as well as computer representation and processing. Currently, 3D reconstruction from multi-view images [2]–[6] is quite popular. Compared with the traditional modeling methods, such as modeling software (3D Max, AutoCAD, etc.), and laser scan methods [7], multi-view approaches are inexpensive and highly automated. The pipeline for conventional 3D reconstruction from unordered images consists of three stages: first, obtaining camera internal and external parameters and sparse point clouds using structure-from-motion (SfM) methods [8]–[12]; second, recovering dense point clouds using multi-view stereo (MVS) methods [13]–[18]; and third, generating surface meshes based on dense point clouds and posting textures through surface reconstruction [19]–[22]. Due to the breakthrough progress of the SfM and MVS methods, the point clouds are often recovered with very high accuracy, and these techniques have achieved impressive results.

However, one of the drawbacks of the classic multi-view geometric method is its over-reliance on the photometric consistency assumption [23] and Lambertian reflection assumption [24], which can be frequently violated in the real world. These situations are always present when building poorly textured regions and transparent or highly reflective surfaces. The iPad screen shown in Figure 1 is an example of this. The reconstructed point clouds of the laptop screen either disappear or appear in the wrong location, making it impossible to further create the corresponding meshes. This kind of region is called a weakly supported plane surface, the reconstruction of which is a challenging problem. In fact, currently, even the state-of-the-art multi-view algorithms cannot obtain a satisfactory result for this surface.

Most of the methods reported in the literature concentrate on flossy surface reconstruction and address the problem by adding extra hardware (e.g., coded pattern projection [25] and

The associate editor coordinating the review of this manuscript and approving it for publication was Dong Wang.

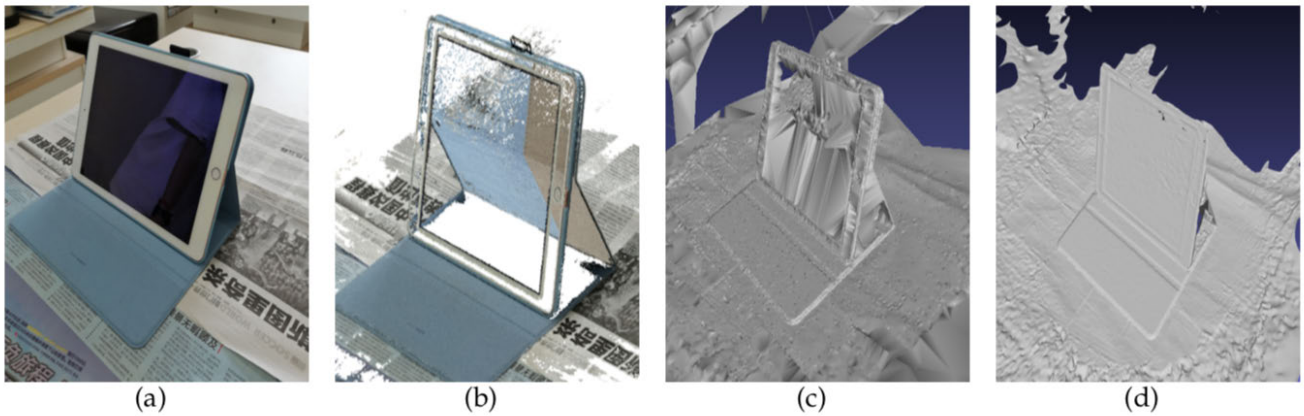

FIGURE 1. Results for the 'iPad' data set. (a) Input image, (b) dense point clouds recovered with COLMAP, and (c) mesh reconstructed with COLMAP; (d) the technique presented in this work better reconstructs weakly supported surfaces (the iPad screen).

a two-layer LCD [26]), filtering the non-Lambert region [27], [28] or translating multi-view images of the objects with specular reflection to diffuse images [29]. Although these methods have made great progress towards non-diffuse surface reconstruction, they cannot handle textureless images, making them unable to realize general weakly supported surface reconstruction. Unlike the above approaches, the methods proposed in [30]–[33] working with Delaunay tetrahedralization of the dense point clouds provide solutions that can reconstruct a weakly supported surface of any kind. These methods separate the tetrahedra labeled as ''free'' (empty space) and the tetrahedra labeled as ''full'' (full space) by minimizing the s-t cut, and the target surface can be viewed as an interface between the free and full spaces. However, these techniques often break down when the percentage of missing and irrelevant point clouds is high because, in this case, Delaunay tetrahedra cannot be built at all.

Recently, data-driven approaches that learn priors to tackle 3D reconstruction problems have been widely applied to generate 3D scenes from single or multiple input images [34], [35]. These methods can effectively reconstruct poorly textured surfaces to some extent. Additionally, some learning-based methods can accomplish image plane segmentation in advance to obtain a better 3D prediction [1], [36], [37]. Even though these approaches expand the solution to the 3D reconstruction problem, they can only be reliable for examples similar to the training dataset that they learn from, which means that the practical application scenario is quite limited. Realizing this restriction, a number of researchers have focused on applying deep learning in substages of the stereo reconstruction pipeline, such as pose estimation [38], feature detection and description [39], [40], image retrieval and matching [41], and bundle adjustment [42]. However, these methods still impose the photometric consistency and Lambertian assumptions, which are not suitable for our goal.

Inspired by recent studies, we present a prior-guided approach that adds a compatible but special step to the 3D reconstruction pipeline to augment the point clouds in the weakly supported plane region. Based on this, we can provide sufficient point support in such areas, leading to surface reconstruction with high accuracy and integrity without much effort. We carry out our method in two stages. In the first stage, we combine a data-driven plane segmentation mask with the attributes of depth-map merging based MVS method to refine the initial inaccurate segmentation prior. Then, we propagate the refined mask to all visible images, serving as the projections of the truly 3D plane surface. In the second stage, by carefully analyzing the depth map, the confidence map and the plane segmentation mask, we effectively obtain point clouds that belong to the target plane surface with high confidence. The corresponding 3D plane parameters are obtained by fitting these point clouds with the RANSAC [43] framework. Last, our approach corrects and yields point clouds that not only satisfy the plane parameters mentioned above but also can be projected back into the mask area. A main advantage of our approach is its non-reliance on particularly accurate and complete a priori information - a necessary condition for reconstructing objects from real-world images. With this method, we take a step toward the practical usage of 3D weakly supported plane surface reconstruction with real images.

Our contributions are summarized as follows:

1) We incorporate the multi-view stereo method with data-driven plane segmentation cues to further correct and create the missing point clouds of the weakly supported plane surface.
2) We introduce a novel segmentation process that, provided with an incorrect and incomplete plane segmentation of a certain image, enables all of the masks of the same plane among the visible images to be extracted accurately and efficiently.
3) We evaluate our method on *DTU* and real-life datasets and compare it with other existing methods. The experimental results prove that our method has a better performance than either traditional MVS pipelines or other advanced alternatives.

In the rest of the paper, we first present the related work in Section II and then introduce the mechanism of our method in detail in Section III. Then, we conduct extensive experiments on *DTU* and real-life datasets in Section IV. At the end, we present the conclusion of our findings in Section V.

II. RELATED WORK

A. CONVENTIONAL MULTI-VIEW 3D RECONSTRUCTION

As described in Section I, the conventional multi-view 3D reconstruction pipeline can be divided into three parts, namely, structure-from-motion (SfM), multi-view stereo (MVS), and surface reconstruction. In this section, we focus on the first two processes because our method aims at operating point clouds that come from the end of MVS combined with SfM.

We only focus on the incremental SfM approaches [12] that are adopted by large-scale software packages [10], [16], including commercial packages, to achieve a stable reconstruction effect. Typically, the core process of the incremental SfM is multiple realizations of nonlinear optimization, which is called bundle adjustment [44], [45]. The parameters optimized by bundle adjustment are normally the generated sparse feature point cloud coordinates, registered camera focal length and camera external parameters, where the initial values of these parameters are obtained with multiple view geometry in computer vision. During the initialization process, feature extraction [46]–[49], image matching [50], geometric verification [43], and the PnP algorithm [51] are used to register the new image, and triangulation [43] is applied. The reader is referred to [10] for more details. Since this problem has been well addressed, we can directly regard the output of SfM as reliable data when reconstructing real objects.

