DeepC-MVS: Deep Confidence Prediction for Multi-View Stereo Reconstruction

Andreas Kuhn\(^1\), Christian Sormann\(^2\), Mattia Rossi\(^3\), Oliver Erdler\(^1\), Friedrich Fraundorfer\(^2\)

\(^1\) Sony Europe B.V., \(^2\) Graz University of Technology, \(^3\) École Polytechnique Fédérale de Lausanne

Abstract

Deep Neural Networks (DNNs) have the potential to improve the quality of image-based 3D reconstructions. A challenge which still remains is to utilize the potential of DNNs to improve 3D reconstructions from high-resolution image datasets as available by the ETH3D benchmark. In this paper, we propose a way to employ DNNs in the image domain to gain a significant quality improvement of geometric image based 3D reconstruction. This is achieved by utilizing confidence prediction networks which have been adapted to the Multi-View Stereo (MVS) case and are trained on automatically generated ground truth established by geometric error propagation. In addition to a semi-dense real-world ground truth dataset for training the DNN, we present a synthetic dataset to enlarge the training dataset. We demonstrate the utility of the confidence predictions for two essential steps within a 3D reconstruction pipeline: Firstly, to be used for outlier clustering and filtering and secondly to be used within a depth refinement step. The presented 3D reconstruction pipeline DeepC-MVS makes use of deep learning for an essential part in MVS from high-resolution images and the experimental evaluation on popular benchmarks demonstrates the achieved state-of-the-art quality in 3D reconstruction.

1. Introduction

Multi-View Stereo (MVS) methods render possible the reconstruction of a 3D scene from multiple images for which the camera calibration and the relative camera poses are known. The major challenge is the reconstruction of 3D point clouds as complete as possible while minimizing the number of outliers and maintaining a high accuracy of the 3D points. One way of minimizing outliers is the use of filtering methods which preserve the most accurate measurements and which remove unreliable measurements which can result in rather sparse point clouds. Applications with the goal of photo-realistic 3D rendering have a need for complete 3D models however.

For improving the completeness of reconstructed 3D scenes, regularization techniques are useful. Many conventional MVS methods perform regularization by using a 3D cost volume of matching costs [15, 35, 10]. One drawback of this technique is the large size of the cost volume and subsequently the increased optimization complexity, which depends polynomially on the image resolution limiting their applicability on high resolution images [17]. The 3D cost volumes can also be used as an input to a DNN [12, 39], to optimize the volume employing a learned regularization measure. However, the high memory requirements of DNNs for these approaches makes them rather unsuitable for processing high-resolution image data.

For efficient estimation of depth maps, the Patch Match (PM) method [7] has demonstrated high quality results without the need of handling a global cost volume. PM instead utilizes stochastic search over the depth space. Many efficient implementations exist, which further optimize the efficiency by proposing parallelization schemes [8, 42] and sophisticated view selection schemes [30]. Due to its local optimization, however, PM depth maps lack in completeness, which need further processing. In this paper, we extend multi-scale estimation for PM [38] and show how to filter and optimize noisy depth maps.
2. Related Work

In this section we give an overview of papers that are relevant to our proposed method. Firstly, we review MVS methods. Secondly, confidence prediction methods which mainly focus on disparity maps classification from stereo images are summarized. Finally, depth refinement methods are reviewed as they can naturally make use of our confidence prediction technique.

**Multi-View Stereo:** In a similar approach to classical two-view stereo methods, which build up a cost volume by matching image patches along the epipolar line, plane-sweep based Multi-View Stereo methods construct a cost volume by computing costs for a set of given plane hypothesis [9]. Instead of the number of disparities, the depth of the volume is defined by the number of planes. As this leads to a significant consumption of computational resources, Bleyer et al. [7] make use of the PatchMatch algorithm [5], which tries to reduce the amount of computed matching costs by propagating depth hypothesis across the image. This concept has also been implemented for the MVS case [4]. Zheng et al. [42] make use of a probabilistic scheme for view selection in PatchMatch MVS [4], which is improved upon by Schnberger et al. [30]. In order to increase the completeness of the depth estimates Romanoni and Matteucci [27] introduce a method which propagates depth estimates from local planes estimated from superpixels. This approach is extended upon by Kuhn et al. [17], who propose a region growing for the superpixels and additional outlier filtering strategies. A black-red checkerboard sampling scheme was utilized by Galliani et al. [8] in order to decrease the runtime of PatchMatch based Multi-View Stereo, which was further improved upon by the multi-scale approach of Xu and Tao [38].

