International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

DeepPCF-MVS: Deep Plane Estimation and Filtering for Complete Multi-View Stereo

Andreas Kuhn · Christian Sormann · Tarek Bouamer · Shan Lin · Oliver Erdler · Friedrich Fraundorfer

Received: date / Accepted: date

Abstract Multi-View Stereo (MVS)-based 3D reconstruction is a major topic in computer vision for which a vast number of methods have been proposed over the last decades showing impressive visual results. Long-since, benchmarks like Middlebury [45] numerically rank the individual methods considering accuracy and completeness as quality attributes. While the *Middlebury* benchmark provides low-resolution images only, the recently published *ETH3D* [44] and *Tanks and Temples* [23] benchmarks allow for an evaluation of high-resolution and large-scale MVS from natural camera configurations. This benchmarking reveals that still only few methods can be used for the reconstruction of large-scale models. We present an effective pipeline for large-scale 3D reconstruction which extends existing methods in several ways: (i) We introduce an outlier filtering considering the MVS geometry and make use of machine-learned confidences for filtering [30]. (ii) To avoid incomplete models

Andreas Kuhn
Sony Europe B.V. - Stuttgart Technology Center
E-mail: andreas.kuhn@sony.com

Christian Sormann
Graz University of Technology
E-mail: christian.sormann@icg.tugraz.at

Tarek Bouamer
Graz University of Technology
E-mail: tarek.bouamer@icg.tugraz.at

Shan Lin
Sony Europe B.V. - Stuttgart Technology Center
E-mail: shan.lin@sony.com

Oliver Erdler
Sony Europe B.V. - Stuttgart Technology Center
E-mail: oliver.erdler@sony.com

Friedrich Fraundorfer
Graz University of Technology
E-mail: fraundorfer@icg.tugraz.at

from local matching methods we propose a plane completion method based on growing superpixels allowing a generic generation of high-quality 3D models. We show further improvements by utilizing plane detections from a deep neural network [33] in addition to superpixel segmentation masks to generate improved plane-based segmentation masks. (iii) Finally, we use deep learning for a subsequent filtering of outliers in segmented sky areas. We give experimental evidence on benchmarks that our contributions improve the quality of the 3D model and our method is state-of-the-art in high-quality 3D reconstruction from high-resolution images or large image sets.

Keywords Multi-View Stereo · 3D Reconstruction · Plane Estimation

1 Introduction

Benchmarking 3D reconstruction from real-world high-resolution images has been absent to the community since the unavailability of Strecha et al.'s [47] online service. Even though ground truth models are available for a subset of the datasets, their images basically show well-textured scenes from specific and simple camera configurations. For example, the objects are captured from a constant distance. Similar limited configurations are provided by the *DTU Robot Image* [1] and *Middlebury* [45] datasets, both consisting of images captured in a laboratory environment with relatively low resolutions.

The importance of employing a large variety of scenes and viewpoints in different kind of environments is demonstrated by the recently published benchmark datasets *ETH3D* [44] and *Tanks and Temples* [23]. Both datasets comprise ground truth models generated from high-precision laser scanners. For specific training data the ground truth is publicly available. The *ETH3D* [44] dataset provides images that have been registered and aligned with the laser scans while the *Tanks and Temples* dataset [23] provides the image data only needing a preceding estimation of the camera calibration by means of Structure from Motion (SfM) [46,35,51,41,42]. Since data alignment is a non-trivial task, we use the *ETH3D* training dataset for the evaluation of the proposed contributions in our experiments. Finally, we show our evaluation results on the *ETH3D* Test and *Tanks and Temples* Test and Training datasets from our proposed 3D reconstruction pipeline.

1.1 Related Work

In this paper, we focus on methods allowing a high completeness of 3D models while still preserving details. **CMPMVS** [21] reconstructs surfaces (meshes) in a tetrahedral space [31] derived from noisy point clouds using visibility constraints. CMPMVS is especially strong in the reconstruction of surface parts that have not been directly derived in the Multi-View Stereo (MVS) step. For MVS an efficient plane-sweep approach is used [20]. The underlying optimization scheme has already been shown to be suitable for large-scale 3D reconstruction [50,37]. Instead of tetrahedralization, we reconstruct point clouds allowing a higher level of detail without a complex optimization. **COLMAP** [42,43] is used widely for effective point cloud reconstruction from images with unrestricted configurations. Impressive results were shown with diverse datasets like Photo Community Collections [42], laboratory data [45] and high-resolution imagery [43] One of COLMAPs major contributions is the depth map generation based on PatchMatch (PM) [3,55] including statistics for correspondence search in multiple images resulting in a higher efficiency [43]. Finally, it uses a geometric fusion of noisy depth maps into clean point clouds. The resulting point cloud can be transformed into surface meshes, e.g., by means of Poisson reconstruction [22]. **Gipuma** [14] also provides a GPU implementation of PM stereo matching including a pixel-wise normal estimation. For an independent parallelization a checkerboard propagation scheme is proposed leading to a higher efficiency of the method. Depth maps are fused into a single point cloud by averaging over consistent depth and normal estimates. The Gipuma method is ranked lower in the *ETH3D* benchmark compared to alternative PM methods. **ACMH** [52] demonstrates that the checkerboard sampling allows high-quality reconstruction when using a multi-hypothesis joint view selection. When further employing a multi-scale geometric consistency guidance (**ACMM** [52]) state-of-the art quality is achieved. To this end, the completeness is improved by multi-scale geometric consistency guidance for propagating depth measurements from lower resolution levels. **DeepC-MVS** [30] also combines checkerboard sampling with multi-scale processing for PM [52] including plane-based propagation of the depth hypothesis [43]. For an improved outlier filtering a deep-network is introduced capable of predicting pixel-wise confidences which are subsequently used for outlier filtering and depth map refinement [40]. **LTVRE** [27] uses Semi-Global Matching (SGM) [16] to generate disparity maps and derives a pixel-wise quality estimate using a Total Variation (TV) criterion [26]. The TV criterion has also been successfully applied to PM for stereo images [29]. One key contribution of LTVRE is a 3D error estimation and probabilistic fusion and filtering [25] which which improves standard local volumetric fusion while still allowing a high scalability [24]. Semi-globally optimized depth maps provide a higher completeness but are limited by their fronto-parallel assumptions. **MVE** [11] is also based on a volumetric fusion of implicit functions from 3D point clouds including point-wise quality values [38] [9] [10]. In contrast to linear one dimensional functions, MVE considers a 3D error assigning values to voxels in a spatial neighborhood. Depth maps are estimated with a region-growing approach [15] which is less effective since it lacks completeness as it can be demonstrated with the *ETH3D* benchmark. **PMVS** [13] also employs sparse features expanded by means of a region-growing approach. The semi-dense point clouds are filtered for obtaining a higher accuracy. PMVS does show a similar relatively low quality as other methods based on region growing MVS on the *ETH3D* benchmarks. **TAPA-MVS** [39] recently demonstrated that the completeness of PM depth maps can be significantly improved by applying depth completion with plane-fitting on superpixel level as additional hypothesis.

This paper describes an extended version of PCF-MVS [28]. Our method uses PM depth maps as input for a plane-based completion on superpixels. In contrast to [28], we use the improved PM pipeline and outlier filtering proposed in [30]. We extend TAPA-MVS by introducing hierarchical superpixel clustering and adaption of the plane estimation to generic MVS where no scale of the scene is available. To this end, we make use of LTVREs error propagation. In addition, we apply the plane fitting as a post-processing to PM depth maps and demonstrate a significant improvement for the completeness of 3D models on the *ETH3D* benchmark while still preserving high resolution details.

As an extension to PCF-MVS [28] we additionally integrate planar segments derived from a DNN as geometrical cues from a single RGB image. Recently, deep learning-based methods have shown promising results in reconstructing textureless surfaces from single images. PlaneNet [34] and SVPNet [54] introduced a plane segmentation task with deep learning to model man-made 3D scenes. These networks successfully learn to generate plane masks and plane parameters from ground truth depth maps, however they fail to detect flat surfaces with a small receptive field and are limited to a maximum number of detections. PlaneNet [34], for instance, can detect up to 10 planes per view. PlaneRCNN [33] addresses these two issues and proposes a novel architecture that jointly detects the plane parameters and the corresponding instance segmentation mask. Additionally the network refines the plane detections during the training and enforces geometric consistency between the views for better segmentation recall. PlaneRCNN outperforms PlaneNet and SVPNet in terms of number of detections, accuracy and generalizability.

2 Review of Suitable 3D Reconstruction Methods

As said, our method is based on ideas from MVS, error propagation, plane detection and depth completion. In this section we give a summary and analysis of employed methods.

2.1 Multi-View Stereo

There are promising deep-learning-based stereo methods [18, 53] for accurate depth map generation. However, they are limited with large scenes because they generate very large 3D cost volumes as input strongly limiting the applicability for large disparity ranges. A direct comparison on the high-resolution *ETH3D* benchmark is given by [39] demonstrating their shortcomings. We focus on methods which are feasible for high-resolution image processing.

