Confidence-based Large-scale Dense Multi-view Stereo

Zhaoxin Li, Wangmeng Zuo, Zhaoqi Wang and Lei Zhang

***Abstract*—Albeit remarkable progress has been made to improve the accuracy and completeness of multi-view stereo (MVS), existing methods still suffer from either sparse reconstructions of low-textured surfaces or heavy computational burden. In this paper, we propose a Confidence-based Large-scale Dense Multi-view Stereo (CLD-MVS) method for high resolution imagery. Firstly, we formulate MVS as a multi-view depth estimation problem, and employ a normal-aware efficient PatchMatch stereo to estimate the initial depth and normal map for each reference view. A self-supervised deep learning method is then developed to predict the spatial confidence for multi-view depth maps, which is combined with cross-view consistency to generate the ground control points. Subsequently, a confidence-driven and boundary-aware interpolation scheme using static and dynamic guidance is adopted to synthesize dense depth and normal maps. Finally, a refinement procedure which leverages synthesized depth and normal as prior is conducted to estimate cross-view consistent surface. Experiments show that the proposed CLD-MVS method achieves high geometric completeness while preserving fine-scale details. In particular, it has ranked No. 1 on the ETH3D high-resolution MVS benchmark in terms of F_1-score.**

***Index Terms*—Multi-view stereo, confidence, large-scale, interpolation, static and dynamic guidance, refinement.**

I. INTRODUCTION

MULTI-view stereo (MVS) is an important research topic in computer vision, which aims at reconstructing 3D geometric surface of scenes from multiple overlapped images. MVS allows user to capture images using consumer cameras instead of specialized devices, and can be applied in both indoor and outdoor scenes. Along with the prevalence of cell-phones, digital cameras and unmanned aerial vehicles, there is a steadily increasing demand for MVS in many applications, such as large-scale urban reconstruction [1], [2] and image-based rendering [3]. Driven by a series of MVS benchmarks [4]–[6], the performance of MVS algorithms [1], [7]–[10] has been consistently improved in terms of both reconstruction completeness and accuracy. However, while the state-of-the-art methods have achieved high reconstruction accuracy on highly textured and Lambertian surface, they perform limited for reliable reconstruction in real-world man-made scenes containing low-textured and non-Lambertian surfaces. In particular, the unreliable multi-view matching on these scenes gives rise to gross defects in the form of severe noise and sparse reconstruction, inevitably restricting the deployment of MVS in many applications that visual quality matters [3].

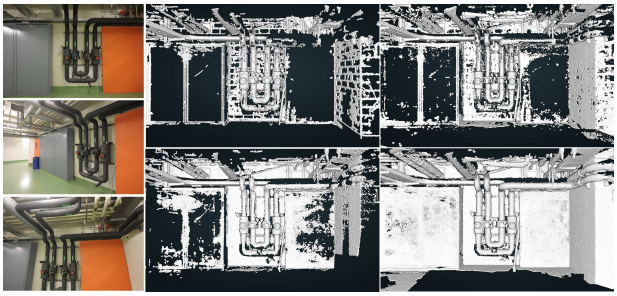


Fig. 1: Reconstruction results of *pipes* from ETH high-resolution MVS benchmark. The scene consists of low-textured and non-Lambertian man-made objects. Three of input images are shown in the left, and 3D point-clouds by PMVS [8], COLMAP [10], LTVRE [2] and our method are shown in upper middle, lower middle, upper right and lower right, respectively. It clearly shows that our method can achieve higher completeness while visually more pleasing.

To improve the completeness and accuracy for reconstructing complex scenes, prior terms have been introduced to alleviate unreliable matching. The regularization terms [11]–[13] have been integrated into a well-posed global labeling problem to impose pairwise smoothness prior for estimating per view depth maps [12] or directly recovering consistent 3D surface on 3D voxels [11], [13]. However, these methods generally require heavy computational efforts and costly memory consumption even on low-resolution imagery, and thus they cannot be applied to large-scale and high resolution imagery. On the other hand, rapid growth of image and video resolution by modern digital camera makes the efficiency another challenging issue for MVS methods. Segmentation-based methods [14] assume that scenes can be represented by a set of planar structures which partition scenes into sub-regions via superpixels, and each superpixel is corresponding to a 3D plane. However, geometric details and thin structures cannot be preserved in the reconstruction due to over-segmentation. Data-driven methods learn shape priors [15], normal distribution [16] and joint reconstruction and semantic labels [17] from training data. These approaches give remarkable results on specific classes of objects with high intra-cluster

This work was supported by National Natural Science Foundation of China under Grants (Nos. 61702482 and 61532002), Hong Kong RGC GRF Grant (PolyU 152216/18E), and the Natural Science Foundation of Beijing (Grant No. L172049).