Almost all MVS algorithms reconstruct the 3D geometry using photo consistency [52] functions. By measuring the agreement between a set of input photographs' information, such as illumination, material and texture, MVS algorithms can invert the image formation process and produce highly a detailed 3D geometry. The MVS process can be implemented in a variety of ways [23], and in this paper, we only introduce depth-map merging-based methods not only because they currently show the best performance but also to obtain more intermediate verbosities to help us generate point clouds at weakly supported planes. These methods consist of four steps: stereo pair selection [16], depth-map computation [53], depth-map refinement [54], and depth-map fusion [55], with each of these processes intensely investigated in previous work. In the refined depth map, the depth is marked as unknown if it is not determined; while this leads to very good performance in terms of accuracy, there is much room for improvement for the reconstruction completeness. Recently, Schönberger *et al.* presented the COLMAP [10] MVS system. Through a tight integration of multiple advanced techniques, COLMAP is one of the best performing algorithms on several public multi-view stereo benchmarks and is useful for real-world reconstruction. However, because it does not eliminate the photo consistency assumption, COLMAP still has difficulty in manipulating poorly textured and reflective surfaces. Additionally, OpenMVG [56] combined with OpenMVS is another mature open-source 3D reconstruction system; even though the 3D reconstruction production is promising, it faces the same problem as COLMAP.

B. WEAKLY SUPPORTED SURFACE RECONSTRUCTION

As mentioned above, most photometric stereo methods assume that the appearance of the object is uniquely identified. However, such assumptions are not valid for specular and textureless objects, and researchers have put forward improved methods. These methods can be roughly classified into two categories with respect to their operating objects. The first are focused on non-diffuse surfaces, while the others aim at reconstructing weakly supported surfaces of any kind. Tin *et al.* [26] adopt a two-layer liquid crystal display (LCD) setup to encode the illumination directions for reconstruction of the mirror-type specular objects. Or-El *et al.* [57] address the same issue by exploiting the built-in monochromatic IR projector and IR images of RGB-D scanners. While such techniques can deal with challenging non-Lambertian effects, they require the use of additional hardware and user expertise. Given only images, Büyükatalay *et al.* [57] directly filter the highlight surface, and Mallick *et al.* [28] separate the specular reflection effects for surfaces that can be modeled with dichromatic reflectance. Wu *et al.* [29] extend a ''specular to diffuse'' generative adversarial network translation for transforming objects with specular reflection into diffuse objects. Since these methods target only one type of weakly supported surfaces, they cannot fully solve the problem raised in this paper.

On the other hand, inspired by Sinha *et al.* [32], who formulated multi-view 3D shape reconstruction as the computation of a minimum s-t cut on the dual graph of a tetrahedral mesh, and Labatut *et al.* [31], who provided an energy function that perfectly fits into the above minimum optimization framework, Jancosek and Pajdla [33] augmented these methods with the ability to cope with any kind of weakly supported surface by merely changing the t-edge weights. Then, they proposed an interface classifier to modify the previous method, obtaining better performance [30]. However, for this type of approach, there is no way to deal with the extreme situation in which the surface is completely free of point cloud support.

C. LEARNING-BASED 3D RECONSTRUCTION

Learning-based 3D reconstruction from images has recently been a quite active research direction. The studied methods mainly follow two technical routes, namely, an end-to-end implementation directly from the images to the final 3D model, and the replacement of some intermediate processes of the traditional multi-view pipeline. In the former framework, Tatarchenko *et al.* [58] present a convolutional network to infer a 3D representation of a previously unseen object with a single image. Choy *et al.* [59] propose a recurrent

neural network to learn a mapping from the pictures of the objects to their underlying 3D shapes. Lin *et al.* [34] propose a 3D generative modeling framework to efficiently generate object shapes in the form of dense point clouds. Along the latter line, we are mainly concerned about the learning-based MVS methods because these methods share similar characteristics with the depth-map merging-based approaches that we adopt. Zbontar *et al.* [60] and Hartmann *et al.* [61] sought to learn a similarity measure for patch matching. MvsNet [62] represents an end-to-end deep learning architecture for depth map inference from multi-view images. DeepMVS [63] predicts high-quality disparity maps taking an arbitrary number of posed pictures as the input. Nevertheless, generally, the reconstruction quality of these methods cannot surpass that of the traditional approaches, particularly for real photos, because of the diversity limitations of the training data.

Moreover, for a scene with many planes, such as an indoor environment, to increase the plane constraint during the reconstruction, PlaneNet [1] propose end-to-end CNNs to directly infer a set of plane parameters and corresponding plane segmentation masks from a single RGB image. PlaneRCNN [64] has been reported to improve the quality of PlaneNet by employing a variant of Mask R-CNN and adopting a novel loss during training, but this method is not open source yet. Yu *et al.* [36] leverage a two-stage method based on associative embedding to detect an arbitrary number of planes. These studies are helpful for our work because they provide the priory segmentation mask of the weakly supported plane in images, even though these segmentation masks are inaccurate for real data.

III. PROPOSED METHOD

Our goal is to generate a complete 3D scene containing a weakly supported surface. The input for our approach is a multi-view image sequence with only manual assignation of the weakly supported plane in a certain image. This enables our approach to process real-world photos, where such additional information is easy to acquire. The output of our method is scene point clouds rectified at the weakly supported plane position, which directly serve as the input to the surface reconstruction pipeline, resulting in integrity-improved 3D reconstruction without additional effort. To accomplish this goal, we add an auxiliary step after the standard dense point cloud reconstruction that modifies the mistaken point clouds on the weakly supported plane and supplements some fresh point clouds in this region. Moreover, we also provide precise image segmentation masks as the projection of the weakly supported planar surface on each visible image. The overall pipeline of our proposed approach is visualized in Figure 2. We discuss the depth-map information and the confidence map information obtained from the multi-view stereo process in Section III-A. In Section III-B we introduce the method for obtaining a complete and precise plane mask on the basis of a pretrained plane segmentation model. The implementation details of the correction and interpolation of the point clouds of a weakly supported plane surface are discussed in Section III-C. Finally, in Section III-D, we apply a minor disturbance to the planar point cloud position, which is necessary for the establishment of the Delaunay tetrahedron during surface reconstruction. All of the parameters in this section are evaluated in the normalized coordinate system.

A. DEPTH MAP AND CONFIDENCE MAP

Suppose that we have N unordered images $\{I_i, P_i\}_{i=1}^N$, where each image I_i is associated with its self-camera matrix P_i. First, the camera matrix $\{P_i\}_{i=1}^N$ is calibrated utilizing SfM algorithms. Specifically, the rotation matrix $\{R_i\}_{i=1}^N$, the camera center position $\{C_i\}_{i=1}^N$ and the intrinsic parameters $\{K_i\}_{i=1}^N$ are obtained, where their relation with camera matrix $\{P_i\}_{i=1}^N$ is described as follows:

$$P_i = K_i R_i [I| - C_i] \tag{1}$$

where I is a 3×3 identity matrix.

Then, based on the relative position of the calibrated cameras $\{P_i\}_{i=1}^N$, the best performing MVS method (depth-map merging based) is applied to recover dense point clouds. This method computes the depth map at each view and then fuses the obtained depth maps together into a single structure. The depth map D_i is determined by the constraints on the polar geometry and photometric consistency, as shown in detail in Figure 3, where there are many ways to measure the similarity between the patches, such as the sum of the squared differences (SSD), sum of the absolute differences (SAD), normalized cross correlation (NCC) and other more advanced methods [65]. NCC is one of the most common and successful photo consistency measures used in multi-view stereo algorithms. It is invariant to changes in gain and bias of the pixel value, so it is mainly used to rebuild the scene of the real world. For a square domain B centered on pixel p and its corresponding domain B' centered on pixel p', the NCC score between p and p' is computed as:

$$NCC(p, p') = 1 - \frac{\sum_{q \in B, q' \in B'} (q - \overline{q})(q' - \overline{q}')}{\sqrt{\sum_{q \in B} (q - \overline{q})^2 \sum_{q' \in B'} (q' - \overline{q}')^2}} \tag{2}$$

where $\overline{q}$ represents the mean value of q, and $\overline{q}'$ represents the mean value of q'.

When reconstructing the weakly supported plane surface, we observe that while the depth inside the plane cannot be estimated, the depth of the plane edge can be accurately computed, and the NCC score is rather high, as shown in Figure 4. This is mainly because the plane edge is rich in distinguishable features. This observation shows us that we may renew the incomplete weakly supported plane by the depth map D_i and the NCC map (which we call the confidence map instead later in the text) C_i, as only three highly confident edge points are enough to span the plane. The important question now is how to determine the plane area.