Benchmarking shows better results for MVS estimation with PM and stereo estimation with SGM than for region-growing methods. SGM performs well on scalability, as only two images (stereo) have to be processed at a time. The final (MVS) disparity map is derived by means of pixel-wise fusion from multiple stereo disparity maps. In general, PM is feasible for processing high-resolution images as the runtime complexity increases only linearly with the image resolution (in overall pixels N), while SGM has an polynomial complexity ($\mathcal{O}(N^{1.5})$) considering the additional disparity dimension. On the other hand, SGM does not need a pixel-wise propagation of each pixel assigned to a patch neighborhood for all iterations as it only needs a scalar representation of pixel neighborhoods using Census matching costs [17]. Hence, PM needs efficient implementations, e.g., on a GPU where memory resources are limited. However, its unrestricted patch-based nature results in a higher quality reconstruction as it does not imply strong geometric priors like the fronto-parallel assumption in SGM. Therefore, we selected PM for depth-map estimation as we found that a standard GPU is sufficient to process the *ETH3D* datasets in full resolution of 25MP.

We extend the COLMAP [43] and DeepC-MVS [30] PM processes by calculating the average baseline b for multi-view configurations which is used for the subsequent completion and fusion. COLMAP [43] and DeepC-MVS [30] PM employ a final consistency checking by projecting pixel-wise estimated depths into source images. If the re-projection error and a NCC-based photometric consistency is below a threshold the pixel passes the consistency checks and is marked valid. For each pixel p in each source image i we use the finally matched pairs to estimate an average baseline:

$$b_i^p = \frac{1}{|J_i^p|} \sum_{j \in J_i^p} b_{i,j},\ J_i^p \subset N \ , \tag{1}$$

with baseline $b_{i,j}$ between images i and j and set J image pairs out of all source images N which have been marked valid. The average baseline is important for our filtering and quality propagation to model 3D uncertainties.

2.2 Error Propagation

In the subsequent filtering, completion and fusion steps, we use the uncertainty of a 3D point as error metric [36]. We derive this uncertainty in the three space dimension from 3D Point $P = (P_x, P_y, P_z)$, focal length f, camera baseline b and expected disparity error with standard deviation Δp for a pair of registered images:

$$\begin{aligned} \Delta P_x &= \Delta p \frac{P_z}{fb} \sqrt{(b - P_x)^2 + P_x^2} \ , \\ \Delta P_y &= \Delta p \frac{P_z}{fb} \sqrt{2P_y^2 + \frac{b^2}{2}}, \\ \Delta P_z &= \Delta p \frac{P_z^2}{fb} \sqrt{2} \ . \end{aligned} \tag{2}$$

Kuhn et al. [27] already successfully applied this concept for MVS and showed that using $\Delta P = \Delta P_z$ as scalar error is valid because it is the dominant error when depth values are larger then two times the camera baseline. To extend the propagation to MVS instead of single stereo only, they are estimating for the average baseline over all images. In this paper we make use of the pixel-wise estimated baseline $b = b_i^p$ (Eq.(1)) from the PM process (see Sec. 2.1).

2.3 Depth Completion

The recently published depth completion method TAPA-MVS [39] has demonstrated that completion on superpixels improves the quality of 3D reconstruction significantly [44]. TAPA-MVS employs SEED superpixels [49] in two varying sizes on the input images. After the first PM iteration they filter out small peaks in the depth images and fit planes in superpixels by applying RANSAC on remaining depth measurements. In the second PM run, the completed maps are taken into account as hypothesis for pixel-wise depth estimation. To handle untextured areas, a texture confidence from the local variance of an image is used [39]. We take this method as a reference and also make use of RANSAC-based plane fitting on superpixels. In contrast to TAPA-MVS, we propose a hierarchical clustering of extracted superpixels, integrate the MVS geometric error (Sec. 2.1) in the RANSAC optimization and apply the depth completion as a post-processing step for PM instead of adding an additional hypothesis. This improves the PM runtime, allows better handling of large untextured areas and the processing of generic MVS datasets. In addition, fine structured details are preserved as we keep the original PM depth values when available. We will demonstrate the qualitative improvement on the *ETH3D* benchmark. The quantitative improvement is demonstrated by processing the *Tanks and Temples* dataset, which do not provide an absolute scale, e.g., in meters like the *ETH3D* datasets.
DeepC-MVS [30] completion strategy is based on confidence prediction used for guiding a regularizer within a global optimization framework [40]. To this end, a confidence prediction network is used [48] which is extended for learning MVS-derived depth maps. The confidences are used for filtering of outliers employing probabilistic clustering and as an input for a joint depth and normal map refinement [40]. The latter makes use of the confidence as a factor in the part of a cost function describing the regularization term. We instead make use of the confidence for filtering depth maps before applying our plane based depth completion.

2.4 Plane Detection

In perceiving 3D Scenes, a global knowledge of repetitive structures such as planes and smooth surfaces can be of advantage in providing an accurate and complete 3D model. Methods based on Manhattan World Stereo (MWS) are used long-since, e.g., utilizing MRF-based optimization [12] or plane-fitting [32]. The latter employs detected planes to create a set of aligned boxes for approximating the geometry of the scene. Moreover, Satoshi et al. [19] introduce scene structure awareness in extracting planar surfaces for detecting small

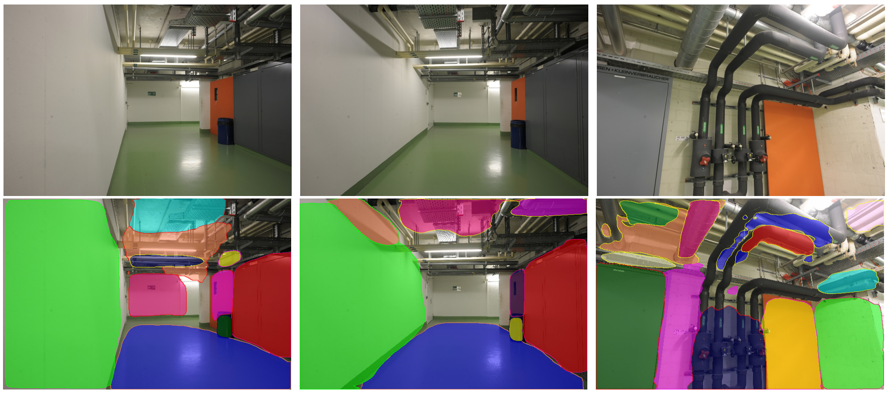

Fig. 1 Piece-wise planar reconstruction results by PlaneRCNN on pipes dataset from *ETH3D* high resolution. Top: Input images from *ETH3D* dataset pipes, bottom: Segmented plane regions.

planar regions and fine structures. The scene is segmented based on geometric representation and structural elements. Recently, deep-learning-based methods are used for a less biased reconstruction. PlaneRCNN proposes a novel end-to-end architecture that jointly infers plane parameters and the corresponding plane mask from a single image. PlaneRCNN's architecture consists of three components. (1) Detection Network that uses a variant of MaskR-CNN to build instance masks over inferred surface normals. The plane offset is subsequently calculated using the inferred normal and an additionally estimated monocular depth map. (2) PlaneRCNN uses a U-network to refine all detected masks together against the ground truth employing a cross entropy loss. Similar to PlaneNet, a simple fitting algorithm was used to extract planes and generate plane ground truth annotations from 3D Scans of the ScanNet datatset [8]. (3) The network enforces the consistency during the training between the nearby views in minimizing the projection and re-projection distance norm. In general, PlaneRCNN outperforms all learning-based plane segmentation networks and traditional MWS-based methods in terms of plane recall and segmentation quality. Although PlaneRCNN gives a better accuracy for plane parameters compared to learning based techniques, the performance is low compared in comparison to MWS-based methods which are more accurate in fitting the plane parameters from depth maps.

In our semantic-aware 3D Reconstruction pipeline, we use the detected plane masks only as an additional input to improve the completeness of the corresponding depth maps. Applying PlaneRCNN on images from the *ETH3D* dataset Fig. 1 achieves a reasonable performance in terms of segmentation recall where most dominant planar surfaces in the scene are detected. Additionally, as shown in Fig. 1, the network achieves a partial segmentation consistency between the views as seen for the left and right side wall segmentations and a robustness against the illumination variations and light reflections. Unlike the SEED [49] segmentation, the network generates a weak plane segmentation as an over-segmentation or under-segmentation mask that requires a refinement process before the depth completion process. To solve this problem, we propose a plane assignment method based on plane semantic information (high level features) and superpixels (low level features).

3 Algorithm

We propose a 3D reconstruction pipeline with three major steps: 1. Depth maps generation and filtering, 2. Completion of depth maps and 3. Final filtering of outliers. In this section, we describe the individual steps.

3.1 Depth Maps Generation and Filtering

For the reconstruction of areas which have been captured by only two cameras, one cannot rely on a filtering with robust statistics as multiple measurements would be needed. To decimate the number of outliers, e.g., TAPA-MVS filters peaks in depth maps considering the depth difference of neighboring pixels. This is not possible for general MVS configurations because a constant depth range has to be defined. Alternatively, filtering in the disparity domain as proposed by Hirschmüller [16] and used in SGM can be utilized: Small peaks are clustered in the disparity map and filtered if they do not exceed a minimum cluster size. The disparity map is segmented by allowing neighboring disparities to vary by only less than one pixel. At this point SGM operates on disparity maps estimated from two images. Such clustering is suitable for SGM with its fronto parallel assumption. However, such depth maps show problems on strongly slanted surface parts. Our employed PM in contrast, allows their reconstruction and neighboring pixels could be connected even though the disparity difference is high. Hence, in [28] we do not cluster the disparity maps directly but use its derivative. We transfer depth values d from the depth maps into disparity space considering the pixel-wise average baseline b_i^p from Eq. (1) and derive the first order deviation for pixel i as follows:

$$\nabla \mathcal{D}_i^p = \nabla \frac{f \ b_i^p}{d_i^p} \tag{3}$$

From the derivated disparity map $\nabla \mathcal{D}$ small clusters of connected components are filtered out. Because the derivative does not penalize depth values on parallel planes we additionally employ the 3D error term (Eq. 2) in the clustering, which means that neighboring values should be in a tolerantly-set noise area ΔP. Our experience shows that the filtering is working well when employing the average baseline, even though Eq. (3) considers standard stereo configurations and we use MVS-derived depth maps in varying configurations.