Z. Li is with Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, and Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong. Email: cszli@hotmail.com

W. Zuo is with School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. Email: cswmzuo@gmail.com

Z. Wang is with Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. Email: zqwang@ict.ac.cn

L. Zhang is with Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong. Email: cslzhang@comp.polyu.edu.hk

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. Contact cszli@hotmail.com for further questions about this work.

similarity, such as cars, humans and buildings. Recently, deep learning-based MVS methods [18], [50] discriminatively train the end-to-end mapping from images to depth maps or 3D volumetric fields. Due to the limitation of training data with available ground-truth, deep learning-based MVS methods are still lacking in generalization ability on reconstructing versatile real-world scenes which usually have quit different data distribution from training data.

For efficient dense reconstruction, interpolation-based methods [19], [21] adopt a two-stage scheme to reconstruct depth maps. In particular, the initial reconstruction stage generates sparse feature points [21] or semi-dense depth map [19] as anchor points, and the interpolation stage propagates confident depth points to form dense depth map by using color image as guidance. However, physically unrealistic interpolations may happen for the regions far from anchor points. To further refine the interpolation results, [23] introduces a large-scale stereo matching method including three stages: a set of anchor points are detected via Sobel features, Delaunay triangulation is used to interpolate anchor points to provide an initial estimation for disparity map, and then refinement is performed with prior constraint. The main drawback of this method is that it relies on salient features which are less reliable on low-texture regions, and the interpolation via Delaunay triangulation on anchor points cannot preserve surface boundaries. Instead of relying on specific feature points, [3] uses a set of handcrafted filters to select semi-dense anchor points from noisy depth maps, then the interpolated maps estimated by the first-order Poisson system is used as an auxiliary term of MRF for global optimization of discrete labels. However, global optimization needs huge computational efforts and it also cannot achieve sub-pixel accuracy. Moreover, color images used as guidance may lead to textured-copy problem [22], because most image textures are actually not corresponding to true surface boundaries.

In this work, we follow the line of interpolation-based methods and make a series of improvements. To begin with, normal-aware PatchMatch stereo is adopted as the backbone of label optimization for maintaining sub-pixel reconstruction accuracy. We then propose a data-driven confidence prediction method that adaptively predicts spatial consistency of depth maps based on self-supervised learning, which does not rely on salient features, hand-crafted filters and ground-truth depth. The proposed method combines both spatial and cross-view consistency to accurately detect confident depth points. Finally, we present a confidence-driven and boundary-aware interpolation stage to estimate physical-realistic depth and normal prior maps by utilizing the estimated confidence map and dynamic guidance, which are then used as constraints for efficient pixel-wise refinement.

In summary, we propose a confidence-based large-scale dense multi-view stereo method (CLD-MVS). The proposed method consists of four key stages: coarse depth and normal estimation, data-driven confidence prediction, plausible depth prior construction via interpolation and pixel-wise refinement, as shown in Fig. 2. The main contributions are summarized as follows:

- We propose a novel CLD-MVS pipeline consisting of initial reconstruction, confidence prediction, confidence-driven and boundary-aware interpolation, and pixel-wise refinement. The pipeline is robust to low-textured surface and efficient for large-scale scenes and images.
- A normal-aware PatchMatch stereo is presented to improve accuracy of normal map.
- A confidence-driven and boundary-aware interpolation is proposed to obtain plausible prior depth maps. We combine both spatial and cross-view confidence measures for detecting reliable ground control points. Moreover, a self-supervised method is proposed to learn spatial confidence of multi-view depth maps. The use of the second order static and dynamic guidance benefits the reconstruction of slanted surface while increasing robustness to noise and inconsistency of boundaries between image and depth.
- The proposed refinement step improves accuracy and cross-view consistency by constraining photometric matching costs with interpolation results, thereby resulting in both higher completeness and better detailed reconstruction.