B. PRECISE 2D PLANE SEGMENTATION

In fact, provided with the depth map D_i of a certain image I_i, if the projection mask M_i of the weakly supported plane is

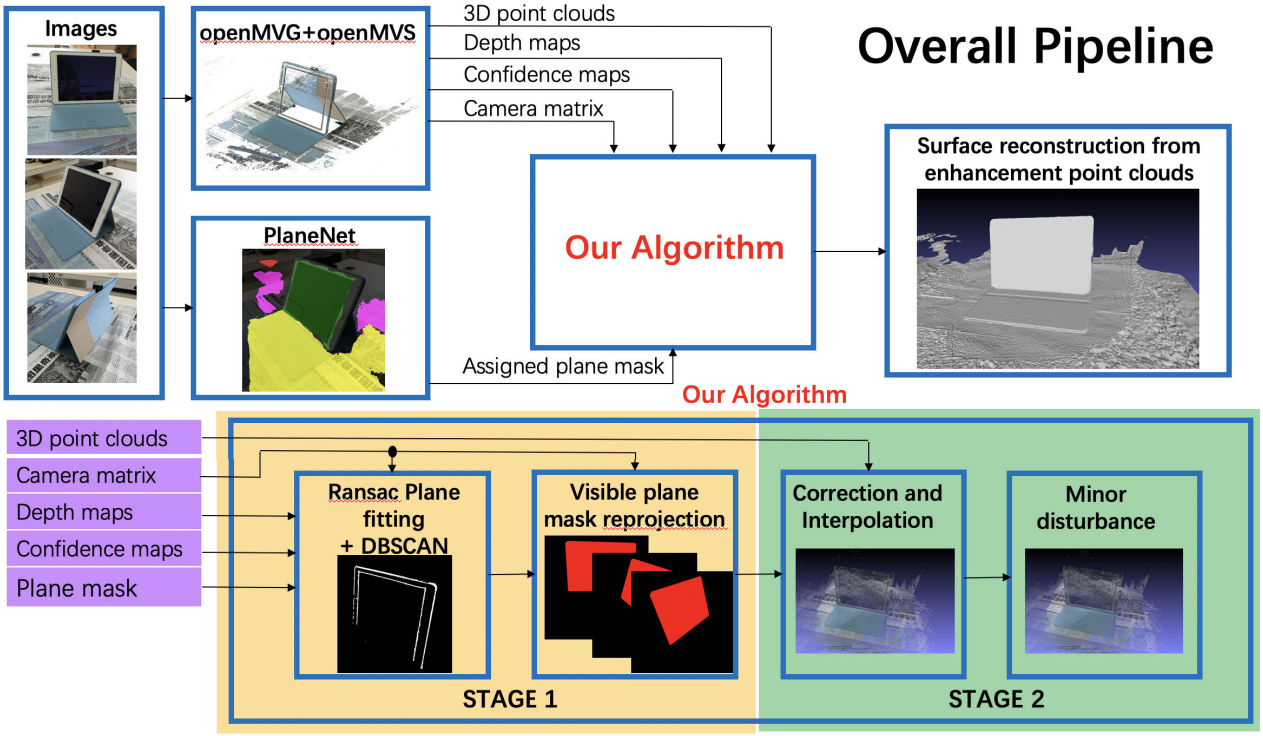

FIGURE 2. Overall pipeline of our proposed method.

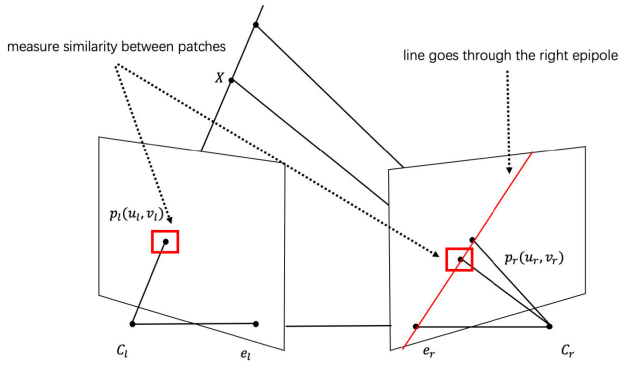

FIGURE 3. The depth of a pixel is determined by the constraints on the polar geometry and photometric consistency. Specifically, the depth is obtained by looking for a patch on the polar line most similar to the patch on the left.

known, then the 3D plane area can be easily calculated by reverse projection. This problem now turns into a problem of weakly supported planar instance segmentation in the image. Most recent approaches leverage convolutional neural networks (CNNs) and achieve state-of-the-art performance on multiple indoor and outdoor datasets in this task. After comparing several CNN-based methods, we choose PlaneNet because, through extensive experiments, we found that it has a better segmentation effect on real data. Readers can refer to Figure 11 in Section IV-D.1 to see the segmentation comparison results of different methods. However, due to the deviation between the real data and the training dataset, the segmentation result from the pretrained model is still not ideal, as shown in Figure 5. In most cases, the segmentation

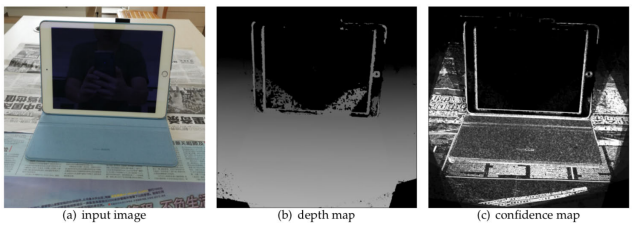

FIGURE 4. Depth map and confidence map priors on the 'iPad' data set. In both maps, the black color indicates that the value of this position is unknown.

mask is a partial section of the integral plane. It is not wise to rebuild a real dataset to fine-tune this model because this task is very large, and it is impossible to accommodate all of the scenes. Nevertheless, due to the prior information of the depth map D_i and the confidence map C_i, we can recover an accurate segmentation mask not only for the reference image I_i but also for the other visible images I_j by an uncomplicated operation.

For a certain registered image I_i containing a weakly supported plane, after piece-wise planar instance segmentation with CNNs, an inaccurate weakly supported plane mask $M_{i_{pre}}$ is obtained. In this mask, there are pixels with known depth value and high confidence that draw our attention. Suppose one of these pixels $p_i = I_i(u, v) \in M_{i_{pre}}$, for which the homogeneous coordinate is:

$$p_i = \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{3}$$

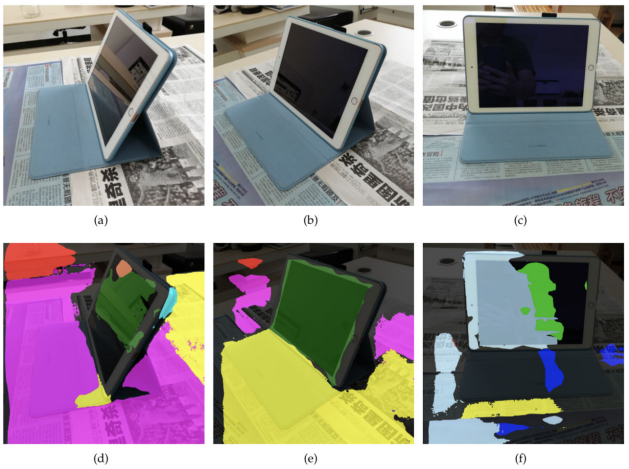

FIGURE 5. Pretrained PlaneNet segmentation results given RGB images (a)-(c) from the 'iPad' dataset, used to generate the three corresponding segmentation masks (d)-(f) according to the pretrained model.

The 3D point X_i must lie in the viewing ray of p_i. Given the depth value $\lambda_i = D_i(u, v)$ of this pixel p_i, X_i is computed in P_i's coordinates as:

$$X_i = \lambda_i K_i^{-1} p_i \tag{4}$$

Next, given the rotation matrix R_i and the camera center position C_i, X_i is transformed to the world coordinates X as:

$$X = R_i^T X_i + C_i \tag{5}$$

Through this process, hundreds or thousands of such points will be formed in space, which come from the mask $M_{i_{pre}}$. Unfortunately, the initial mask $M_{i_{pre}}$ covers not merely pixels inside the plane but also some outliers, making some generated points be distributed beyond the plane in 3D space. Due to the relatively large proportion of outliers, we adopt the RANSAC framework to fit the plane, and the plane parameters are estimated. We define the plane parameters as (A, B, C, D). For a 3D point $Q = (x, y, z)$ lying on this plane, we have $Ax + By + Cz - D = 0$. Then, we traverse the image I_i and back-project all of the pixels to space only if its depth value λ_i is known. According to the distance formula, the distance from point (x_0, y_0, z_0) to plane (A, B, C, D) in three-dimensional space is given by:

$$d = \frac{|Ax_0 + By_0 + Cz_0 - D|}{\sqrt{A^2 + B^2 + C^2}} \tag{6}$$

The points for which the distance d is less than a threshold t are considered to be on the estimated plane. As a result, in the image I_i, all of the pixels that meet the weakly supported plane surface parameters are captured. Since we did not set the spatial extent in the above steps, some points that are not on the weakly supported plane but satisfy the plane equation by chance are also extracted, as is shown in Figure 6. We filter these points by employing the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [66] method. The DBSCAN method is introduced in Algorithm 1. We treat the category with the largest number as the desired result.