In recent years, neural network based approaches to MVS [12, 40], working with cost volumes, have also been established. Yao et al. [41] have extended their approach in order to be able to process higher resolution imagery. However, as described in [41], this method is not able to process resolutions as high as the ones found in the high-resolution version of the ETH3D benchmark [31]. This is why we make use of a PatchMatch [5] based approach for our MVS pipeline, which has the advantage of not having to process a large cost volume for high resolution imagery. Furthermore, we utilize the multi-scale approach of [38], to increase the robustness of the method when processing images with large amounts of non-textured regions.

**Confidence Prediction:** Prediction of confidences is an inherent part of MVS methods. It can be calculated as costs by local patch comparison based on metrics like Normalized Cross Correlation. Detailed analyses of the influence from local matching costs are given by Hirschmüller and Scharstein [11] as well as by Hu and Mordohai [2]. Moreover, the confidence can be derived from a globally optimized cost volume within the cost aggregation process [32]. The latter allows the consideration of global cost terms like the overall smoothness of a disparity map [15, 35, 10].

It has been shown that machine learning methods can improve the quality of confidence prediction, e.g., by employing hand-crafted features as input for a random forest classifier [26, 2, 19, 22]. The use of automatically learned features for confidence prediction was firstly proposed by Seki et al. [1] while Poggi et al. [20] presented the first end-to-end confidence prediction giving the raw disparity map as input to a DNN. Further improvements are possible by exploiting local consistencies [18], adding the image as additional input to the network [36] or even using extended information from the entire cost volume as input data [13].

Due to scalability issues we avoid processing on global cost volumes and focus on methods that can be applied in the image domain. Unfortunately, all proposed 2D methods work on disparity maps from standard stereo configurations only. In this paper we present the first confidence prediction network for generic MVS-derived depth maps.

**Depth Refinement:** MVS methods tend to fall short in untextured areas, where the matching becomes ambiguous. Therefore, most 3D reconstruction pipelines include a refinement step meant to remove depth outliers or even estimate missing depth areas. Most of these methods rely on a confidence map, as it is critical to understand which depth map areas are reliable and which to extend.

Local refinement methods, such as Tosi et al. [37], binarize the confidence map in order to classify pixels as reliable; then the depth at non reliable pixels is inpainted using an interpolation heuristic relying exclusively on the neighboring reliable pixels. Unfortunately, local approaches tend to fall short in the presence of large unreliable areas.

Global methods, instead, compute the refined depth map as the minimizer of a global cost function. Some methods, e.g., the Fast Bilateral Solver [6], promote a smooth refined depth map while making sure that this is close to the input depth in those areas recognized as reliable by the confidence map. Although smoothing is avoided across color edges, where depth discontinuities could take place, simple smoothness is a weak regularization. In fact, other methods rely on stronger assumptions. Park et al. [24] assume a piece-wise planar world: a set of candidate planes is estimated a priori and a Markov Random Field is used to assign each pixel to a plane, from which the depth can be derived. However, the a priori selection of the possible planes, first, fixes the complexity of the scene in front of the camera and, second, assumes that the selected planes are correct. To overcome the previous limitations, in [28] the authors propose to adopt a regularizer which promotes depth maps fulfilling a piece-wise planar world assumption, but without any a priori candidate plane selection. Also in this method, the confidence plays a fundamental role, as the method im-
plicitly fits planes based on the reliable depth areas.

Recently, some authors addressed the depth refinement problem within the deep learning framework [34, 25]. However, the performances of these methods tend to decrease significantly when applied to data even slightly different from that used in the training phase. Therefore, in this article we adopt the methods in [28] and show that, coupled with our proposed confidence, it can improve the 3D reconstruction both qualitatively and quantitatively.

3. Algorithm

In this section we give a detailed explanation of our proposed MVS algorithm based on PatchMatch [5]. The generated depth maps are subsequently used as an input to our confidence prediction network. For the latter we propose extended input and a new synthetically-derived dataset.