In this paper, we also evaluate the use of the clustering in the normal map domain as recently proposed in [30]. Instead of depths or disparities, the normal vectors are clustered which generally have a unit scale. If the normal vector of neighboring pixels is similar the same cluster is assigned to them. This is rendered possible even on noisy normal maps when taking a confidence into account [30]. To this end, inlier probabilities for the clusters are estimated by means of the Binary Bayes fusion:

$$p_I = 1 - \frac{1}{1 + e^{l_I}}, \; l_I = \sum_{i \in I} \frac{p_i}{1.0 - p_i} \; , \tag{4}$$

where cluster probability p_I is estimated by fusing pixel probabilities p_i from confidences. Note that for this clustering option we also make use of the 3D error (Eq. 2) as described above. Fig. 2 shows a direct comparison of the proposed filtering methods. Neighboring depth values on the wall have varying differences in depth (≈ 0.5m in near and ≈ 5m in far areas). Clustering in the disparity or normal vector space can handle such scale differences.

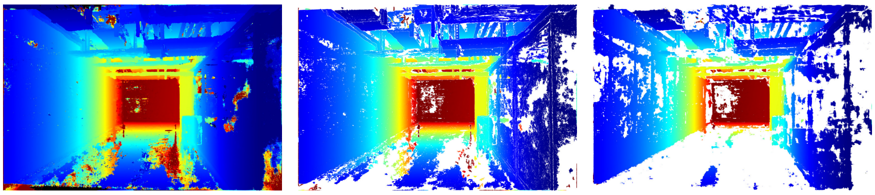

Fig. 2 From left to right: input PM depth map, filtered depth map based on disparity clustering, filtered depth map based on normal clustering.

3.2 Depth Completion

The locally derived depth maps lack in completeness as PM disregards textureless areas. Inspired by recently published depth filling on superpixels TAPA-MVS [39] we propose an important extension of the filling of depth maps. TAPA-MVS extracts superpixels on two levels and considers the filled depths as additional hypothesis in PM. We, in contrast, estimate the superpixels on the finer resolution and cluster them subsequently. To this end, we set a minimum number of valid depth measurements per superpixel. If the number is below a threshold the superpixel is merged with the most similar neighboring superpixel (see Fig. 5). As proposed by TAPA-MVS, the Bhattacharya distance of RGB histograms is used as similarity metric. TAPA-MVS considers this metric for selecting measurements of neighboring superpixels for the RANSAC fitting depending on a similarity norm. In their method only one neighboring superpixel can be taken into account, hence, our method is more adaptive for larger untextured areas.

Having a sufficient number of depth measurements, we run the plane fitting employing a RANSAC optimization. More precisely we use RANSAC employing an M-estimator (MSAC). In TAPA-MVS the RANSAC considers inliers to

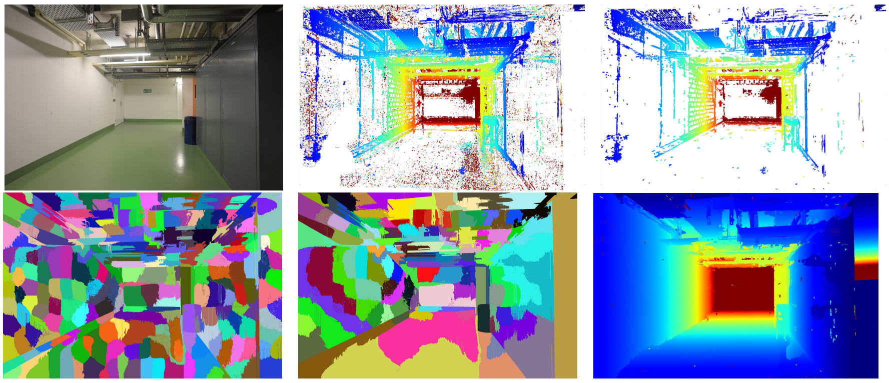

Fig. 3 Filtered (top right) and unfiltered (top centre) PM depth maps. The bottom left images shows the initial and the bottom centre the clustered superpixels. On the bottom right the depth map is filledfrom the RANSAC-derived planes.

be in a fixed range of 10 cm. At this point, we make use of the pixel-wise 3D error estimate. The inlier range is set relatively to the expected error defined in Eq. (2) and the summed cost of the relative distance is minimized:

$$\text{argmin}_K \sum_{k \in K} \frac{|| < P_k, \mathcal{P} > ||}{\Delta P_k} \ , \tag{5}$$

with set K as selected inlier points which minimize the sum of relative distances considering the distance $|| < P_k, \mathcal{P} > ||$ from point P to plane $\mathcal{P}$ and uncertainty ΔP (Eq. (2)) as normalization. If the distance from 3D point P and plane $\mathcal{P}$ is above a threshold its influence is fixed as proposed by MSAC. The normalization concerning the 3D uncertainty allows the processing of generic data where no scale is know, e.g., on the *Tanks and Temples* datasets. In addition, considering the 3D error is beneficial when having varying baseline and distances to the scene (see Fig. 3). Having a plane estimate for each superpixel, we fill the initial PM depth map with depth values by intersecting the line of sight with the estimated 3D plane. Next to runtime optimization the post-processing allows the preservation on fine-structured details from the original depth maps. PM also extracts normal vectors, which are required for our fusion method. Because our depth filtering method is implemented as postprocessing, we fill the PM normal maps by the individual normals of the superpixel-wise estimated plane.

3.3 Plane-based Depth Completion

As an extension to the depth filling based on superpixels by PCF-MVS [28] (Sec. 3.2), we propose a semanticly-aware depth completion method based on

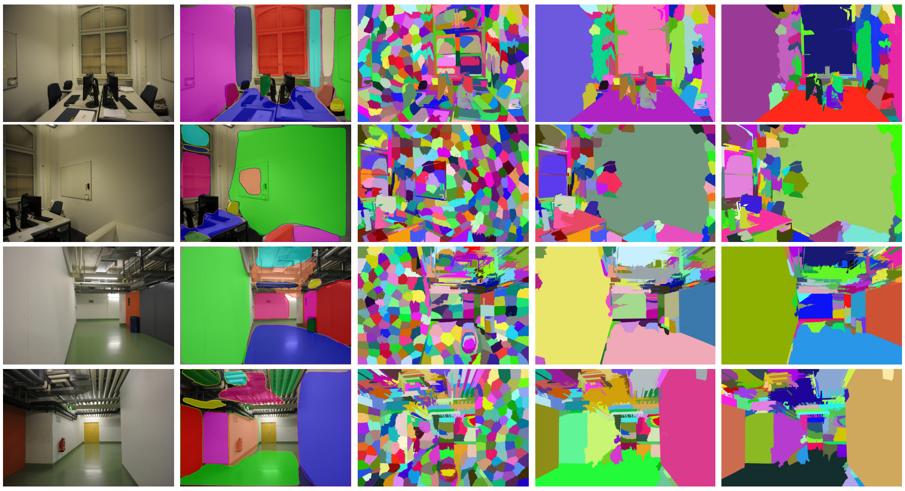

Fig. 4 Plane segmentation refinement on pipes dataset from *ETH3D* high resolution. from Left to Right: (a) input RGB image, (b) segmented plane regions, (c) superpixels produced by SEED, (d) semantic planarity labeling, (e) final plane segmentation after plane growing.

planarity detection and robust fitting aiming to complete the filtered depth maps even further. Relying on low-level features (appearance) in assigning fine superpixels to the same planarity surfaces is not sufficiently robust. More precisely, the neighbouring similarity metric fails to reconstruct large surfaces with wide variability in surface appearance. For that, we use an additional planar semantic segmentation as high level features to extend the definition of planar surfaces.

As a pre-processing step to our plane semantic-aware depth completion, we propose a local boundary refinement strategy based on plane segments and superpixels. For each image in the scene, we use PlaneRCNN [33] to extract a set of planes $P = (p_1, ..., p_m)$, where m denotes the number of extracted plane masks per image. Further, we use SEED [49] to segment the image into a set of n superpixels $S = (s_1, ..., s_n)$, where $n >> m$ (see Fig. 4 (b-c)). Our algorithm consists of two main steps: 1) Semantic majority voting and 2) Appearance based plane growing.

Semantic majority voting: Due to the fact that the plane mask may exceed the real plane boundaries and include partially other planar or non-planar regions, we define a semantic cost $C_{s_i}^{p_j}$ for each superpixel s_i as an overlapping ratio between pixels which lie in plane p_j and the total pixel count $|s_i|$. The semantic cost C is in the range of $0 \leq C_i^{\,j} \leq 1$. We additionally define a cost for non-planarity with $C_{s_i}^{\neg p} = 1 - \sum_{k \in P} C_{s_i}^{p_k}$ and perform a majority voting with the highest semantic cost to assign superpixel s_i to plane p_j, where $\text{argmax}_{j \in P}\, C_j$, or to the background as non-planar region.