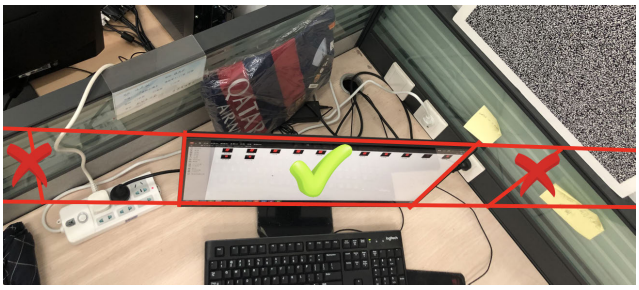

FIGURE 6. An example of mistaken point clouds accidentally lying on the target plane. The captured point clouds that do not belong to the display area are not what we need, and they should be removed.

We note that the edge of the plane is filled with significant features, and the edge point clouds are always well created. We re-project these clustered point clouds into image I_i to obtain their pixel positions x_i, and then, we obtain the rectified plane segmentation mask M_i using the envelope contour provided by openCV [67]. Moreover, we also project these point clouds into other visible images $\{I_j\}_{j\in Vis \bigcap j\neq i}$ to collect precise plane segmentation mask $\{M_j\}_{j\in Vis \bigcap j\neq i}$. The projection function is defined as:

$$x_i = P_i X = K_i R_i [I | - C_i] X \tag{7}$$

This plane mask information $\{M_i\}_{i\in Vis}$ is used to advance the subsequent operation.

C. CORRECTION AND INTERPOLATION

From experiments, we found that weakly supported plane reconstruction failure is due to two factors: first, some point clouds are built in a wrong location, and second, some point clouds are not built at all. To address these issues, we proposed two corresponding options, namely, correction and interpolation.

As shown in Figure 7, due to reasons such as specular reflection imaging, some point clouds that should be on the weakly supported plane are generated behind the surface. These kind of points have a strong negative effect. This occurs because even though a new point cloud is created at the weakly supported plane from the other view, it will be removed to obey the occlusion rule when conducting depth-map fusion. We decided to directly move these points to the weakly supported plane by perspective transformation. These points are selected based on a simple condition, i.e., whether they can be projected back into a mask area of a visible image.

Apparently, a correction in and of itself is not sufficient because such points only occupy a part of the whole weakly supported plane. We utilize linear interpolation to complete the rest of the point clouds that should have existed in the weakly supported plane region. The spatial boundary range of these interpolated points are found by the following steps. After leveraging the Sobel operator [68] for edge extraction on a certain mask map M_i, we inverse project these edge pixels to three-dimensional space to form the bound using depth map D_i. The interpolation density ρ is set to control

Algorithm 1 DBSCAN, a Density-Based Clustering Method

Input :
A dataset containing n objects D;
The radius parameter ϵ;
The field density threshold *Minpts*;
Output:
The density-based cluster sets;

```
mark all objects as unvisited;
while there are unvisited objects do
    randomly select an unvisited object p;
    mark p as visited;
    if there are at least Minpts objects in the ϵ field of p then
        create a new cluster C, and add p to C;
        let N be the set of all objects in the ϵ field of p;
        for each p′ in N do
            if p' is unvisited then
                mark p as visited;
                if there are at least Minpts objects in the ϵ field of p′ then
                    add these objects to N
                end
                if p′ is not a member of any cluster then
                    add p′ to C
                end
            end
        end
        output C;
    end
    else
        mark p as noise
    end
end
```

FIGURE 7. (a) Specular reflection imaging of a specular plane surface; (b) point clouds with black color that should be on the 'iPad' screen are generated behind the surface.

the sparseness of the created point clouds, which is measured by the pixel density projected onto image I_i.

D. MINOR DISTURBANCE OF PLANE POINT CLOUDS

We plan to directly generate plane meshes from the corrected and interpolated point clouds via surface reconstruction. To our surprise, although the reconstructed plane already has sufficient support in the point clouds, the reconstruction result is not perfect and is accompanied by a small number of holes. After careful analysis and consideration, we found that this problem may be caused by the fact that the plane is too flat to induce the building of the Delaunay tetrahedron. Therefore, we apply a minor disturbance to these plane point clouds in the direction of the plane normal vector. The disturbance distance e is defined on the basis of the interpolation density ρ, where $e = 0.1\rho$. Finally, the reconstructed mesh model of the weakly supported plane surface appears to be complete and error-free.

IV. EXPERIMENTS AND RESULTS

In this section, we present qualitative and quantitative experimental results and evaluations of our proposed approach on datasets that contain specular or textureless images. In Section IV-A, We briefly introduce the datasets information as well as the implementation details. In Section IV-B, we report 3D reconstruction effect on standard *DTU* dataset and we also perform an evaluation on real-world data in Section IV-C. At last, in Section IV-D, we provide two meaningful additional experiments. In the first experiment, we compare our segmentation refine results with several learning-based baselines, including PlaneNet [1] and an associative embedding-based method [36]. In the second experiment, we report on a number of ablation studies carried out to validate our method, where each action cannot be ignored.

A. DATASET AND IMPLEMENTATION DETAILS

Our approach works towards reconstructing practical scenes dominated by a weakly supported plane structure. We first evaluate our method on the *DTU* dataset [71]. This dataset consists of 80 different scenes of large variability. Each scene consists of 49 or 64 accurate camera positions and reference structured light scans, all acquired by a 6-axis industrial robot. Most importantly, the *DTU* dataset not only focuses on the Lambertian surface objects, but also provides some weakly textured cases that are just right for testing our methods. We select 5 suitable scans inside it for comparative experiments and evaluation. In addition, to prove the feasibility of our algorithm in real scenes, we decide to compile a real-world test set ourselves. We choose 5 objects common in life and take pictures of them using a cellphone from 25 different camera positions. The resolution of each image is 3840 × 3840. These objects are 'iPad', 'display','box','mirror' and 'charger'. The provided dataset contains non-Lambertian surfaces and texture-free structures, which is sufficient for proving the feasibility range of our method.

The proposed method has seven parameters, and we have discussed their settings in Section III. We give a summary of the set parameter values in Table 1.

B. BENCHMARKING ON DTU DATASET

We first evaluate our method on the 5 weakly supported plane surface scans of the *DTU* dataset. By adopting camera

TABLE 1. This table shows the results of the quantitative analysis of the generated point clouds of the scene.

Parameter	Section	Description	Value
m	III-B	minimum number of points to define the plane in RANSAC	3
w	III-B	percentage of inliers in the point cloud in RANSAC	0.6
p	III-B	probability that the best fitting model in one of the iterations in RANSAC	0.9
t	III-B	threshold value of the distance of points belonging to the selected model	0.01
ϵ	III-C	radius parameter in DBSCAN	0.1
$Minpts$	III-C	field density threshold in DBSCAN	100
ρ	III-C	interpolation density	1/4 pixels

TABLE 2. This table shows the results of the quantitative analysis of the generated point clouds of the *DTU* dataset.