3.1. Multi-View Stereo

The proposed MVS pipeline makes use of the PatchMatch [5] algorithm with a red-black checkerboard propagation scheme, as proposed in [8]. A visualization of the used sampling pattern is shown in Figure 2. We have removed samples close to the central pixel, due to the intuition that these hypothesis can be replaced by a perturbed estimation which can be utilized in addition to samples from the pattern [30]. The depth and normal estimates are perturbed according to the scheme presented in [30], where the parameter $\epsilon$ is calculated as follows $\epsilon = \frac{1}{2^i}$, analogous to [30]. The variable $i$ denotes the current red-black iteration. We also test hypothesis which result from combinations of the current, perturbed and random depth estimates with their respective normal vectors, as suggested by [30, 38]. However, as opposed to [38], we compute the matching costs and test these hypothesis during the main red-black iterations and do not perform these calculations in a separate step. We are drawing from six potential samples in each direction yielding a total of 24, as we have found that sampling from a lot of hypothesis increases the chance of using incorrect estimates. This is because we follow the view selection scheme proposed by Xu and Tao [38], which selects the eight best candidates from the pattern based on their respective matching costs. This means that outliers are potentially selected due to ambiguities in the photometric matching cost metric we are using, which is the bilateral weighted normalized cross correlation proposed in [30]. In contrast to the scheme described in [38], the number of candidates which are considered when updating the current estimate is increased to 16. It should be noted that the view-selection matrix is still based on the eight candidates with the lowest matching costs [38]. Furthermore, we are incorporating a plane-based depth propagation strategy, as proposed in [30]. This means that instead of directly using the depth estimate at a given sampling location as a new hypothesis, one makes use of the local plane defined by the depth and normal estimate at this location. By means of intersecting the viewing ray of the destination pixel with the local plane defined at the sampling location, we are able to propagate rapid changes in depth values more effectively. We also make use of the multi-scale estimation, geometric consistency and detail restorer proposed by [38] on three hierarchy levels with a downsampling factor of 0.5. For the view selection and detail restorer, we use the parameter settings from [38]. Example depth maps are shown in Fig. 3.

3.2. Confidence Prediction

We present the first confidence prediction method feasible of handling complex MVS scenarios using a DNN. While for disparity maps from stereo configurations datasets with ground truth disparities are available [29, 21] for MVS only semi-dense lidar ground truth [31] and data captured in a laboratory environment with relatively low resolutions and nonvarying distance to the scene [32, 3] exist. Because training data in sufficiently large numbers is a requirement for training a DNN, we propose a new dataset generated from synthetic 3D scenes. We use the AirSim framework [33] for the simulation of a drone flight to capture multiple images with ground truth depth maps and camera calibration from varying viewpoints. The dataset includes representative scenarios with large baselines, perspective deformations, varying distance to the scene, specular reflections and weakly textured surfaces. Examples from the dataset are shown in Fig. 3. Trees in the simulation framework can lead to erroneous ground truth depth maps because branches are simulated simplified as partially transparent planes. Hence, we masked out vegetation areas from the ground truth data. For the final learning of the network parameters, we combine publicly available semi-dense real world dataset [31] with our proposed synthetic dataset to alleviate the influence of the domain gap.

Our confidence prediction is inspired by ConfNet [36] which allows the prediction of confidences on disparity maps by employing derived global features. The ma-
Figure 3. Four examples from our synthetic dataset and two examples from the ETH3D training dataset [31]. From top to bottom: Input image, ground truth depth map, estimated depth map and ground truth confidence maps. Depth estimates which do not fit into the minimum/maximum range defined by the ground truth depth map are marked white. The confidence maps are estimated employing the geometrically estimated 3D error. We combine semi-dense real-world with dense synthetic ground truth to lighten the domain gap.

Major component of ConfNet is an encoder-decoder network which takes as input the RGB image and the disparity map derived from a stereo pair and outputs a pixel-wise confidence. In contrast to [36] we are solely using the encoder-decoder network ConfNet as we found that the originally proposed combination with a local confidence network does not improve the results in our application. The network was originally trained on the Middlebury [29] and Kitti [21] datasets. Because of a binary classification into valid and non-valid measurements the cross entropy loss function on binarized ground truth is used. For these datasets the ground truth can be created in a straightforward way from disparities by generating binary confidence maps by setting a maximum disparity error of, e.g., one pixel.

We, in contrast, work on depth maps derived from multiple viewpoints as it is typical for MVS scenarios. Basically, confidence prediction is about assigning a probability value for each pixel specifying if the corresponding depth value is either an outlier or an inlier. Positive confidences would support a depth value to be arranged in a reasonable noise range. The constant setting of a fixed noise level, e.g., in meters is not possible, because of varying baselines, focal lengths and distances to the scene. Hence, we make use of a pixel-wise estimation of the expected noise in the 3D space. Therefore, we use a well-known model derived from analytical error propagation [23] which was already used in the 3D fusion of depth maps [16]. The simplified expected 3D standard deviation \( \sigma_P \) as scalar value is described as:

\[
\sigma_P = \sigma_p d^2 fb \sqrt{2},
\]

with distance to the scene \( d \), baseline \( b \), focal length \( f \) and pixel uncertainty \( \sigma_p \). Note that the error model is based on the assumption of classical stereo cameras. In MVS configurations we would have multiple baselines, hence we use the average baseline from all cameras for \( b \) following [17]. Having a scalar depth range, we can perform a binary classification of the depth maps generating ground truth for valid confidences (see Figs 3 and 4).