This normalized semantic cost also represents the likelihood of an individual superpixel belonging to a plane. We define then a threshold parameter τ_s to assign superpixels to the most probable plane mask if the maximum semantic

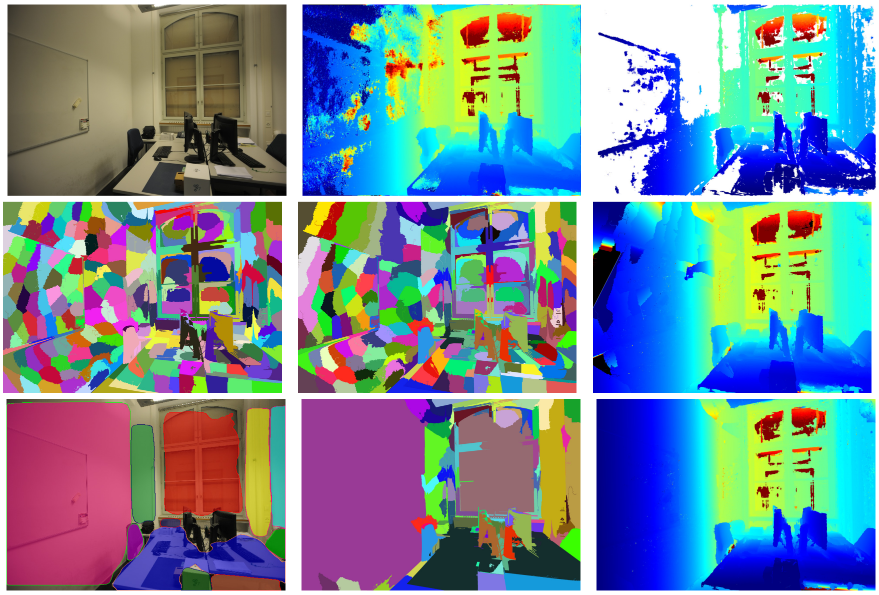

Fig. 5 Filtered (top right) and unfiltered (top centre) PM depth maps. The middle left images shows the initial SEED and the middle centre the PCF clustered superpixels, the middle right the completed depth using PCF. The bottom left images shows the initial plane masks by PlaneRCNN and the bottom centre the refined plane segments. On the bottom right the depth map is filled from the RANSAC-derived planes.

cost exceeds $C_{max} > \tau_s$. At this step, the successfully labeled superpixels are subsequently merged constructing a new plane segmentation set P^*. Superpixels which are shared between several planes with low scoring are not assigned to any plane class at this point, even though they are part of a plane.

Appearance based plane growing: To solve the under-segmentation problem we use the appearance similarity growing mechanism using Bhattacharya distance of RGB histograms [39] on the new plane segmentation set P^* to merge the weakly supported superpixels with its most similar neighbouring one, pushing the plane segments to grow to its boundaries. Non-assigned similar superpixels in textureless areas are also merged forming larger planar clusters extending the network detection sets P_+^* with more planar clusters and reducing the non assigned superpixels set S_-.

Additionally to the refinement objective, semantic majority voting groups the assigned superpixels in weakly supported textureless region into one plane (see Fig 4 (d)) where the appearance growing extends the support of initial plane segments with bordering superpixels which has a sufficient depth measurement and a good recall to scene object boundaries (see Fig 4 (e)). The thresholding parameter τ_s discourages the assignment of wrong detections and uncertain planar candidates providing a good initial segmentation to grow safely.

Having a final segmentation for each view $\Psi_{I_N} = \{P_+^* \cup S_-\}_{I_N}$, we ensure a sufficient support to each plane segment in estimating robustly the plane parameters. With a sufficient number of depth measurements, we estimate the plane parameters for each plane segment $\mathcal{P} \in \Psi$ using MSAC and we fill the filtered depth and normal map as described in section 3.2

The network shows shortcomings when detecting sky area as planar surface in outdoor scenes; and since the sky superpixels are weakly supported by non-valid depths values; the growing process merges the sky segment with other valid plane segments (e.g., road, building roof) seeking for a sufficiently enough plane support. To solve this issue, we use a sky detector described in section 3.4 to remove the sky part from the processed plane keeping only the valid part from the plane mask to assign superpixels afterwards.

3.4 Sky Filtering

Filtering of depth maps does not guarantee the removal of outliers on surfaces with low degree of texturedness. In particular, sky areas for outdoor scenes lead to strong artifacts that cannot be filtered geometrically. To solve for this problem, we introduce a sky area detection by means of semantic segmentation (see Fig. 6) based on DeepLabV3+ [4] which demonstrated a stable semantic segmentation on images. In order to retrain the network for binary segmentation, an enhanced dataset for sky is used comprising the following datasets: Cityscapes [7],

ADE20K [56] and SkyFinder [6]. They all provide the class sky in a various of outdoor scene which we use for a binary labeling. The datasets Cityscape and ADE20 maintain a large variety of outdoor configurations while SkyFinder strengthens the stability of the retrained network for different illuminance and

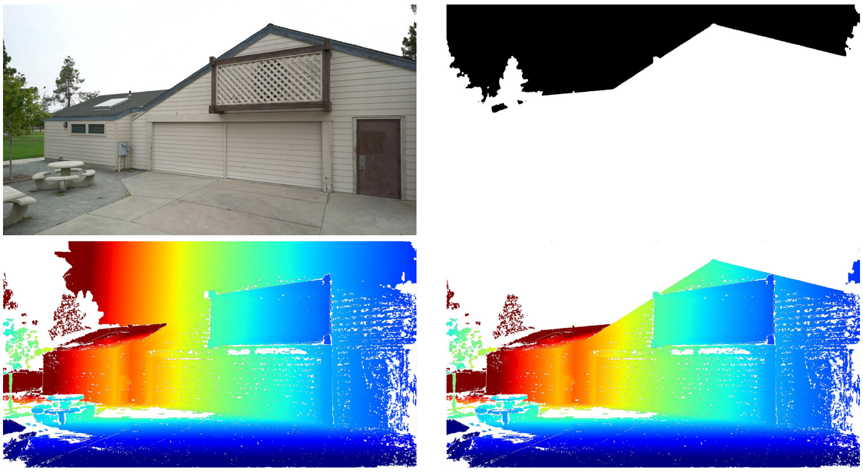

Fig. 6 The upper row shows an input image with the binary sky map where black marked pixels represent detected sky areas. The mask is used to filter the depth map (bottom left). The bottom right image shows final depth map.

weather conditions. This helps to improve the ability of the network to distinguish sky and non-sky parts of each image. In order to avoid overfitting we augmented the data by cropping, rotating and flipping images. Furthermore, DeepLabV3+ is modified for binary segmentation by adjusting the loss function. We penalize wrong labeling of sky areas by a factor of 10 because sky areas appear less frequently in the training dataset.

4 Experiments

We run experiments using the *ETH3D* and *Tanks and Temples* training datasets to evaluate the proposed steps of the pipeline and validated the full pipeline on the test datasets. The 13 *ETH3D* training datasets contain from 14 to 76 registered images with 24MP resolution while the low-res training datasets contain 660 to 1200 registered images with 0.36 MP resolution. We processed the input PM depth maps on a 28-core 2.6GHz machine with four Geforce GTX 1080 Ti GPUs and 11GB memory. The depth completion was performed on a Core i7 8700k CPU and RTX 2080Ti GPU. The initial depth maps are generated using the MVS pipeline and parameter settings from [30]. The PM depth maps are derived from half resolution images for the high-resolution and in full resolution for the low-resolution images. The number of source images to match the reference image against is set automatically dependent on the GPU memory. Because PCF-MVS [28] employs PM depth maps from COLMAP and because it it is a very popular method, representing the state of the art in 3D reconstruction, we use it as a direct comparison for the individual evaluations. In addition TAPA-MVS is compared as the baseline for our depth completion method.

4.1 Depth Completion

In this experiment, we filter depths, by setting a minimum cluster size of connected pixels which neighboring disparity gradients (see Eq. (3)) do not exceed the threshold of 1.0 pixel. We empirically found that a minimum cluster size of 100 pixels is a good trade off preserving completeness and accuracy.

A major contribution of our paper is the depth completion on growing superpixels: We generate superpixels from the input images and combine neighboring superpixels when having an insufficient number of depth values. The minimum number of depth values per superpixel is set to 4000. Similar to TAPA-MVS we use SEED superpixels [49] with a parameter of 200 superpixel per image for the initialization. Note that in contrast to TAPA-MVS we make use of the superpixel-based completion as post-processing which avoids computational complex hypothesis propagation within the MVS process and allows higher quality in high resolution areas because the original PM depth map is preserved. For each clustered superpixel we run the RANSAC-based plane estimation allowing a maximum error of two times the standard deviation (see Eq. (2)). We set the minimum inlier rate to 30%, the maximum

number of trials to 10000 and the confidence to 99.99%.
To allow a direct comparison of the depth completion part, we employ our completion method on COLMAP PM depth maps similar to TAPA-MVS in this experiment. In addition, we use the depth map fusion method provided by COLMAP which was also used by TAPA-MVS. Similar to TAPA-MVS, we changed the standard fusion parameters to maximum re-projection error of 0.5 pixels for the high-res and 0.25 pixels for the low res images and the maximum difference of the normal angle of 20°. Furthermore, we disabled the sky filtering procedure at this point to allow a direct comparison of the depth completion part. Table 1 shows the results over all *ETH3D* training datasets and a direct comparison to TAPA-MVS and COLMAP which are based on the same PM method. The latter employs the PM depthmaps without completion. As evaluation metric the standard F-Score is used combining completeness and accuracy. We evaluated the metrics on two resolutions: 1 *cm* and 10 *cm* to cover fine and coarse resolution scores. Our approach has the best F-Score values for the fine and competitive values for the coarse resolution evaluation. Note that for the final evaluation (Sec. 4.5) we do not use the COLMAP fusion parameters suggested by TAPA-MVS. Table 2 shows the evaluation table for the *ETH3D* video datasets containing large sets of low resolution images. Again, our method has best scores concerning the F-score when evaluating fine-structured details (1 cm distance).