Among all published results on the ETH3D high-resolution MVS benchmark [6], the proposed CLD-MVS method has achieved the highest rank in terms of F_1-scores, which is a comprehensive metric for both accuracy and completeness. A qualitative comparison of CLD-MVS with the state-of-the-art methods on a challenging ETH3D dataset is shown in Fig. 1. We also achieve the best performance on a set of challenging datasets from DTU benchmarks [5], including vegetations, and man-made objects with textureless and specular parts.

The paper is organized as follows. Section II introduces the related work. Section III presents the framework of the proposed method in detail. Section IV provides the experimental results. Finally, the paper is concluded in Section V.

II. Related Work

According to the taxonomy proposed by [4], MVS approaches can be divide into four categories: volumetric-based [13], mesh-based [24], feature-based [1], [8] and depth map-based methods [9], [10], [25]. Actually, these methods are not independent to each other. Some practical MVS pipelines [2], [26], [27] combine two or more categories of methods in different stages to achieve high-quality reconstruction.

A. Volumetric-based MVS

Volumetric-based MVS methods assume that 3D scenes embed in a pre-defined 3D volume. By labeling 3D voxels as outside or inside true surface, 3D reconstruction can be treated as a 3D segmentation problem. In the early studies, voxels are labeled by photoconsistency [28], and inconsistent voxels are then removed from 3D volume. This kind of methods may lead to noisy reconstruction since it highly depends on the photoconsistency which is sensitive to matching ambiguity. By enforcing surface smoothness in energy function [13], smoothing 3D reconstruction can be attained. There are some methods [2], [11] that fuse multi-view depth maps in 3D volumetric grids. These methods initialize voxel labels using noisy depth maps, and output a fused 3D surface by labeling

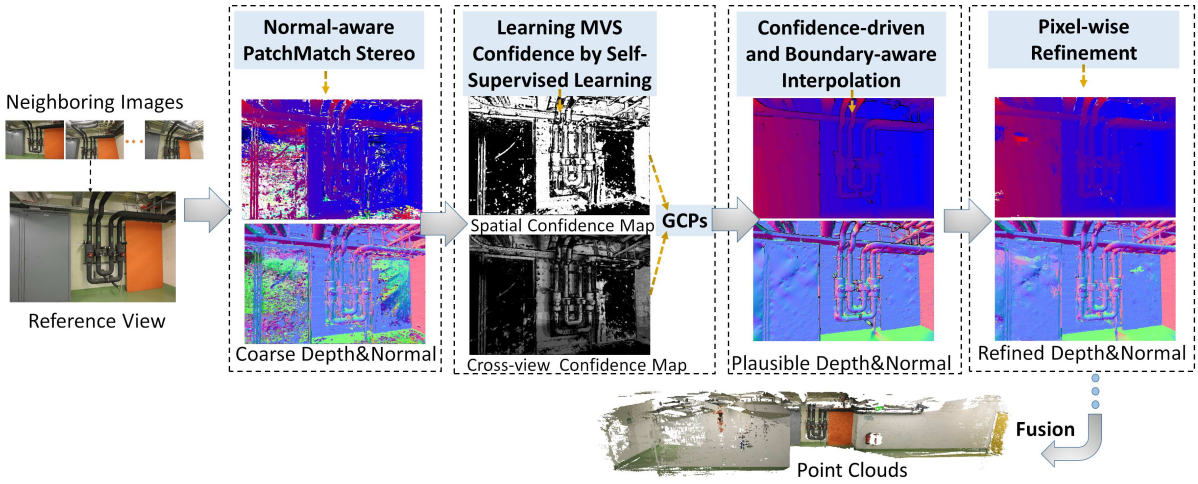


Fig. 2: Overview of the proposed CLD-MVS method

voxels via the accumulation of votes from multiple depth maps or optimization of energy function. Some semantic methods [16], [17] use volumetric methods as backbone for optimizing semantic 3D reconstruction while incorporating normal prior as an additional constraint term [16] or joint optimization of 3D reconstruction and semantic segmentation [17]. However, volumetric-based methods generally suffer from huge computation and memory cost: (i) resolving energy minimization on 3D volume requires huge computational cost; (ii) volumetric-based surface representation needs huge memory to represent detailed surface and thin structures.