The input of ConfNet is a four channel tensor consisting of the RGB image and the disparity map. In our case we have depth maps, which are generally not normalized as the scale of a scene is not necessarily given. In addition, the unconsidered varying noise levels in MVS-derived depth maps can be problematic for the network. Therefore, the depth map input is replaced by two normal channels created by the estimation of spherical coordinates of a normal vector. An example input normal map is shown in Fig. 5.

As already mentioned, we cannot use an entire cost volume from MVS as input for the network because of mem-
ory issues. Nonetheless we give additional information for each pixel to the network derived from MVS. As done by our baseline MVS methods, we perform consistency checks employing the geometric and photometric error of depth measurements. Instead of using a fixed threshold for the minimum number of successfully matched images, we keep this information for every pixel and generate a counter map as additional network input (see Fig. 5).

![Figure 5](image1.png)

Figure 5. Beside the RGB image (Fig. 3), we enter the network with normal map (upper right) and a counter map (lower left). The counter map illustrates from how many source images the depth has been verified color coded in dark blue (0 matches) to red (5 matches). On the lower right the output of our network is shown color coding low confidences (blue) to high confidences (red).

4. Applications

In the following we describe how the proposed learned confidence maps can be integrated in a 3D reconstruction pipeline to improve the quality of depth maps.

4.1. Filtering

MVS generates dense depth maps including outliers due to occlusion and wrongly matched patches. It is common practise that individual depths are filtered pixel-by-pixel within a post-processing step. State-of-the-art MVS methods make use of geometric filtering by firstly estimating the re-projection error when projecting the depth from the reference image into overlapping source images. In addition, photometric consistency is checked [30]. If there is an insufficient number of source images that fulfill the geometric and photometric requirements the depth is filtered. We, in contrast, use that information as additional input for our confidence prediction network. Subsequently, the depth is filtered depending on the optimized confidence additionally considering RGB image and normal map.

An advanced filtering can be useful to decimate clusters of outliers. To this end, small islands in depth image [27] or disparity space [10, 17] are removed if they have an insufficiently large size. In contrast to existing methods, we cluster normals instead of disparities or depths to overcome the problem of an unknown scale. In addition, we use probabilistic fusion to estimate for an overall confidence of a finally clustered island. More precisely, we cluster neighboring pixels incrementally, if its normal vectors are similar with respect to a given threshold. At this point, we consider our confidences $m$ as an inlier probability $p$. The overall probability of an island $p_I$ is estimated by fusion of individual pixel probabilities $p_i$. For the fusion of probabilities for binary states the Binary Bayes Fusion is used:

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

(2)

4.2. Depth Refinement

Most human made environments either consist of piece-wise planar surfaces or can be well approximated as such. Moreover, in the pinhole camera model, the inverse depth map of a planar surface imaged by the camera results in a planar function. Formally, if the point $P_j \in \mathbb{R}^3$ belongs to the same plane of the point $P_i$ in the scene, with $i$ and $j \in \mathbb{R}^2$ the pixel coordinates where the two points are imaged, then the inverse depth $d$ at $j$ can be expressed as:

$$d_j = d_i + u_i^\top (j - i),$$

(3)

where $u_i \in \mathbb{R}^2$ is a vector defining the plane orientation at $i$. Based on these observations, the authors in [28] propose to refine a noisy and possibly incomplete depth map $\bar{z}$ by enforcing that the refined inverse depth map $d = 1/z$ is piece-wise planar. The method in [28] requires a depth map confidence: we adopt our proposed one.

The refinement of each inverse depth map $\bar{d} = 1/\bar{z}$ is cast into the minimization of a cost function comprising two terms. The first one is a simple data term, which penalizes those solutions deviating from the input inverse depth map $\bar{d}$ in those areas where this is considered as reliable by our confidence map $m$:

$$f(d) = \sum_i |d_i - \bar{d}_i| m_j.$$  

(4)

The second one is a regularization term, which promotes piece-wise planar solutions explicitly.

In particular, the regularization term models the inverse depth map as a weighted graph, where the pixels are the nodes, and where the weight $w_{ij} \in \mathbb{R}_{>0}$ of the edge between the pixels $i$ and $j$ encodes the probability that $P_i$ and $P_j$ belong to the same plane in the scene. The higher $w_{ij}$, the more the regularizer will enforce the fulfillment of the constraint in Eq. (3). The regularizer reads as follows:

$$g(d, u) = \sum_i \sqrt{\sum_{j \sim i} w_{ij}^2 (d_j - d_i - u_i^\top (j - i))^2} + \alpha \sum_i \sum_{j \sim i} w_{ij} \|u_j - u_i\|_2,$$

(5)

(6)
where \( j \sim i \) is the set of pixels connected to the pixel \( i \) in the graph. The weight \( w_{ij} \) is computed based on the color difference between two pixels and on their Euclidean distance. In particular, the weight \( w_{ij} \) is computed for all the pixels \( j \) in a searching window centered at \( i \), but only the \( K \in \mathbb{N} \) pixels \( j \) with the highest weights \( w_{ij} \) are selected. This permits to avoid edges between points that are not likely to correspond to the same plane in the scene, as these would just increase the graph size unnecessarily. We refer to [28] for more details on the regularizer in Eqs. (5)–(6) and on the graph construction.