Method	AVG	courty.	delivery	electro	facade	kicker	meadow	office	pipes	playgr.	relief	relief2	terrace	terrain
Ours	**68.16** 93.34	**67.70** 96.56	**74.03 97.70**	**75.73** 95.82	**52.31 94.55**	69.00 88.33	**57.96 89.38**	**61.09** 86.18	**73.73 94.33**	**55.57** 93.70	**71.09** 91.51	**67.76** 91.54	**78.92** 97.63	**81.16 96.22**
TAPA	60.85 **93.69**	47.38 **96.89**	65.33 97.62	65.35 **96.15**	36.51 91.67	**75.16 94.94**	48.82 85.97	54.70 **87.72**	63.51 91.96	53.31 **94.40**	68.36 **93.76**	64.97 **93.06**	73.37 **98.30**	74.27 95.58
COLMAP	51.99 87.61	49.13 95.54	61.73 94.48	60.53 91.77	36.57 90.14	53.14 87.16	32.95 75.50	37.10 73.41	38.68 76.86	40.49 87.33	65.72 90.05	63.08 89.87	72.11 96.48	64.60 90.46

Table 1 F-Score [%] combining completeness and accuracy at a distance of 1cm and 10cm for all *ETH3D* high-res training datasets and their average mean (AVG). The first row show the result of our completion method with COLMAP fusion with same parameters as TAPA-MVS. The second and third row show the results of TAPA-MVS and COLMAP which are based on the same depthmaps. Best results are marked bold. Our method outperforms the the baseline methods especially in fine-structured areas.

4.2 Evaluation of disparity and normal based clustering

In the following, we show evaluation results for the two clustering techniques described in Section 3.1 for filtering our input depth maps. The input PM

Method	AVG	delivery	electro	forest	playground	terrains
Ours	**42.66 82.11**	**32.50 79.81**	**40.83** 80.72	**46.11 88.57**	**28.69 76.43**	65.17 85.02
TAPA	38.87 81.65	22.75 77.80	34.37 **82.80**	45.53 87.23	24.02 73.54	**67.70 86.89**
COLMAP	32.32 76.00	16.26 74.84	28.69 78.71	39.99 85.35	17.68 59.56	58.96 81.54

Table 2 F-Score [%] combining completeness and accuracy at a distance of 1cm and 10cm for all *ETH3D* low-res training datasets as in Table 1.

depth maps used in this evaluation are obtained from the pipeline used in [30]. For all further experiments performed with these depth maps on ETH3D high resolution data, we set the fusion parameters as follows: The maximum normal difference is set to 20°, the maximum re-projection error to 1.0 and the minimum amount of measurements required to full-fill this condition are 2. For *ETH3D* low resolution data the maximum normal difference is set to 10°, the re-projection error to 0.25 and the minimum amount of measurements to 4. For *Tanks and Temples* we use the same settings as for *ETH3D* low resolution except for a maximum normal difference of 20°. For the *Tanks and Temples* dataset we also enable the proposed 3D consistency check. This constraint allows a fusion of 3D points considering the uncertainty in 3D (see Eq. (2)). The original COLMAP fusion considers the inverse depth as a filtering criteria. We found that the improvement does not have a significant influence on the numerical benchmarking, but reduces the final point cloud size because a larger set of redundant low quality points is fused resulting in a clean point cloud. In Table 3, we provide an ablation study on the *ETH3D* high resolution training dataset.

Method	AVG	courty.	delivery	electro	facade	kicker	meadow	office	pipes	playgr.	relief	relief2	terrace	terrain
disp. clust.	73.72 **96.93**	69.25 **98.89**	78.67 98.39	77.61 **98.32**	**54.98 96.09**	79.66 **96.68**	62.80 **94.00**	67.44 **91.40**	**79.08 97.21**	**63.58 96.67**	78.11 97.83	77.13 97.75	**82.65 98.50**	87.39 98.30
norm. clust.	**74.30** 96.51	**69.47** 98.83	**80.05 98.46**	**78.11** 97.48	54.74 95.73	**82.15** 96.36	**63.00** 93.48	**69.98** 89.89	77.99 95.74	63.37 96.26	**78.48** 97.83	**77.59 97.79**	82.62 98.31	**88.36 98.44**
TAPA	60.85 93.69	47.38 96.89	65.33 97.62	65.35 96.15	36.51 91.67	75.16 94.94	48.82 85.97	54.70 87.72	63.51 91.96	53.31 94.40	68.36 93.76	64.97 93.06	73.37 98.30	74.27 95.58
COLMAP	51.99 87.61	49.13 95.54	61.73 94.48	60.53 91.77	36.57 90.14	53.14 87.16	32.95 75.50	37.10 73.41	38.68 76.86	40.49 87.33	65.72 90.05	63.08 89.87	72.11 96.48	64.60 90.46

Table 3 Ablation study for disparity based clustering and normal based clustering on the *ETH3D* high resolution datasets. Results are shown for 1cm tolerance (left) and 10cm tolerance (right) It can be seen that both the disparity and normal based clustering perform similar, however for a tolerance of 1cm the normal based clustering yields better results on average.

It can be observed that, on average, the clustering based on normals performs better on the fine grained tolerance level of 1cm, while the disparity based clustering performs better on the high tolerance level of 10cm. However, the quantitative results show no significant differences. In Figure 7, we show visualizations of the resulting point clouds for the different clustering methods. It can be seen that the clustering based on normals improves visual quality of the reconstruction as less outliers and noise are present in the point cloud.

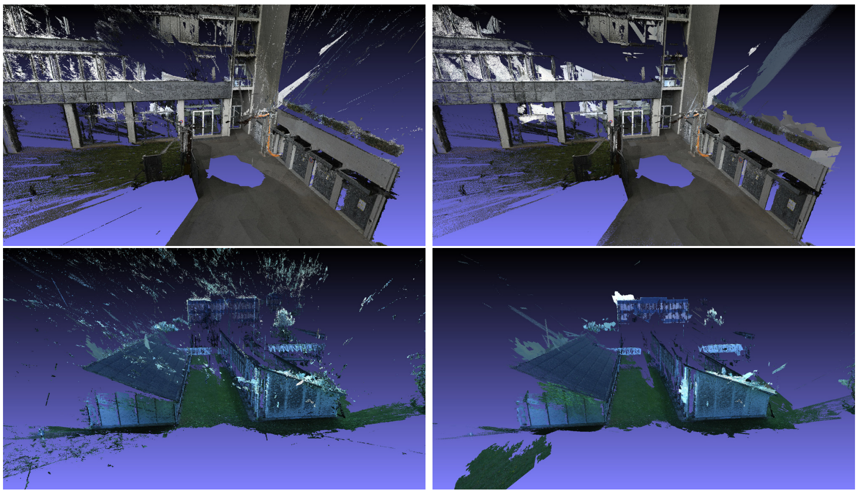

Fig. 7 Point cloud results for disparity based clustering (left) and normal based clustering (right) from the *ETH3D* training dataset. The first row shows the Electro dataset and the second row shows the Meadow dataset. The normal based clustering approach (right) significantly reduces the amount of outliers present.

4.3 Semantic Plane Completion

We also evaluate the performance impact of using the plane-based segmentation masks which are created from plane detections and superpixel segmentation masks. The ablation study for measuring this impact was conducted on the high resolution and low resolution *ETH3D* training datasets. Table 4 and Table 5 show the quantitative results on 1cm and 10cm tolerance. In Table 4, one can see that the average result on 1cm is slightly improved when using the segmentation masks based on plane detections. A more significant improvement can be observed when comparing the scores of datasets Office, Pipes and Kicker. These are datasets where a lot of planar structures and surfaces with poor texturing are present and as such our plane based segmentation mask augmentations have a high performance impact on these datasets. Figure 8 shows these improvements qualitatively on the reconstructed point clouds of

Office and Kicker. It is observable that planar structures such as the floor, ceiling and walls are more complete when using the plane-based segmentation masks. Furthermore, an improvement in terms of the average F1 score can be observed for the low resolution training datasets on both the 1cm and 10cm tolerance level.