Method.	Mean Acc.	Mean Comp.	Med Acc.	Med Comp.	Overall
Camp [69]	0.7458	1.2664	0.5094	**0.2958**	1.0061
Furu [18]	0.6514	1.9670	0.6514	0.7302	1.3092
Tola [70]	**0.3606**	1.9819	**0.2446**	0.6749	1.1712
Shen [53]	0.4982	1.3812	0.2459	0.5237	0.9397
Ours	0.4709	**1.0881**	0.2740	0.4608	**0.7795**

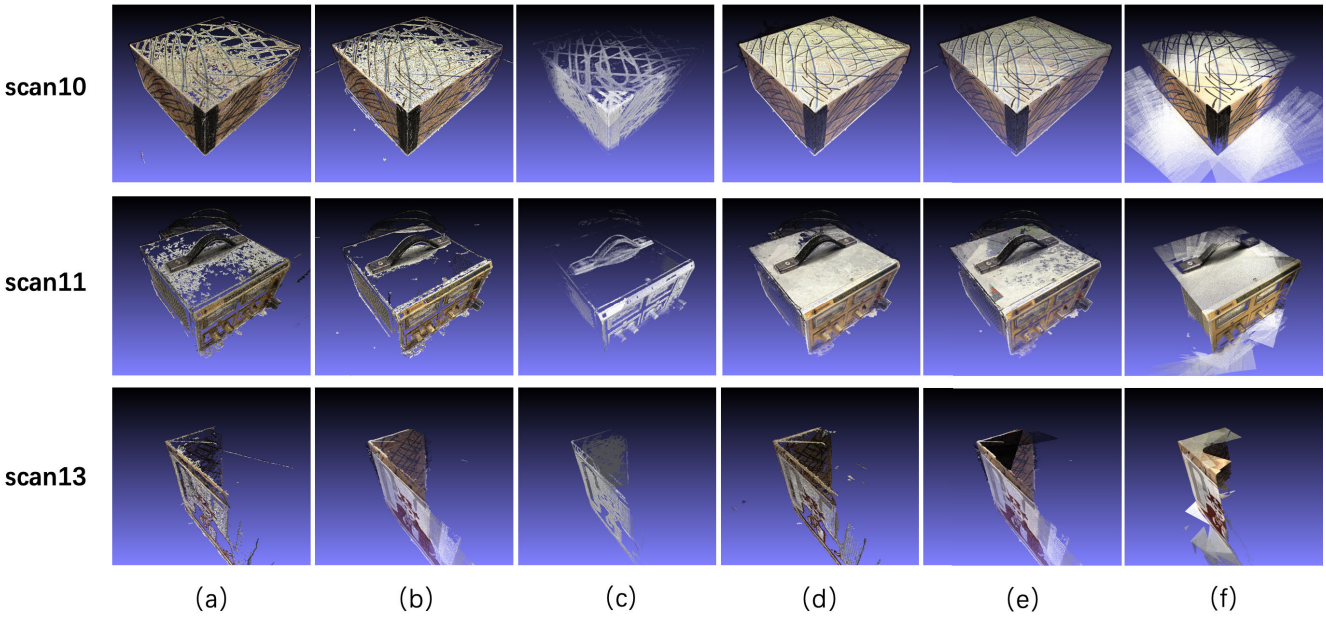

FIGURE 8. Qualitative results of scans 10, 11 and 13 of *DTU* dataset between our method and other MVS methods. (a) Camp [69], (b) Furu [18], (c) Tola [70], (d) Shen [53], (e) our method and (f) ground truth. Our method generates the most complete point clouds especially in those textureless areas.

calibration(internal and external) information from dataset for MVS reconstruction, we ensure that the generating point cloud result is in the same coordinate system with the ground truth point cloud. For quantitative evaluation, we calculate the *accuracy* and the *completeness* of the distance metric [72]. *Accuracy* is measured as the distance from the MVS reconstruction to the structured light reference, and the *completeness* is measured from the reference to the MVS reconstruction. For each reconstruction, the distances of 3D point are condensed into comparable statistics by computing the mean and median for the *accuracy* and *completeness*. This is, however, first done by removing all distances over 20 mm to avoid biasing by outliers.

The MVS methods of Campbell *et al.* [69], Furukawa and Ponce [18], Tola *et al.* [70], and Shen [53] have been evaluated by comparing the point clouds. A summary of the overall quantitative performance is shown in Table 2. While Tola *et al.* [70] performs best in the accuracy, our approach outperforms all other methods in both the completeness and the overall quality **with a significant margin**. As shown in Figure 8, our method corrects and generates point clouds in the weakly supported plane region which leads to better completeness.

C. RECONSTRUCTION ON REAL-WORLD DATASET

The *DTU* scans are taken under well-controlled indoor environment with fixed camera trajectory. To further demonstrate the generalization ability of our method, we test the proposed method on the more complex real world dataset. We perform comparisons with state-of-the-art large-scale

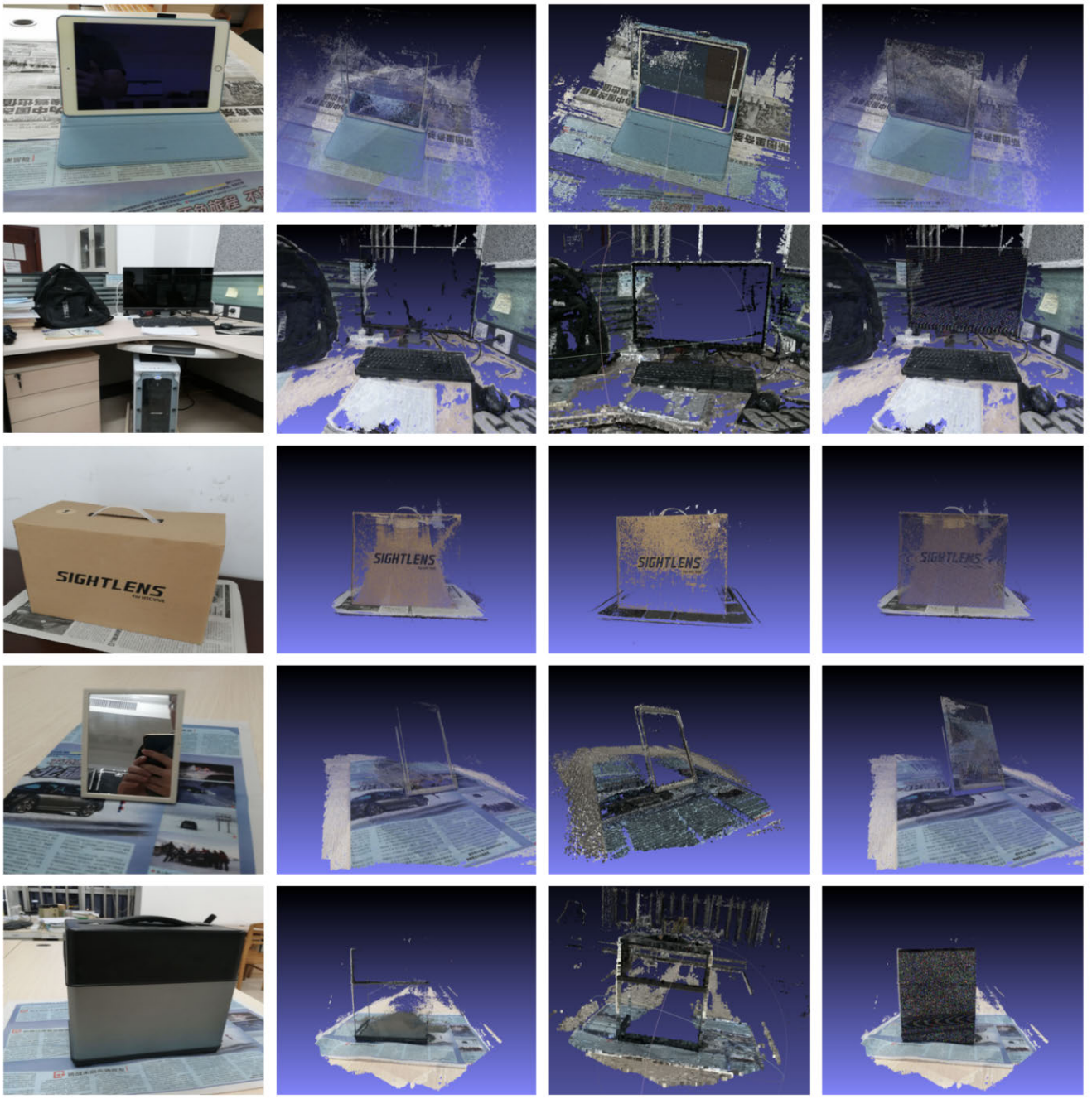

FIGURE 9. Qualitative point cloud comparisons between our method and well-known multi-view stereo approaches. (a) Input images, (b) dense point clouds recovered with openMVS [53], (c) dense point clouds recovered with COLMAP [16], and (d) dense point clouds recovered through our correction and interpolation of the openMVS result.

reconstruction systems [16], [53] and the most recently reported approach [30] designed for preserving weakly supported surfaces. These methods can effectively reconstruct real-life 3D scenes.