The resulting optimization problem is the following:

\[
\arg\min_{d,u} f(d) + \lambda g(d,u),
\]

where \( \lambda \in \mathbb{R}_{>0} \) balances the two terms. Thanks to the direct relation between \( u \) and the normals of the planes in the scene, the solution of the optimization problem in Eq. (7) provides both the refined depth map and the corresponding normal map. An example is provided in Figure 6.

![Figure 6. Refined depth (left) and normal map (right). The noisy input depth and normal maps is shown in Fig. 5.](image)

5. Experiments

In this section, we report the evaluation and validation of our proposed method on confidence prediction and demonstrate the improvement by applying our method for outlier filtering and depth map refinement on popular MVS datasets. We trained the confidence prediction network on the entire high-res dataset and tested on the ETH3D low-res and test data [31] which were not included in the training procedure. All network configurations are trained with the same parameters for 800 epochs with a learning rate of 0.0001 and ADAM [14] as the optimizer to minimize the Binary Cross Entropy on our ground truth label maps.

5.1. MVS

We evaluate the performance of our vanilla MVS pipeline (prior to outlier filtering and depth map optimization) by benchmarking on the high-resolution and low-resolution training datasets of the ETH3D benchmark [31]. This benchmark provides ground truth laser scan point clouds. These are used to evaluate a reconstructed point cloud based on an F-score metric which is computed as the harmonic mean of a completeness and accuracy term [31]. These terms represent the coverage of a given reconstruction with respect to the ground truth and the accuracy of the point locations measured against the laser scan point cloud. The reconstructed point cloud is generated by utilizing the depth fusion algorithm implemented in COLMAP [30] using the parameter settings from [38]. For our experiments, we perform 8 red-black iterations of the algorithm and downsample the input images to half of the original resolution. In Table 1 we first show the baseline performance of the pipeline, which is denoted as ours. The results are compared with both ACMM [38] and COLMAP [30], as our algorithm is based on features proposed in both pipelines. It can be observed that our pipeline achieves a better F1-score than both ACMM [38] and COLMAP [30]. The improvement is based on an increase in the completeness score and competitive results in terms of accuracy. As seen in Table 1, the plane based depth propagation is the main contributor to this increase, as highly varying depth estimates along surfaces can now be propagated more effectively. However, the increased completeness also results in a lower F1-score on the low-res datasets which contain a lot of images, thus inaccuracies have a bigger impact. We have also evaluated the performance of the sampling pattern compared to the one used by ACMM [38]. Table 1 shows that the sampling pattern used in our pipeline leads to an improvement in terms of the F1-score as well, due to an increase in accuracy.

| Method          | F1  | compl | acc. | F1  | compl | acc. |
|-----------------|-----|-------|------|-----|-------|------|
| COLMAP [30]     | 67.66 | 55.13 | 91.85 | 49.91 | 40.86 | 67.58 |
| ACMM [38]       | 78.86 | 70.42 | 90.67 | 55.12 | 57.01 | 54.69 |
| ours            | 83.62 | 83.25 | 84.17 | 55.04 | 62.55 | 50.66 |
| ours with ACMM [38] pattern | 82.77 | 82.90 | 82.85 | 53.66 | 62.72 | 47.30 |
| ours without plane prob. | 81.23 | 78.96 | 83.97 | 56.18 | 61.74 | 52.51 |

Table 1. Ablation study for different configurations of the MVS pipeline on the ETH3D [31] high-res (left) and low-res training set (right). The results represent the average F1-score, completeness and accuracy [%] over the training set at a distance of 2 cm. Best scores are marked bold. The scores for ACMM [38] and COLMAP [30] were obtained from the ETH3D benchmark [31] website. It can be seen that our pipeline outperforms the other two methods as a result of a significant increase in the completeness.