Method	AVG	courty.	delivery	electro	facade
norm clust	74.30 **96.51**	**69.47 98.83**	**80.05 98.46**	**78.11** 97.48	**54.74 95.73**
norm clust + planes	**74.73** 96.47	68.02 97.41	79.09 98.28	77.92 **97.55**	53.86 95.06
TAPA	60.85 93.69	47.38 96.89	65.33 97.62	65.35 96.15	36.51 91.67
COLMAP	51.99 87.61	49.13 95.54	61.73 94.48	60.53 91.77	36.57 90.14

Method	kicker	meadow	office	pipes	playgr.
norm clust	82.15 96.36	**63.0 93.48**	69.98 89.89	77.99 95.74	**63.37 96.26**
norm clust + planes	**85.00 97.00**	59.01 90.52	**78.61 93.65**	**81.23 97.07**	62.46 95.90
TAPA	75.16 94.94	48.82 85.97	54.70 87.72	63.51 91.96	53.31 94.40
COLMAP	53.14 87.16	32.95 75.50	37.10 73.41	38.68 76.86	40.49 87.33

Method	relief	relief2	terrace	terrain	
norm clust	78.48 **97.83**	**77.59 97.79**	**82.62 98.31**	88.36 98.44	
norm clust + plane	**78.51** 97.76	77.36 97.42	81.87 97.92	**88.61 98.56**	
TAPA	68.36 93.76	64.97 93.06	73.37 98.30	74.27 95.58	
COLMAP	65.72 90.05	63.08 89.87	72.11 96.48	64.60 90.46	

Table 4 Ablation study for the impact of plane-based segmentations on the *ETH3D* high resolution training dataset. The quantitative results are provided with 1cm tolerance (left) and 10cm tolerance (right). It can be seen that there are significant improvements in the F1 score for Office, Pipes and Kicker.

Method	AVG	delivery	electro
norm. clust	45.41 84.75	30.82 83.38	43.99 85.42
norm. clust + planes	**46.25 85.17**	**31.46 83.97**	**44.91 86.24**
TAPA	38.87 81.65	22.75 77.80	34.37 82.80
COLMAP	32.32 76.00	16.26 74.84	28.69 78.71

Method	forest	playground	terrains
norm. clust	**52.34 87.69**	**30.78 78.60**	69.14 88.64
norm. clust + planes	52.14 87.50	30.50 78.25	**72.26 89.87**
TAPA	45.53 87.23	24.02 73.54	67.70 86.89
COLMAP	39.99 85.35	17.68 59.56	58.96 81.54

Table 5 Ablation study for the impact of plane-based segmentations on the *ETH3D* low resolution training dataset. The quantitative results are provided with 1cm tolerance (left) and 10cm tolerance (right). The plane-based segmentations yield an increase in the average F1 score both on 1cm and 10cm tolerance.

4.4 Depth Filtering

Sky areas do not have a strong influence on the *ETH3D* benchmarking as they are mostly assigned to undefined areas. We found that for the *Tanks and*

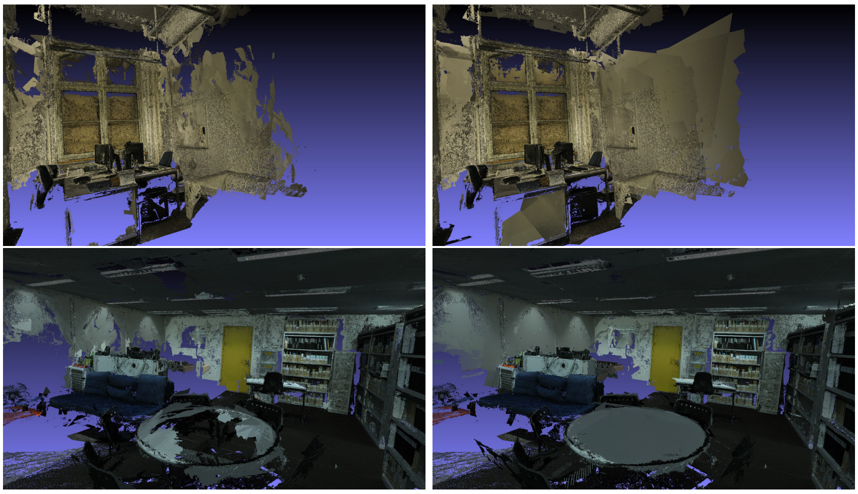

Fig. 8 Point cloud results comparing the use of standard superpixel-based segmentations (left) with the use of plane-based segmentations (right). The two datasets are Office (top) and Kicker (bottom) from the *ETH3D* high resolution training set. It can be seen that the method utilizing the plane-based segmentations is able to complete more of the walls in Office. Furthermore the table is more complete for the Kicker dataset.

Temples benchmark, in contrast, artifacts appear around surfaces which have an influence on the final quality. This can be traced back to the large amount of images showing the same scene leading to a higher rate of outliers. In addition, the evaluation is calculated from cropped parts of the scene without undefined areas. As described in Sec. 3.4, we retrained the segmentation network DeepLabV3+ for a binary labeling of sky areas. After testing with two backbones and training steps the best performance was obtained by retraining 30000 steps with Xception [5] as backbone. The mean intersection over union (mIOU) of our sky segmentation network scores 91% accuracy for the validation dataset. In general, the sky segmentation network has the robustness against complex weather and illuminance conditions and generates stable and precise segmentation result.

We generate binary maps for each image and filter sky-labeled depth values in the final depth maps before the fusion. Even though, for some areas false positive appear, the overall improvement is obvious. Fig. 9 visually shows the resulting point cloud with and without sky-labeled filtering. For a numerical evaluation we run the *Tanks and Temples* training datasets on both point clouds and show the results in Table 6. The point clouds were generated using the fusion settings from our clustering evaluation described in Section 4.2. For a fair comparison we use the sky filtering in all datasets for the final evaluation, even for the *ETH3D* and *Tanks and Temples* indoor data which slightly lowers the accuracy.

Method	AVG	Barn	Caterp.	Church
DeepPCF-MVS	**64.98**	**72.45**	**62.02**	**62.23**
DeepPCF-MVS $\setminus\{SF\}$	64.92	71.94	61.73	62.22
COLMAP	53.03	47.26	54.71	52.37

Method	Courth.	Ignatius	Meetingr.	Truck
DeepPCF-MVS	49.98	89.67	44.15	74.35
DeepPCF-MVS $\setminus\{SF\}$	**50.06**	89.67	**44.26**	**74.59**
COLMAP	38.37	78.06	34.45	64.98

Table 6 F-Score [%] combining completeness and accuracy at for all *Tanks and Temples* training datasets. For the evaluation of Ours $\setminus\{SF\}$ the sky filtering was disabled.

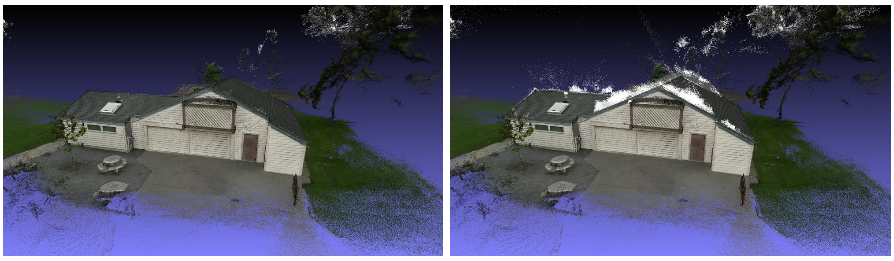

Fig. 9 The images show the rendered 3D point cloud from the Barn dataset with and without filtering of sky elements. The outliers are obviously decimated.

4.5 Final Evaluation

We run our pipeline also on the *ETH3D* and *Tanks and Temples* test datasets. The 12 *ETH3D* test datasets contain from 7 to 110 registered images captured with a 24MP camera while the *Tanks and Temples* datasets contain larger sequences containing from 251 to 1105 images of 12MP. For this evaluation, we use the parameter settings described in Section 4.2 and additionally enable sky filtering. Concerning the *ETH3D* high-res dataset (Table 7) our method outperforms TAPA-MVS, PCF-MVS and ACMM for 2cm which is the standard evaluation on the *ETH3D* homepage. Table 8 shows the results for the *ETH3D* video datasets, where an improvement in terms of the average F1-score and completeness compared to the other methods is achieved as well. These improvements are the results of a more dense initial prediction of the MVS pipeline from [30] compared to COLMAP, as [30] use a hierarchical multi-scale approach [52] which yields better results for surfaces with poor texturing. A more robust outlier filtering strategy based on normals which also takes confidence metrics into account [30] helps to remove additional outliers for the depth completion input depth maps and thus also yields more reliable points during plane fitting. Furthermore, the plane-based segmentations help to provide the MSAC based plane fitting with more point measurements on large planar surfaces. The datasets in the *Tanks and Temples* benchmark do not have many flat walls which reduces the influence of our plane prior. In addition, the evaluation strongly depends on the SfM results as no ground truth camera poses is given. To demonstrate the influence, we registered the datasets with COLMAP SfM and Altizure SfM [2] for PCF-MVS. In the case of DeepPCF-MVS we use the SfM results from Altizure SfM [2] as they improved the results for PCF-MVS. Note that for the experiment on the *Tanks and Temples* test set, we also increase the allowed re-projection error to 1.0, as this yields more dense results for difficult scenes. Furthermore the minimum amount of measurements is set to 3. We also filter the completed depth maps such that the re-projection error needs to be less or equal to 2.0 for at least 3 source images. Table 9 shows the quantitative results of DeepPCF-MVS on the *Tanks and Temples* dataset. In terms of the average F-score metric on the advanced and intermediate datasets, the results from DeepPCF-MVS are slightly worse compared to PCF-MVS, however there are no significant differences. For some datasets such as Family and Playground DeepPCF-MVS outperforms PCF-MVS.