B. Mesh-based MVS

Mesh-based MVS methods optimize energy functions defined on triangular meshes to directly deform estimated surface to true surface by minimizing reprojection errors [24], [26]. The surface regularization terms are defined on vertices of triangular meshes to explicitly control reconstruction quality. Given an initial surface obtained from visual hull or other MVS reconstructions, mesh-based methods can refine coarse 3D surface to achieve detailed reconstruction while eliminating surface artifacts. Moreover, mesh-based methods [29], [30] can incorporate with shading cues to recover fine-scale geometric structures. However, mesh-based methods are limited in handling topological changes.

C. Feature-based MVS

Feature-based MVS methods focus on reconstructing the regions with discriminative salient features. Furukawa et al. [8] proposed the PMVS method that initially estimates a set of seeds based on Harris corners and Blob detection, then propagates the estimated depth to neighbors in a region growing manner. Tola et al. [1] suggested to use Daisy feature for reliable multi-view matching to generate quasi-dense depth maps. This type of methods can efficiently reconstruct large-scale scenes and process high-resolution images. However, reconstructions by feature-based methods lead to large irregular holes for the regions where salient feature points are hard to be detected and matched. The low geometric completeness has prevented them from the application tasks where dense reconstruction is an important concern.

D. Depth map-based MVS

Depth map-based MVS methods reconstruct 3D surface in two stages, i.e. multi-view depth estimation and fusion. Depth map for each input view is estimated based on view selection and local correspondence search, and then outlier removal and fusion are conducted based on point-cloud representation [9], [10], [31] or volumetric representation [2], [11]. Geselese et al. [25] proposed an efficient multi-view depth estimation method to estimate depth only for regions with strong image gradients. Bailer et al. [32] presented a scale robust MVS for unstructured image datasets that can deal with large variations in surface sampling rate. Galliani et al. [9] proposed an GPU-friendly PatchMatch stereo by alternating between hypothesis propagation and hypothesis refinement based on black-red pattern, and fused depth maps based on an efficient cross-view depth and normal consistency checking. Xu et al. [33] improved propagation speed of [9] by utilizing an asymmetric pattern that gives priority to hypotheses with smaller photometric costs. Schönberger et al. [10] proposed an occlusion-aware multi-view stereo that jointly estimates pixel-wise visibility, normal maps and depth maps, then consistency filters based on photometric cost and geometric consistency are used to remove the outliers when fusing the depth maps. Although depth map-based MVS methods can achieve accurate and efficient reconstruction on textured regions, they usually result in low reconstruction completeness of low-textured surface, such as man-made scenes. In contrast, we present a novel depth map-based MVS method which effectively exploits confident reconstruction parts to help the reconstruction of unreliable parts, thereby resulting in significant improvement of the entire reconstruction quality, especially on low-textured surface.

E. Confidence Prediction

The confidence measures have been demonstrated to be helpful in improving disparity map in stereo matching. The

high confidence estimates are used to modulate the data term in MRF [34] and semi-global stereo [35], [36]. A number of confidence measures have been suggested in stereo matching. Instead of hand-crafted features, the state-of-the-art methods usually adopt the learning-based strategy [35], [36], which train random forest [34], [35] or deep CNNs [36], [37] based on ground-truth disparities. Besides the supervised learning methods, several methods [38], [39] have attempted to design self-supervised strategy for training. Mostegel et al. [38] proposed a self-supervised method that constructs a set of training samples based on self-contradiction of depth maps estimated from continuous stereo frames sampled from video sequences. Tosi et al. [39] conservatively selected a set of samples based on a serial of weak classifiers, e.g., left-right consistency, median deviation of disparity, winner margin, uniqueness constraint, and then trained a CNN-based model for confidence prediction. In contrast, confidence prediction has attracted less attention on multi-view stereo [2], [40], partially because accumulation of photometric costs from multiple views is more robust than stereo matching, and cross-view consistency provides a strong evidence for probable noise and outliers. However, these measures become less reliable for sparse input views and low-textured surfaces. In this work, we propose a self-supervised strategy for learning spatial consistency of multi-view depth maps without the requirement of ground-truths information. The predicted spatial consistency can then be integrated with cross-view consistency to detect high-quality ground control points.