5.2. Confidence Prediction

Confidence prediction aims a separation between correct and incorrect measurements. As described in Sec. 3.2, we generate pixel-wise ground truth labels for each depth map by assigning depth measurements to an expected range of the 3D standard deviation. We found that the amount of positive confidences is higher then the amount of negative confidences in the training dataset. Hence, we re-weight the influence by scaling the loss of negative values by factor 3 in the Binary Cross Entropy loss function. For expected uncertainty (Eq.(1)) we use three times the standard deviation \( \sigma_p \) assuming a pixel uncertainty of \( \sigma_p = 1.0 \).
The effectiveness of confidence prediction is generally measured using the receiver operating characteristic (ROC) curve as metric. The area under the ROC curve (AUC) gives a scalar value representing the separability of two functions and, hence, the quality of the predicted confidences. As already mentioned, our method is the first able to process depth maps from Multi-View Stereo configurations and cannot be compared to existing work directly as they cannot handle depth maps from complex configurations. Only when employing the estimated 3D error, which is one of our contributions, the network is able to learn for complex depth errors. Another contribution of our work is proposal of a novel multi-channel data input comprising the RGB image, normal vectors and counter information from MVS (see Sec. 3.2). For an evaluation, we firstly analyse the performance of our method with varying input data. Firstly, we use solely the depth maps as input data of our network which is comparable to [20] which employs disparity maps. Secondly, we use the RGB and depth maps which is comparable to [36]. Thirdly we add the counter map and finally replace the depth by normal maps from our method. Note, that for depth map input, we perturbed the depth by a random scale, because on generic scenes no scale is given. Our method gives an AUC of 98.21, while the baseline approaches have lower scores; depth: 92.99, RGB+depth: 95.44, RGB+depth+counter: 97.58. In general, the use of the normal map allows the prediction of confidences for generic scenes as no scale is learned implicitly. The counter map information from MVS makes the network robust against complex configurations.

### 5.3. Filtering

In this experiment we optimized the 3D fusion parameters by allowing a maximum normal difference of 20° and a maximum re-projection error of 1.0 pixel. Additionally the minimum number of consistent points in 3D is set to 4 and the minimum angle is set to half for the dense sequences (low-res ETH3D) as they provide a higher redundancy.

Dense depth maps from MVS suffer from outliers due to occlusion or wrongly matched patches. Therefore, depth maps are generally filtered in a post-processing step. State-of-the-art methods project the depth values of the reference image to the employed source images and estimates for the geometric re-projection error and a photo-consistency measure [30, 38]. If a sufficient number of supporting source images is found the depth value is kept valid otherwise it is filtered. In the following, we call it mask filtering. In practice, a number of 1 (mask1) or 2 (mask2) images is used for datasets with limited number of images as provided by the ETH3D datasets [31]. We generate a counter map including the pixel-wise count as additional input to the network which takes into account additional information as the degree of texturedness and the level of noise implicitly. We use a confidence threshold of 0.1 for filtering all depths below that confidence threshold. Sophisticated methods employ the filtering of clustered islands (peaks) with small sizes [17, 27]. We filter islands with taking into account the clustered confidence threshold (Eq.(2)) of 0.3 and compare it to filtering clusters with sizes smaller 150 (clust150) which we empirically found to be the best value. Table 2 shows a direct comparison of the competitive methods.

| Method     | F1  | compl | acc. | F1  | compl | acc  |
|------------|-----|-------|------|-----|-------|------|
| Ours       | 84.80 | 81.63 | 88.51 | 61.35 | 63.14 | 60.17 |
| +conf      | 84.56 | 79.87 | 90.32 | 61.48 | 59.83 | 63.80 |
| +mask1     | 84.09 | 79.77 | 89.38 | 61.35 | 60.15 | 63.16 |
| +mask2     | 79.57 | 71.78 | 90.88 | 61.31 | 60.03 | 63.19 |
| +clust conf| 84.28 | 79.66 | 89.97 | 61.39 | 59.52 | 63.98 |
| +clust150  | 82.01 | 75.43 | 90.77 | 60.80 | 53.83 | 70.67 |
| +conf+clust| 84.16 | 78.87 | 90.83 | 61.48 | 59.45 | 64.26 |

Table 2. Scores as described in Table 1. Our confidence-based filtering (+conf) gives better F1-scores when comparing to mask filtering [30, 38]. Using the clustered confidences (+clust conf) also improves the score compared to clustering small islands [17, 27] of 150 depths (+clust150). The pure MVS baseline without filtering is denoted as (ours). Best of each category is underlined.

Filtering improves the accuracy of the point cloud, and hence, is minimizing the number of outliers in the 3D point cloud. Even though mask2 filtering has a higher accuracy the F-score falls drastically. Our confidence-based filtering allows a better regulation of the parameter. The confidence-based clustering allows a stable filtering of islands and gets better scores then number-based island filtering. In comparison to our method without filtering, the F-Score slightly decreases and combining confidence and clustered confidence is numerically best only for the F1 scores on video data. On the other hand, visually the point cloud is obviously improved comparing the results in Fig. 7. Note, that the unfiltered point cloud has best F-Score as outliers behind the walls are not considered in the evaluation.