We use the number of reconstructed point clouds as an indicator to quantitatively evaluate the reconstruction completeness. The number of generated point clouds is listed in Table 3. Our proposed method has the most point clouds, implying better integrity. The reconstructed dense point clouds of different methods are shown in Figure 9. We found that both COLMAP and openMVG+openMVS cannot produce point clouds in the weakly supported plane. The point clouds that should have been generated in the weakly supported plane disappear in pieces. Our method instead creates

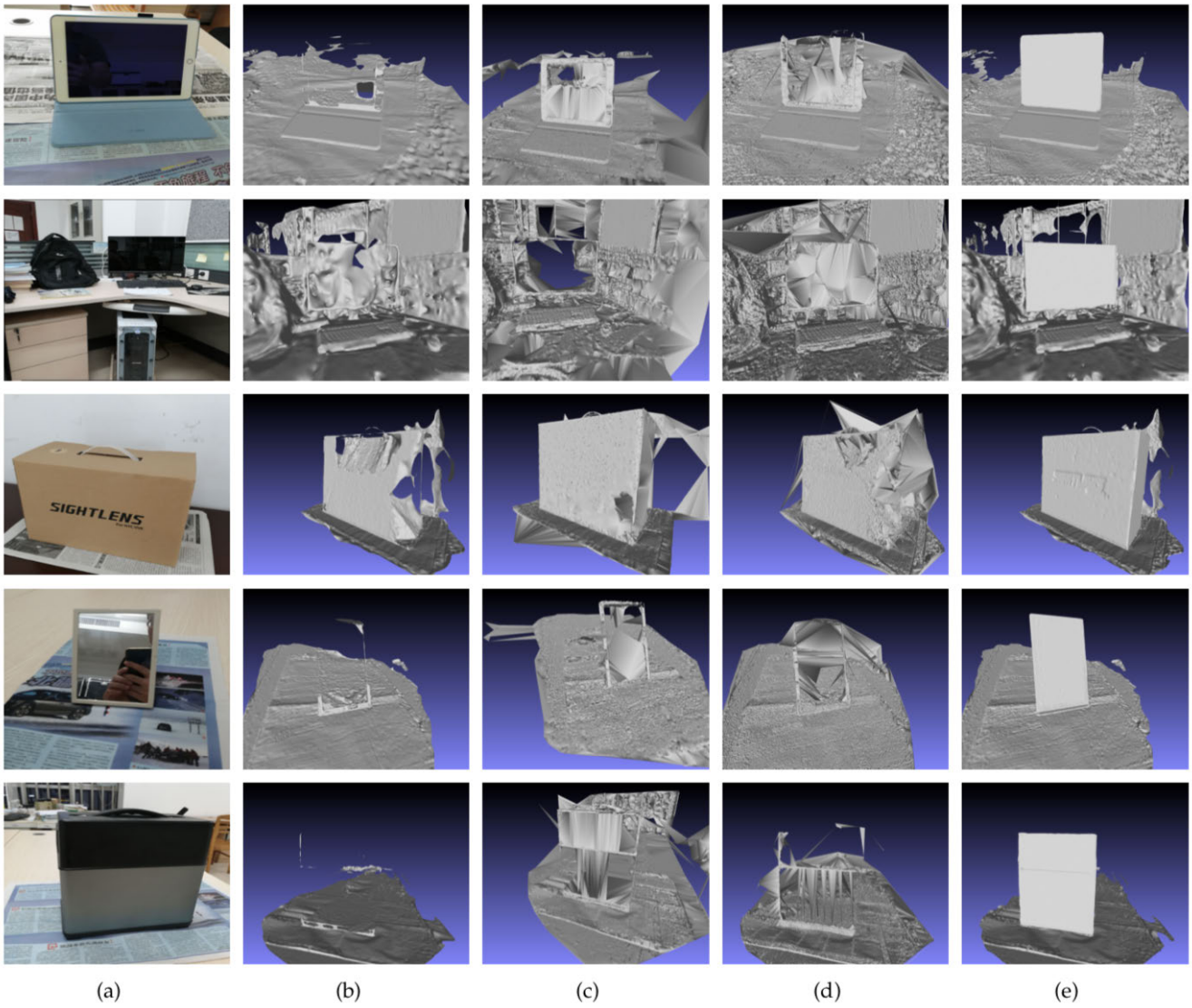

FIGURE 10. Qualitative mesh reconstruction comparisons between our method and other methods aimed at weakly supported surfaces. (a) Input images, (b) mesh reconstruction with openMVS [53], (c) mesh reconstruction with COLMAP [16], (d) mesh reconstruction with s-t cut [30], and (e) mesh reconstruction with our method.

TABLE 3. This table shows the results of the quantitative analysis of the generated point clouds of the scene.

Data	OpenMVS [55]	COLMAP [45]	Ours
iPad	11, 392, 051	1, 811, 062	**11, 663, 592**
Display	3, 280, 502	786, 294	**3, 343, 622**
Box	2, 920, 123	595, 883	**3, 251, 985**
Mirror	4, 928, 217	786, 223	**5, 050, 287**
Charger	4, 675, 291	873, 241	**5, 068, 105**

complete point clouds in such regions. We attribute the point cloud completeness in the weakly supported plane surface to the use of the correction and interpolation methods.

We introduce the mesh reconstruction results obtained by building on the point clouds in Figure 10. The mesh reconstructions of COLMAP and openMVS are filled with mistakes and holes. Although the s-t cut based method alleviates this problem, the obtained results still show much room for improvement. Generally, these reconstructed plane mesh results are too poor to see. Only our proposed method can reconstruct the flat and complete weakly supported plane surface. These results prove that our method can effectively solve the problem of the inability to generate weakly supported planes in real scenes.

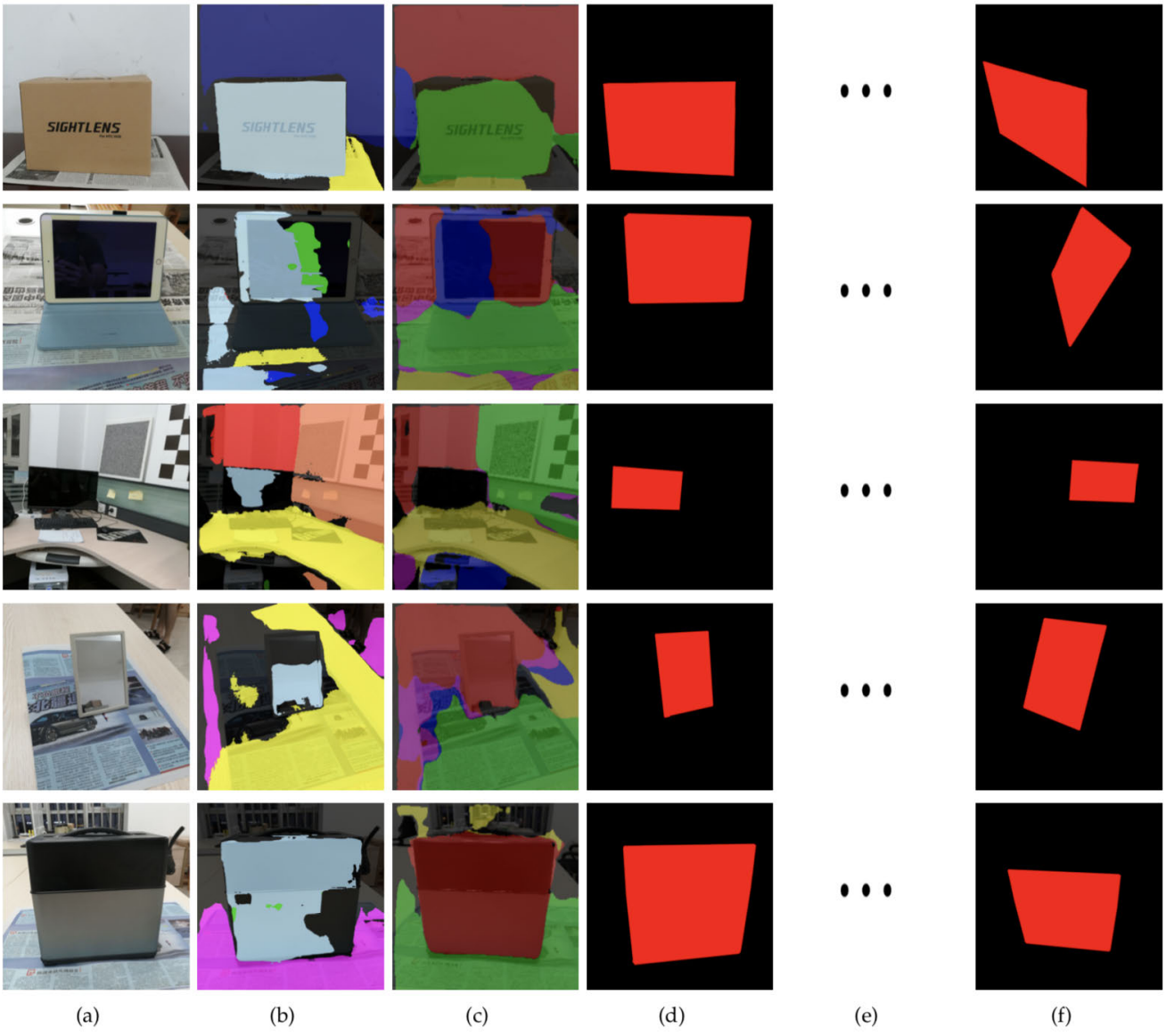

FIGURE 11. Qualitative segmentation comparisons between our refinement method and existing plane segmentation methods. (a) Input images, (b) plane segmentation results for PlaneNet [1], and (c) plane segmentation results for the associative embedding-based method [36]; (d)(e)(f) are weakly supported plane masks of the visible images after the manual assignment and refinement process.