III. Proposed Approach

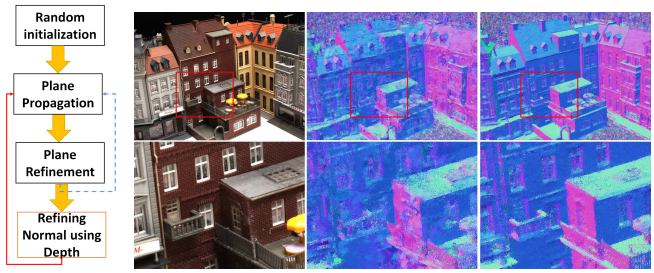
The overview of the proposed CLD-MVS method is illustrated in Fig. 2. The main idea of the proposed method is based on the assumption that real-world scenes contain many low-textured surfaces which are difficult to be reliably reconstructed via multi-view matching, and we propose to make full use of reliable reconstruction parts to help the reconstruction of unreliable parts. To achieve this goal, we first need an initial dense reconstruction and a method to reliably detect the confident reconstruction parts, then use a reliable interpolation method for filling missing surfaces while utilizing a refinement step for improving cross-view consistency and geometric details. Given a reference view, our method estimates both dense depth and normal map in four stages.

(1) The coarse depth and normal maps are estimated using a normal-aware PatchMatch stereo algorithm.

(2) The confidence measures in terms of cross-view consistency and spatial consistency are integrated to reliably predict the correctness of each depth hypothesis. A self-supervised strategy is proposed to generate labeled samples for training deep CNNs (DCNNs) to predict spatial consistency.

(3) Plausible depth and normal estimation are obtained by a confidence-driven and boundary-aware interpolation using both dynamic and static guidance.

(4) Depth and normal are further refined to eliminate possible interpolation errors, improve cross-view consistency and recover geometric details.



(a) (b)

Fig. 3: (a) The pipeline of normal-aware PatchMatch stereo. Compared to baseline method (blue dotted line), we further refine normal via cross-product of depth map (red solid line). (b) First row, from left to right, input image, normal map by baseline method and normal map by the proposed normal-aware variant. Second row shows the corresponding close-ups.

It is worth noting that all steps in the proposed MVS pipeline collaborate closely, and each step builds upon the previous one and is essential for high-quality dense reconstruction. The proposed normal-aware PatchMatch stereo significantly improves accuracy of normals, and the initial reconstruction increases the density of confident points and reduces the errors in the subsequent interpolation step. For confidence prediction, high false positive rate will propagate the reconstruction errors while high false negative rate will lead to the unreliable interpolation due to the lack of reference depth information. The proposed confidence prediction utilizes both spatial confidence and cross-view consistency to achieve more reliable confidence prediction for MVS. The interpolation step uses confident points predicted by the proposed confidence prediction step for reasonable interpolation of the unreliable parts. The proposed pixel-wise refinement step utilizes interpolation results as observed maps to estimate more faithful 3D surface, especially for fine-scale structures, depth discontinuities and regions far from confident points. After estimating depth and normal maps for each of input views, the entire set of depth and normal maps are fused to oriented point clouds.

A. Multi-view Depth Estimation via Normal-aware PatchMatch Stereo

Let $S \subset \mathbb{R}^3$ denote a 3D surface of a scene, and $I_i : \Omega_i \subset \mathbb{R}^2 \rightarrow \mathbb{R}^\xi$ be an observed image in camera i ($\xi = 1$ for grayscale images, and $\xi = 3$ for color images). The goal of multi-view depth map estimation (MVDE) is to assign a dense label map $U : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}^4$ for reference view I given its neighboring views $\mathcal{J} = \{J_k | k = 1, 2, ..., K\}$, where U consists of one channel for depth map $D : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ and three channels for normal map $N : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$. Stereo matching can be treated as a special case of MVDE, where two input images are rectified and $K = 1$.