Figure 7. Point clouds from ETH3D office sequence. Upper left: ours - unfiltered, bottom Left: ours+mask1 filtered [30, 38], Upper right: ours+clust150 filtered [17, 27], Bottom right: ours+deep-confidence-based filtered (prob+clust). The unfiltered point cloud gives the best F-Score but needs improved outlier filtering.

### 5.4. Optimization

Besides for filtering of outliers, confidences can be used for a refinement of depth maps.
As described in Sec. 4.2 we are using our confidence map as input of an optimization framework, which strongly regularizes depth and normals on areas marked as low confidence while preserving fine details on high confidence areas. We use the same parameters as proposed by the authors in [28]. The finally optimized depth and normal maps are fused with the same method as the filtered maps besides that the minimum normal consistency was set to 5 because the refined normal maps are of a higher accuracy (see Fig. 6). Table 3 shows the results for ETH3D high and low-resolution training datasets.

Table 3. Scores as described in Table 1. Our confidence-based refinement allows an improved reconstruction for many sequences.

| Method       | train F1 | compl | acc. | test F1 | comp | acc. |
|--------------|----------|-------|------|---------|------|------|
| Ours         | 84.80    | 81.63 | 88.51| 61.35   | 63.14| 60.17|
| Ours opt     | 85.86    | 82.70 | 89.65| 61.45   | 58.44| 65.33|

The numerical improvements results especially from closed holes in depth maps. As demonstrated in Fig. 8 the optimization is also suitable for refinement of thin structures which does not influence the evaluation score significantly.

5.5. Final Evaluation

As a final experiment, we test our proposed method with subsequent filtering and refining on the ETH3D test [31] datasets. We use the same parameters for all datasets for the final evaluation as described in Sec. 5.3 and 5.4. As our refinement method generates complete depth maps, geometric consistency can lack in the filtering of outliers in sky areas. Hence, we enable a sky filtering method [17] to minimize visual artifacts. The sky filtering does influence the score on ETH3D marginally (<0.01% on the high-res F-Score) data but gives visually better results.

Our proposed refinement method is effective but computational complex while the confidence-based filtering allows an efficient processing. Hence, we call the refinement method DeepC-MVS and the filtering DeepC-MVS fast. Table 4 shows a direct numerical comparison against the leading methods on the ETH3D high-resolution and low-resolution datasets [31]. Both of our methods outperform the current state of the art. In comparison to ACMM which is considered as baseline method the improvement on the ETH3D high-resolution data [31] is > 6%. By visually comparing the refinement against the filtering (Fig. 1 and 8) improvements are obvious on flat surface parts. Numerically they perform almost similar because there is currently no extended filtering enabled after the refinement. To this end, our confidence prediction network could be re-trained on the refined depth maps in the future. Furthermore, thin structures are highly unrepresented in the datasets. When evaluating the F-Score on a fine resolution (1cm) the refinement method gives better result 78.02 (refined) vs 77.40.

Table 4. F-Score [%] as described in Table 1. The individual rows show the results of the currently leading methods on ETH3D.

6. Summary

We have presented the first confidence prediction method of depth maps derived from challenging Multi-View Stereo (MVS) configurations. To allow training on dense depth maps, we propose a new dataset with dense ground truth generated from synthetic scenes. Binarized ground truth labels are estimated from ground truth and estimated depth maps employing an analytical error propagation. The finally predicted confidence maps are used for the probabilistic filtering of outliers as well as for the refinement of depth maps based on global optimization. Our method improves the state of the art in terms of qualitative and quantitative results when employing our proposed Patch-Match-based Multi-View Stereo method, as demonstrated on popular 3D reconstruction benchmarks.
References

[1] Seki A. and Pollefeys M. Patch based confidence prediction for dense disparity map. In BMVC, 2016. 2

[2] Spyroupolous A., Komodakis N., and Mordohai P. Learning to detect ground controlpoints for improving the accuracy of stereo matching. In CVPR, 2014. 2

[3] Henrik Aanæs, Rasmus Ramsbol Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. International Journal of Computer Vision (IJCV), 120(2):153–168, 2016. 3

[4] Christian Bailar, Manuel Finckh, and Hendrik P. A. Lensch. Scale robust multi view stereo. In ECCV, 2012. 2

[5] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. In ACM SIGGRAPH 2009 Papers, SIGGRAPH ’09, 2009. 2, 3

[6] Jonathan T. Barron and Ben Poole. The fast bilateral solver. In European Computer Vision Conference, ECCV 2016, Amsterdam, The Netherlands, pages 617–632, 2016. 2

[7] M. Bleyer, C. Rhemann, and C. Rother. PatchMatch stereo - stereo matching with slanted support windows. In CVPR, 2007. 2