D. ADDITIONAL STUDIES

1) SEGMENTATION REFINEMENT

To evaluate the benefit of our proposed segmentation refinement operation, we perform a comparison with plane segmentation baselines by implementing a pretrained model. PlaneNet [1] and the associative embedding-based method [36] are used in the comparison as the state-of-the-art open source algorithms on standard datasets. In Figure 11, we show the segmentation results. The segmented plane mask using the pretrained network is completely unusable, but the results obtained by PlaneNet are relatively better, and we therefore take them as priors. After manually specifying the weakly supported plane and effectively refining, the segmentation area is much more accurate. The results of this experiment fully prove that our proposed refinement operation is useful and necessary.

2) ABLATION STUDIES

To evaluate the effectiveness of the entire process, we show the reconstruction results step by step (with correction, interpolation and minor disturbance) in this experiment. As visualized in Figure 12, each step plays an important role in restoring the complete weakly supported plane surface. In Figure 12 (a), the reconstructed plane surface meshes are completely empty. In Figure 12 (b), some of the faithful meshes have been built. In Figure 12 (c), while more meshes are generated on the plane, the result of the reconstruction still retains a small part of the cavity. In Figure 12 (d), the results of the reconstruction are complete and smooth.

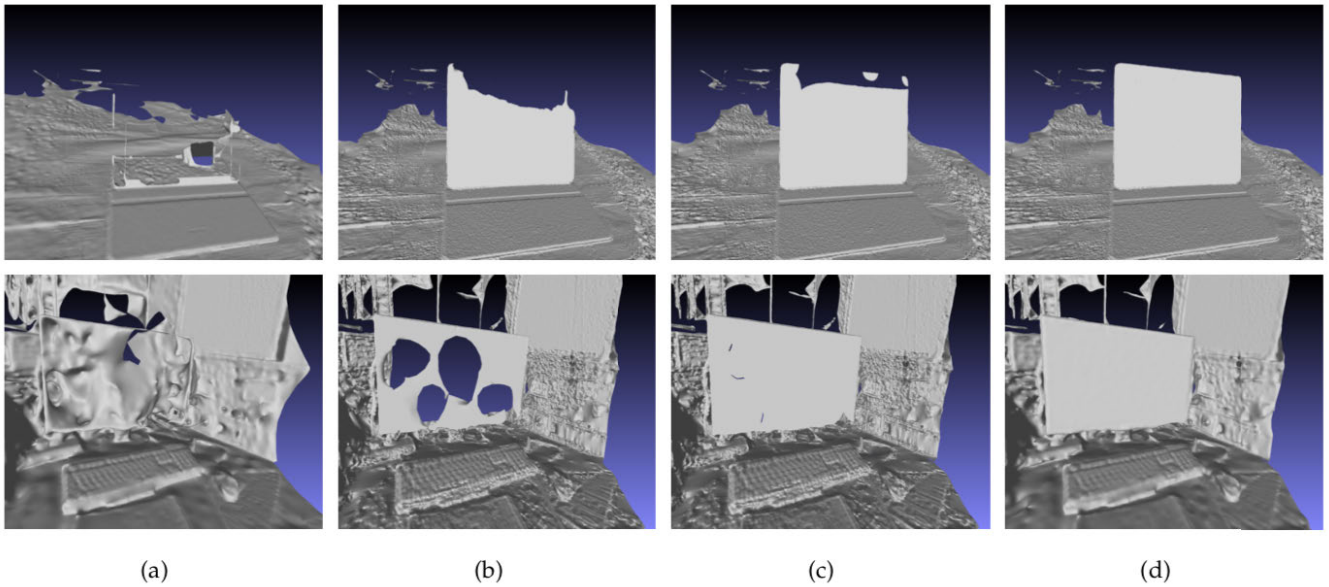

FIGURE 12. Ablation study of the reconstruction pipeline on the 'iPad' and 'display' test sets. (a) Mesh reconstruction without any operations, (b) mesh reconstruction with only correction, (c) mesh reconstruction with correction and interpolation, and (d) mesh reconstruction with correction and interpolation plus a minor disturbance.

V. CONCLUSION

In this paper, we propose a point cloud enhancement method for 3D reconstruction of weakly supported plane surfaces. Based on correcting and integrating inaccurate prior information from a pretrained CNN model and depth-map merging methods, we successfully repair incorrectly generated points and further interpolate more points on the weakly supported plane surface. The proposed approach significantly outperforms state-of-the-art MVS systems in the weakly supported plane surface reconstruction task. It also advances recent methods focusing on preserving weakly supported surfaces. An interesting future direction is to go beyond the plane hypothesis and tackle the structured geometry prediction problems of an arbitrary surface shape.

ACKNOWLEDGMENT

(Shen Yan and Yang Peng are co-first authors.)

REFERENCES

[1] C. Liu, J. Yang, D. Ceylan, E. Yumer, and Y. Furukawa, "PlaneNet: Piecewise planar reconstruction from a single RGB image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2579–2588.

[2] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski, "Building Rome in a day," *Commun. ACM*, vol. 54, no. 10, pp. 105–112, Oct. 2011.

[3] J.-M. Frahm, P. Fite, G. D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, and S. Lazebnik, "Building rome on a cloudless day," in *Proc. 11th Eur. Conf. Comput. Vis. (ECCV)*, Crete, Greece, Sep. 2010, pp. 368–381, doi: 10.1007/978-3-642-15561-1_27.

[4] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, 2006.

[5] J. Heinly, J. L. Schonberger, E. Dunn, and J.-M. Frahm, "Reconstructing the world* in six days*(as captured by the yahoo 100 million image dataset)," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3287–3295.

[6] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from Internet photo collections," *Int. J. Comput. Vis.*, vol. 80, no. 2, pp. 189–210, 2007.

[7] C. Strecha, W. Von Hansen, L. Van Gool, P. Fua, and U. Thoennessen, "On benchmarking camera calibration and multi-view stereo for high resolution imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[8] S. Ullman, "The interpretation of structure from motion," *Proc. Roy. Soc. London B, Biol. Sci.*, vol. 203, no. 1153, pp. 405–426, Jan. 1979.

[9] M. J. Westoby, J. Brasington, N. F. Glasser, M. J. Hambrey, and J. M. Reynolds, "'Structure-from-Motion' photogrammetry: A low-cost, effective tool for geoscience applications," *Geomorphology*, vol. 179, pp. 300–314, Dec. 2012.

[10] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4104–4113.

[11] O. Özyeşil, V. Voroninski, R. Basri, and A. Singer, "A survey of structure from motion," *Acta Numer.*, vol. 26, pp. 305–364, May 2017.

[12] C. Wu, "Towards linear-time incremental structure from motion," in *Proc. Int. Conf. 3D Vis. (3DV)*, Jun. 2013, pp. 127–134.

[13] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2006, pp. 519–528.

[14] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz, "Multi-view stereo for community photo collections," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[15] M. Goesele, B. Curless, and S. M. Seitz, "Multi-view stereo revisited," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 2402–2409.

[16] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 501–518, doi: 10.1007/978-3-319-46487-9_31.

[17] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3260–3269.

[18] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2008.

[19] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proc. 4th Eurograph. Symp. Geometry Process.*, vol. 7, 2006, pp. 1–10.

[20] P. Su and R. L. S. Drysdale, "A comparison of sequential delaunay triangulation algorithms," *Comput. Geometry*, vol. 7, nos. 5–6, pp. 361–385, 1997.

[21] J. R. Shewchuk, "Triangle: Engineering a 2D quality mesh generator and Delaunay triangulator," in *Proc. Workshop Appl. Comput. Geometry (WACG)*, Philadelphia, PA, USA, May 1996, pp. 203–222, doi: 10.1007/BFb0014497.