Method	AVG(train)	AVG(test)	courty.	delivery	electro	facade	kicker	meadow	office
DeepPCF-MVS	**85.30 83.08**	**88.10 89.72**	**88.77**	**90.48**	85.46	**73.08**	**90.97**	**73.77**	**84.81**
PCF-MVS	79.42 75.73	80.38 79.29	84.88	88.17	**86.08**	69.85	75.23	68.43	68.03
ACMM	78.86 70.42	80.78 74.34	86.89	83.40	86.02	70.50	75.28	71.49	63.01
TAPA-MVS	77.69 71.45	79.15 74.49	80.68	84.52	81.36	63.14	84.77	64.82	68.72
LTVRE	61.82 49.41	76.25 66.27	72.83	77.19	64.37	58.97	33.56	28.00	52.59
COLMAP	67.66 55.13	73.01 62.98	80.49	77.98	75.29	62.95	63.62	49.96	47.32

Method	pipes	playgr.	relief	relief2	terrace	terrain	botani	boulde.	bridge
DeepPCF-MVS	**85.95**	**77.43**	**88.22**	**87.73**	87.85	**94.34**	**92.24**	**69.03**	89.57
PCF-MVS	78.38	71.76	81.26	80.65	88.56	91.18	87.71	68.99	83.65
ACMM	69.26	73.57	84.11	83.98	**89.76**	87.84	89.31	68.37	**89.99**
TAPA-MVS	75.91	71.86	81.62	79.55	87.80	85.24	89.59	62.99	88.16
LTVRE	42.21	63.93	74.52	76.28	77.15	82.13	88.60	64.38	79.24
COLMAP	50.72	58.57	76.87	75.50	84.94	75.33	87.13	65.63	88.30

Method	door	exhibi.	lectur	living.	lounge	observ.	old co.	statue.	terrace.
DeepPCF-MVS	**93.39**	**77.53**	**89.79**	**94.12**	**78.98**	**96.58**	**85.66**	**95.72**	**94.58**
PCF-MVS	91.46	63.00	77.77	90.28	66.10	95.09	61.40	88.22	90.94
ACMM	91.60	70.28	77.25	89.66	53.37	93.53	74.24	82.85	88.85
TAPA-MVS	91.51	65.77	77.14	91.09	60.91	93.21	50.26	87.05	92.17
LTVRE	89.12	70.76	69.79	87.86	49.09	93.20	56.21	80.16	86.65
COLMAP	84.19	62.96	63.80	87.69	38.04	92.56	46.66	74.91	84.24

Table 7 F-Score [%] at a distance of 2cm which is the standard setting for the *ETH3D* high-res benchmarking. For the AVG also the completeness is listed (right) The individual rows show the results of the currently leading methods. Our extended method DeepPCF-MVS achieves both a better F1 score and completeness metric compared to the other methods.

Method	AVG(train)	indoor	outdoor	AVG(test)	indoor	outdoor
DeepPCF-MVS	**62.76 60.97**	**66.22 62.83**	**60.45 59.74**	**63.41 60.64**	**53.95** 53.51	**69.71** 65.40
PCF-MVS	57.32 58.17	59.66 57.60	55.76 58.56	57.06 58.42	48.10 **54.11**	63.03 61.29
TAPA-MVS	55.13 55.77	58.21 61.18	53.07 52.17	58.67 58.89	52.34 51.21	62.89 64.01
ACMM	55.12 57.01	54.88 55.57	55.28 57.97	55.01 58.27	43.19 46.31	62.89 **66.24**
ACMH	51.50 53.77	53.46 49.88	50.20 56.37	47.97 52.68	38.24 35.79	54.45 63.93
LTVRE	53.52 41.68	58.21 44.05	51.36 40.11	53.52 43.60	45.46 37.31	58.89 47.80
COLMAP	49.91 40.86	51.76 40.09	48.68 41.37	52.32 45.89	42.45 37.03	58.89 51.79

Table 8 F-Score (left) and completeness (right) [%] at a distance of 2cm which is the standard setting for *ETH3D* low-res (video) benchmarking. DeepPCF-MVS outperforms the other methods on the test set as well as the training set.

Method	AVG(int.)	AVG(ad.)	Family	Francis	Horse	Lighth.	M60	Panther
Altiz.+DeepPCF-MVS	56.33	34.80	**72.88**	48.52	39.19	62.99	**60.09**	58.60
Altiz.+PCF-MVS [28]	55.88	**35.69**	70.99	49.60	40.34	**63.44**	57.79	58.91
COLM.+PCF-MVS [28]	53.39	34.59	67.32	43.28	34.45	61.17	50.59	**61.20**
ACMM	**57.27**	34.02	69.24	**51.45**	**46.97**	63.20	55.07	57.64
ACMH	54.82	33.73	69.99	49.45	45.12	59.04	52.64	52.37
COLMAP	42.14	27.24	50.41	22.25	25.63	56.43	44.83	46.97

Method	Playgr.	Train	Auditor.	Ballr.	Courtr.	Museum	Palace	Temple
Altiz.+DeepPCF-MVS	57.50	50.85	25.88	38.35	34.99	47.97	24.93	36.64
Altiz.+PCF-MVS [28]	56.59	49.40	**28.33**	**38.64**	35.95	**48.36**	**26.17**	**36.69**
COLM.+PCF-MVS [28]	55.93	53.14	26.87	31.53	**44.70**	47.39	24.05	32.97
ACMM	**60.08**	**54.48**	23.41	32.91	41.17	48.13	23.87	34.60
ACMH	58.34	51.61	21.69	32.56	40.62	47.27	24.04	36.17
COLMAP	48.53	42.04	16.02	25.23	34.70	41.51	18.05	27.94

Table 9 F-Score [%] at a employing varying distances as defined by the evaluation software and the average mean (AVG) for the intermediate and advanced *Tanks and Temples* datasets. For a comparison the currently best methods are shown. Our method generates state-of-the-art results especially when using Altizure SfM.

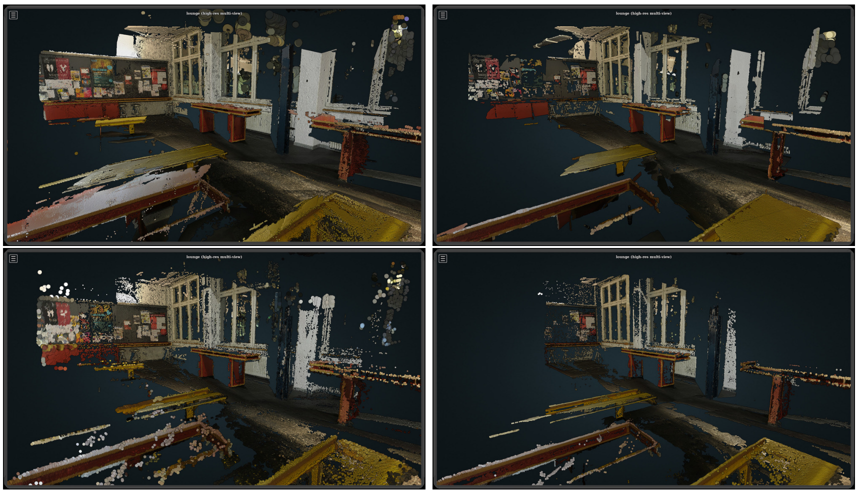

Fig. 10 Qualitative results from the lounge dataset from the high resolution test data. Top left: DeepPCF-MVS. Top right: PCF-MVS. Bottom left: TAPA-MVS. Bottom right: ACMM.

5 Conclusion and Outlook

In this paper we have presented a pipeline for dense reconstruction of 3D point clouds from large sets of high-resolution images. The extension of our previous depth completion method resulted in an improvement over the state of the art in 3D reconstruction concerning completeness and the preservation of fine details. Three major contributions are made: 1) depth completion on growing superpixels, 2) plane-based segmentation 3) filtering of sky areas. The individual steps are extended by considering MVS geometry. We have shown the improvement visually and numerically on datasets from standard benchmarks in large-scale reconstruction. A more robust MVS solution used for input depth

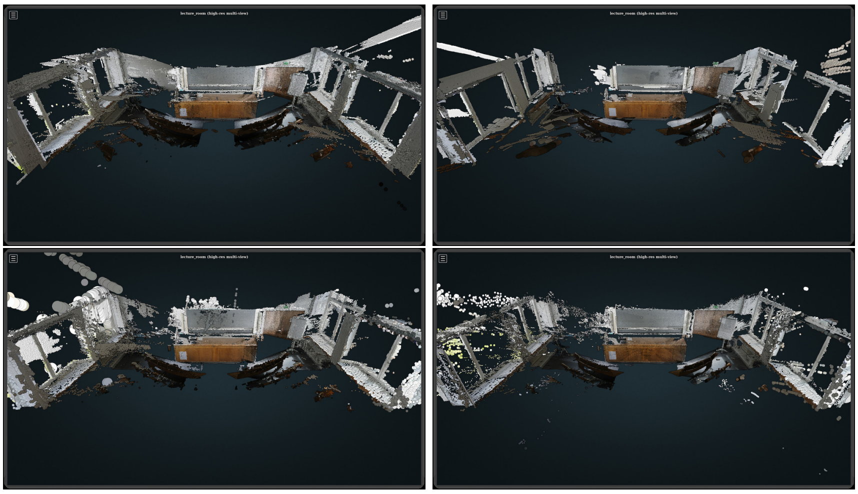
lecture_room (high-res multi-view)
lecture_room (high-res multi-view)
lecture_room (high-res multi-view)
lecture_room (high-res multi-view)

Fig. 11 Qualitative results from the lecture room dataset from the high resolution test data. Top left: DeepPCF-MVS. Top right: PCF-MVS. Bottom left: TAPA-MVS. Bottom right: ACMM.