[8] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In ICCV, 2015. 1, 2, 3

[9] D. Gallup, J. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In CVPR, 2007. 2

[10] Heiko Hirschm¨uller. Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 30(2):328–341, 2008. 1, 2, 5

[11] H. Hirschm¨uller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 31:1582–1599, 2008. 2

[12] P.-H. Huang, K. Maten, J. Knop, N. Ahuja, and J.-B. Huang. DeepMVS: Learning multi-view stereopsis. In CVPR, 2018. 1, 2

[13] Sunok Kim, Seungryong Kim, Dongbo Min, and Kwanhoon Sohn. L af-net: Locally adaptive fusion networks for stereo confidence estimation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 2

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. 6

[15] Vladimir Kolmogorov and Ramin Zabih. Multi-camera scene reconstruction via graph cuts. In Proceedings of the 7th European Conference on Computer Vision-Part III, ECCV ’02, pages 82–96, Berlin, Heidelberg, 2002. Springer-Verlag. 1, 2

[16] Andreas Kuhn, Heiko Hirschm¨uller, Daniel Scharstein, and Helmut Mayer. A TV prior for high-quality scalable multi-view stereo reconstruction. International Journal of Computer Vision (IJCV), 124(1):2–17, 2017. 4

[17] Andreas Kuhn, Shan Lin, and Oliver Erdler. Plane completion and filtering for multi-view stereo reconstruction. In GCPR, 2019. 1, 2, 4, 5, 7, 8

[18] Poggi M., Tosì F., and Mattoccia S. Quantitative evaluation of confidence measure sin a machine learning world. In ICCV, 2017. 2

[19] Poggi M. and Mattoccia S. Learning a general-purpose confidence measure based on(1) features and a smarter aggregation strategy for semi global matching. In 3DV, 2016. 2

[20] Poggi M. and Mattoccia S. Learning from scratch a confidence measure. In BMVC, 2017. 2, 7

[21] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 3, 4

[22] Park M.G. and Yoon K.J. Leveraging stereo matching with learning-based confidence measures. In CVPR, 2015. 2

[23] Nicholas Molton and Michael Brady. Practical structure and motion from stereo when motion is unconstrained. International Journal of Computer Vision (IJCV), 39(1):5–23, 2000. 4

[24] Min-Gyu Park and Kuk-Jin Yoon. As-planar-as-possible depth map estimation. Computer Vision and Image Understanding, 181:50–59, Apr. 2019. 2

[25] Knöbberreiter Patrick and Pock Thomas. Learned collaborative stereo refinement. In German Conference on Pattern Recognition, GCPR 2019, Dortmund, Germany, 2019. 3

[26] Haeusler R., Nair R., and Kondermann D. Ensemble learning for confidence measures in stereo vision. In CVPR, 2013. 2

[27] A. Romanoni and M. Matteucci. TAPA-MVS: Textureless-aware PatchMatch multi-view stereo. CoRR, abs/1903.10929, 2019. 2, 5, 7

[28] Mattia Rossi, Mireille El Gheche, Andreas Kuhn, and Pascal Frossard. Joint graph-based depth refinement and normal estimation. In arXiv:1912.01306 [cs.CV], 2019. 2, 3, 5, 6, 8

[29] D. Scharstein, H. Hirschmiller, Y. Kitajima, G. Krathwohl, N. Niesic, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In GCPR, 2014. 3, 4

[30] Johannes Lutz Schönb¨urger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In ECCV, 2016. 1, 2, 3, 5, 6, 7

[31] Thomas Schöps, Johannes L. Schönb¨urger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In CVPR, 2017. 1, 2, 3, 4, 6, 7, 8

[32] Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In CVPR, 2006. 2, 3

[33] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In Field and Service Robotics, 2017. 3
Gidaris Spyros and Komodakis Nikos. Detect, replace, refine: Deep structured prediction for pixel wise labeling. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA*, 2017.

Jian Sun, Nan-Ning Zheng, and Heung-Yeung Shum. Stereo matching using belief propagation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(7):787–800, July 2003.

Fabio Tosi, Matteo Poggi, Antonio Benincasa, and Stefano Mattoccia. Beyond local reasoning for stereo confidence estimation with deep learning. In *ECCV*, 2018.

Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Leveraging confident points for accurate depth refinement on embedded systems. In *IEEE Embedded Vision Workshop, EVW 2019, Long Beach, CA, USA*, 2019.

Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. *CoRR*, abs/1904.08103, 2019.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. *ECCV*, 2018.

Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. *Computer Vision and Pattern Recognition (CVPR)*, 2019.

E. Zheng, E. Dunn, V. Jojic, and J.M. Frahm. PatchMatch based joint view selection and depthmap estimation. In *CVPR*, 2014.