[22] M. Berger, A. Tagliasacchi, L. M. Seversky, P. Alliez, G. Guennebaud, J. A. Levine, A. Sharf, and C. T. Silva, "A survey of surface reconstruction from point clouds," *Comput. Graph. Forum*, vol. 36, no. 1, pp. 301–329, 2017.

[23] Y. Furukawa and C. Hernández, "Multi-view stereo: A tutorial," *Found. Trends Comput. Graph. Vis.*, vol. 9, nos. 1–2, pp. 1–148, 2015.

[24] R. J. Woodham, "Photometric stereo: A reflectance map technique for determining surface orientation from image intensity," *Proc. SPIE*, vol. 155, pp. 136–143, Jan. 1979.

[25] M. Tarini, H. P. A. Lensch, M. Goesele, and H.-P. Seidel, "3D acquisition of mirroring objects using striped patterns," *Graph. Models*, vol. 67, no. 4, pp. 233–259, 2005.

[26] S.-K. Tin, J. Ye, M. Nezamabadi, and C. Chen, "3D reconstruction of mirror-type objects using efficient ray coding," in *Proc. IEEE Int. Conf. Comput. Photogr. (ICCP)*, May 2016, pp. 1–11.

[27] S. Büyükatalay, Ö. Birgül, and U. Halici, "Surface reconstruction from multiple images filtering non-Lambert regions," in *Proc. 16th IEEE Int. Conf. Image Process. (ICIP)*, Nov. 2009, pp. 533–536.

[28] S. P. Mallick, T. E. Zickler, D. J. Kriegman, and P. N. Belhumeur, "Beyond lambert: Reconstructing specular surfaces using color," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2005, pp. 619–626.

[29] S. Wu, H. Huang, T. Portenier, M. Sela, D. Cohen-Or, R. Kimmel, and M. Zwicker, "Specular-to-diffuse translation for multi-view reconstruction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 183–200.

[30] M. Jancosek and T. Pajdla, "Exploiting visibility information in surface reconstruction to preserve weakly supported surfaces," *Int. Scholarly Res. Notices*, vol. 2014, Aug. 2014, Art. no. 798595.

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

[32] S. N. Sinha, P. Mordohai, and M. Pollefeys, "Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[33] M. Jancosek and T. Pajdla, "Multi-view reconstruction preserving weakly-supported surfaces," in *Proc. IEEE CVPR*, Jun. 2011, pp. 3121–3128.

[34] C.-H. Lin, C. Kong, and S. Lucey, "Learning efficient point cloud generation for dense 3D object reconstruction," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[35] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik, "Multi-view supervision for single-view reconstruction via differentiable ray consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2626–2634.

[36] Z. Yu, J. Zheng, D. Lian, Z. Zhou, and S. Gao, "Single-image piece-wise planar 3D reconstruction via associative embedding," 2019, *arXiv:1902.09777*. [Online]. Available: https://arxiv.org/abs/1902.09777

[37] F. Yang and Z. Zhou, "Recovering 3D planes from a single image via convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 85–100.

[38] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DOF camera relocalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2938–2946.

[39] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "LIFT: Learned invariant feature transform," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 467–483, doi: 10.1007/978-3-319-46466-4_28.

[40] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 224–236.

[41] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, Jul. 2019.

[42] R. Zhu, C. Wang, C.-H. Lin, Z. Wang, and S. Lucey, "Object-centric photometric bundle adjustment with deep shape prior," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 894–902.

[43] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[44] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—A modern synthesis," in *Proc. Int. Workshop Vis. Algorithms (ICCV)*, Corfu, Greece, Sep. 1999, pp. 298–372, doi: 10.1007/3-540-44480-7_21.

[45] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz, "Multicore bundle adjustment," in *Proc. CVPR*, Jun. 2011, pp. 3057–3064.

[46] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[47] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. CVPR*, vol. 4, Jul. 2004, pp. 506–513.

[48] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. 9th Eur. Conf. Comput. Vis. (ECCV)*, Graz, Austria, May 2006, pp. 404–417, doi: 10.1007/11744023_32.

[49] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *Proc. VISAPP*, vol. 2, 2009, pp. 331–340.

[50] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 2161–2168.

[51] Y. Hao, F. Zhu, J. Ou, Q. Wu, J. Zhou, and S. Fu, "Robust analysis of P3P pose estimation," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2007, pp. 222–226.

[52] H. Hirschmüller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8. doi: 10.1109/CVPR.2007.383248.

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

[54] D. Tian, P. B. Pandit, and P. Yin, "Refined depth map," U.S. Patent 9 179 153, Nov. 3, 2015.

[55] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.-M. Frahm, R. Yang, and D. Nistér, and M. Pollefeys, "Real-time visibility-based fusion of depth maps," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[56] P. Moulon, P. Monasse, R. Perrot, and R. Marlet, "OpenMVG: Open multiple view geometry," in *Proc. 1st Int. Workshop Reproducible Res. Pattern Recognit. (RRPR@ICPR)*, Cancún, Mexico, Dec. 2016, pp. 60–74, doi: 10.1007/978-3-319-56414-2_5.

[57] R. Or-El, R. Hershkovitz, A. Wetzler, G. Rosman, A. M. Bruckstein, and R. Kimmel, "Real-time depth refinement for specular objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4378–4386.

[58] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Multi-view 3D models from single images with a convolutional network," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 322–337, doi: 10.1007/978-3-319-46478-7_20.

[59] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3D-R2N2: A unified approach for single and multi-view 3D object reconstruction," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 628–644, doi: 10.1007/978-3-319-46484-8_38.

[60] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, nos. 1–32, p. 2, 2016.

[61] W. Hartmann, S. Galliani, M. Havlena, L. Van Gool, and K. Schindler, "Learned multi-patch similarity," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1586–1594.

[62] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 767–783.

[63] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, "DeepMVS: Learning multi-view stereopsis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2821–2830.

[64] C. Liu, K. Kim, J. Gu, Y. Furukawa, and J. Kautz, "PlaneRCNN: 3D plane detection and reconstruction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4450–4459.

[65] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[66] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, vol. 96. 1996, pp. 226–231.

[67] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision With the OpenCV Library*. Newton, MA, USA: O'Reilly Media, 2008.

[68] S. Gupta and S. G. Mazumdar, "Sobel edge detection algorithm," *Int. J. Comput. Sci. Manage. Res.*, vol. 2, no. 2, pp. 1578–1583, Feb. 2013.

[69] N. D. F. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, "Using multiple hypotheses to improve depth-maps for multi-view stereo," in *Proc. 10th Eur. Conf. Comput. Vis. (ECCV)*, Marseille, France, Oct. 2008, pp. 766–779, doi: 10.1007/978-3-540-88682-2_58.

[70] E. Tola, C. Strecha, and P. Fua, "Efficient large-scale multi-view stereo for ultra high-resolution image sets," *Mach. Vis. Appl.*, vol. 23, no. 5, pp. 903–920, 2012.
[71] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs, "Large scale multi-view stereopsis evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 406–413.
[72] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 153–168, 2016.

SHEN YAN received the B.S. and M.S. degrees in system engineering from the National University of Defense Technology, Changsha, China, in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree in system engineering. His current research interests include image-based 3D reconstruction and computer vision.

YANG PENG received the B.S. and Ph.D. degrees in system engineering from the National University of Defense Technology, Changsha, China, in 2012 and 2017, respectively. He is currently a Lecturer with the Department of System Engineering, National University of Defense Technology. His research interests include single pixel camera and sparsity cluster regularization.

GUANGYUE WANG received the B.S. degree in system engineering from the Zhengzhou Aviation Industry Management College, Zhengzhou, China, in 2017. He is currently pursuing the M.S. degree in system engineering from the National University of Defense Technology. His current research interests include image-based 3D reconstruction and machine learning.

SHIMING LAI received the B.S. and Ph.D. degrees in system engineering from the National University of Defense Technology, Changsha, China, in 2008 and 2014, respectively. He is currently a Lecturer with the Department of System Engineering, National University of Defense Technology. His research interests include computer vision, computational photography, and imaging systems.

MAOJUN ZHANG received the B.S. and Ph.D. degrees in system engineering from the National University of Defense Technology, Changsha, China, in 1992 and 1997, respectively. He is currently a Professor with the Department of System Engineering, National University of Defense Technology. His current research interest include computer vision, information system engineering, and system simulation.