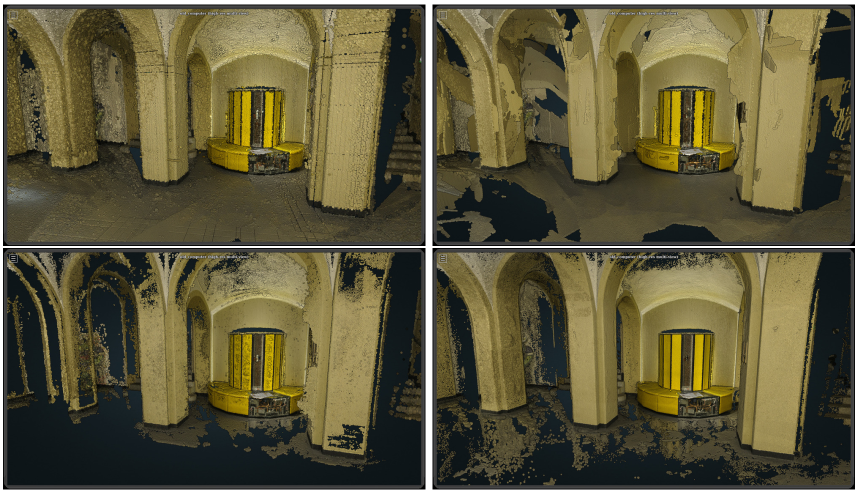
old_computer (high-res multi-view)
old_computer (high-res multi-view)
old_computer (high-res multi-view)
old_computer (high-res multi-view)

Fig. 12 Qualitative results from the old computer dataset from the high resolution test data. Top left: DeepPCF-MVS. Top right: PCF-MVS. Bottom left: TAPA-MVS. Bottom right: ACMM. It can be seen that DeepPCF-MVS improves the result compared to PCF-MVS as there are fewer outliers present.

maps [30] has also allowed a more robust plane fitting reducing the number of outlier planes which was deteriorating the accuracy on PCF-MVS. Figure 12 shows a direct comparison between PCF-MVS and DeepPCF-MVS on the old computer dataset, where fewer wrong plane fittings corrupting the result are visible for DeepPCF-MVS. Future work could include an extension which uses geometric constraints to further refine the initial plane detections.

References

1. Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. International Journal of Computer Vision (IJCV) **120**(2), 153–168 (2016)
2. Altizure: Altizure the portal for realistic 3d modeling (2019). URL https://www.altizure.com/
3. Bleyer, M., Rhemann, C., Rother, C.: PatchMatch stereo - stereo matching with slanted support windows. In: British Machine Vision Conference (BMVC) (2011)
4. Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
5. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: CVPR (2017)
6. Chu, W.T., Zheng, X.Y., Ding, D.S.: Camera as weather sensor: Estimating weather information from single images. Journal of Visual Communication and Image Representation **46**, 233–249 (2017)
7. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
8. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017)
9. Fuhrmann, S., Goesele, M.: Fusion of depth maps with multiple scales. In: ACM Transactions on Graphics (2011)
10. Fuhrmann, S., Goesele, M.: Floating scale surface reconstruction. ACM Transactions on Graphics **33**(4), 46:1–46:11 (2014)
11. Fuhrmann, S., Langguth, F., Goesele, M.: MVE: A multi-view reconstruction environment. In: Eurographics Workshop on Graphics and Cultural Heritage (2014)
12. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Manhattan-world stereo. In: CVPR (2009)
13. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multi-view stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **32**(8), 1362–1376 (2010)
14. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: ICCV (2015)
15. Goesele, M., Snavely, N., Curless, B., Hoppe, H., Seitz, S.M.: Multi-view stereo for community photo collections. In: ICCV (2017)
16. Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **30**(2), 328–341 (2008)
17. Hirschmüller, H., Scharstein, D.: Evaluation of stereo matching costs on images with radiometric differences. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **31**, 1582–1599 (2008)
18. Huang, P.H., Maten, K., Knop, J., Ahuja, N., Huang, J.B.: DeepMVS: Learning multi-view stereopis. In: CVPR (2018)
19. Ikehata, S., Yang, H., Furukawa, Y.: Structured indoor modeling. In: ICCV (2015)
20. Jancosek, M., Pajdla, T.: Hallucination-free multi-view stereo. In: ECCV (2010)
21. Jancosek, M., Pajdla, T.: Multi-view reconstruction preserving weakly-supported surfaces. In: CVPR (2011)
22. Kazhdan, M., Hoppe, H.: Screened Poisson surface reconstruction. pp. 29:1–29:13. ACM (2013)
23. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and Temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics **36**(4) (2017)
24. Kuhn, A., , Mayer, H.: Incremental division of very large point clouds for scalable 3d surface reconstruction. In: International Conference on Computer Vision Workshop (ICCVW) (2015)
25. Kuhn, A., Hirschmüller, H., Mayer, H.: Multi-resolution range data fusion for multi-view stereo reconstruction. In: GCPR (2013)

26. Kuhn, A., Hirschmüller, H., Scharstein, D., Mayer, H.: A TV prior for high-quality local multi-view stereo reconstruction. In: International Conference on 3D Vision (3DV) (2014)
27. Kuhn, A., Hirschmüller, H., Scharstein, D., Mayer, H.: A TV prior for high-quality scalable multi-view stereo reconstruction. International Journal of Computer Vision (IJCV) **124**(1), 2–17 (2017)
28. Kuhn, A., Lin, S., Erdler, O.: Plane completion and filtering for multi-view stereo reconstruction. In: GCPR (2019)
29. Kuhn, A., Roth, L., Frahm, J.M., Mayer, H.: Improvement of extrinsic parameters from a single stereo pair. In: IEEE Winter Conference on Application of Computer Vision (WACV) (2018)
30. Kuhn, A., Sormann, C., Rossi, M., Erdler, O., Fraundorfer, F.: Deepc-mvs: Deep confidence prediction for multi-view stereo reconstruction. In: arXiv:1912.00439 [cs.CV] (2019)
31. Labatut, P., Pons, J., Keriven, R.: Robust and efficient surface reconstruction from range data. In: Computer Graphics Forum (2009)
32. Li, M., Wonka, P., Nan, L.: Manhattan-world urban reconstruction from point clouds. In: ECCV (2016)
33. Liu, C., Kim, K., Gu, J., Furukawa, Y., Kautz, J.: Planercnn: 3d plane detection and reconstruction from a single image. In: CVPR (2019)
34. Liu, C., Yang, J., Ceylan, D., Yumer, E., Furukawa, Y.: Planenet: Piece-wise planar reconstruction from a single rgb image. In: CVPR (2018)
35. Mayer, H., Bartelsen, J., Hirschmüller, H., Kuhn, A.: Dense 3d reconstruction from wide baseline image sets. In: International Conference on Theoretical Foundations of Computer Vision: Outdoor and Large-scale Real-world Scene Analysis (2012)
36. Molton, N., Brady, M.: Practical structure and motion from stereo when motion is unconstrained. International Journal of Computer Vision (IJCV) **39**(1), 5–23 (2000)
37. Mostegel, C., Prettenthaler, R., Fraundorfer, F., Bischof, H.: Scalable surface reconstruction from point clouds with extreme scale and density diversity. In: CVPR (2017)
38. Muecke, P., Klowsky, R., Goesele, M.: Surface reconstruction from multi-resolution sample points. In: Proceedings of Vision, Modeling, and Visualization (VMV 2011) (2011)
39. Romanoni, A., Matteucci, M.: TAPA-MVS: Textureless-aware PatchMatch multi-view stereo (2017)
40. Rossi, M., Gheche, M.E., Kuhn, A., Frossard, P.: Joint graph-based depth refinement and normal estimation. In: arXiv:1912.01306 [cs.CV] (2019)
41. Schönberger, J., Fraundorfer, F., Frahm, J.: Structure-from-motion for mav image sequence analysis with photogrammetric applications. The international archives of photogrammetry, remote sensing and spatial information sciences **40**(3), 305–312 (2014)
42. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
43. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: ECCV (2016)
44. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: CVPR (2017)
45. Seitz, S., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: CVPR (2006)
46. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: Exploring photo collections in 3d. ACM Trans. Graph. **25**(3), 835–846 (2006)
47. Strecha, C., von Hansen, W., Gool, L.J.V., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: CVPR (2008)
48. Tosi, F., Poggi, M., Benincasa, A., Mattoccia, S.: Beyond local reasoning for stereo confidence estimation with deep learning. In: ECCV (2018)
49. Van den Bergh, M., Boix, X., Roig, G., Van Gool, L.: SEEDS: superpixels extracted via energy-driven sampling. International Journal of Computer Vision (IJCV) **111**(3), 298–314 (2015)
50. Vu, H.H., Labatut, P., Pons, J.P., Keriven, R.: High accuracy and visibility-consistent dense multiview stereo. IEEE Trans. Pattern Anal. Mach. Intell. **34**(5), 889–901 (2012)
51. Wu, C.: Towards linear-time incremental structure from motion. In: International Conference on 3D Vision (2013)

52. Xu, Q., Tao, W.: Multi-scale geometric consistency guided multi-view stereo. In: CVPR (2019)
53. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: MVSNet: Depth inference for unstructured multi-view stereo. In: ECCV (2018)
54. Yu, Z., Zheng, J., Lian, D., Zhou, Z., Gao, S.: Single-image piece-wise planar 3d reconstruction via associative embedding. In: CVPR (2019)
55. Zheng, E., Dunn, E., Jojic, V., Frahm, J.: PatchMatch based joint view selection and depthmap estimation. In: CVPR (2014)
56. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20K dataset. In: CVPR (2017)