Given a label assignment $\mathbf{u}(\mathbf{p}) \in U$, for an $m \times m$ patch R_p centered on $\mathbf{p}$ in I, the corresponding patch in one of K neighboring views is $R_{p'_k}$, and $\mathbf{p}'_k$ is the correspondence of $\mathbf{p}$. Let $\rho(\mathbf{p}, \mathbf{p}'_k)$ be photometric matching cost between R_p and $R_{p'_k}$. To improve robustness to radiometric distortion and depth discontinuities, we use negative adaptive normalized cross-

correlation (ANCC) [41] to measure photometric matching cost:

$$\rho(\mathbf{p}, \mathbf{p}'_k) = -\frac{\sum\limits_{q\in R_p, q'_k\in R_{p'_k}} w_p^q w_{p'_k}^{q'_k} \left(I(\mathbf{q}) - \bar{A}_p\right)\left(J_k(\mathbf{q}'_k) - \bar{A}_{p'_k}\right)}{\sqrt{\sum\limits_{q\in R_p} \left|w_p^q \left(I(\mathbf{q}) - \bar{A}_p\right)\right|^2} \sqrt{\sum\limits_{q'_k\in R_{p'_k}} \left|w_{p'_k}^{q'_k} \left(J_k(\mathbf{q}'_k) - \bar{A}_{p'_k}\right)\right|^2}} \quad (1)$$

where $\mathbf{q} \in R_p$ and $\mathbf{q}'_k \in R_{p'_k}$ are a pair of correspondences between reference image and neighboring image, $\bar{A}_p = \sum_{q\in R_p} w_p^q I(\mathbf{q})$ and $\bar{A}_{p'_k} = \sum_{q'_k\in R_{p'_k}} w_{p'_k}^{q'_k} J(\mathbf{q}'_k)$ are weighted means of intensity values of pixels within patch R_p and patch $R_{p'_k}$, respectively. We define the affinity weight as $w_p^q = \exp\left(-\frac{\|I(p)-I(q)\|_1}{\gamma}\right)$, where γ is a user-defined parameter and $\|I(\mathbf{p}) - I(\mathbf{q})\|_1$ computes the L_1-distance between $I(\mathbf{p})$ and $I(\mathbf{q})$. The affinity weight decreases the influence of pixels that differ a lot from the central one.

To increase the robustness to occlusion, accumulation of photometric costs with respect to all corresponding patches in neighboring images is defined as a self-weighted average of the best K_b costs:

$$\Phi(\mathbf{p}, \mathbf{u}(\mathbf{p})) = \frac{\sum_{k=1}^{K_b} \frac{1}{\rho(\mathbf{p},\mathbf{p}'_k)} \rho(\mathbf{p}, \mathbf{p}'_k)}{\sum_{k=1}^{K_b} \frac{1}{\rho(\mathbf{p},\mathbf{p}'_k)}} = \frac{K_b}{\sum_{k=1}^{K_b} \frac{1}{\rho(\mathbf{p},\mathbf{p}'_k)}} \quad (2)$$

This correspondence problem is solved by PatchMatch stereo for sub-pixel accuracy. The PatchMatch stereo was initially introduced in stereo matching [42], and then adopted in MVS methods [9], [10], [32] due to its effectiveness for sub-pixel accuracy and slanted surface. PatchMatch stereo consists of random hypothesis initialization, hypothesis propagation and hypothesis refinement. While the original algorithm sequentially propagates hypotheses, several GPU-friendly variants have been proposed to accelerate computation by using scanlines along diagonals [32] and black-red pattern [9], [10]. Based its highly efficiency for high-resolution images, we use the black-red alternating optimization framework proposed by Galliani et al. [9] as the backbone for the PatchMatch stereo. To further accelerate the hypothesis propagation, we use an improved red-black pattern proposed by [33]. To select neighboring view set $\mathcal{J}$ for a reference view, Goesele et al. [25] proposed a global view selection method followed by a local view selection based on sparse features from outputs of SFM algorithm, triangulation angles and matching cost in MVS, and Bailer et al. [32] improved global view selection method of [25] by prioritizing views with big view angles to reference view and views that capture regions which are not visible to those already selected images. Schönberger et al. [10] further proposed a pixel-wise view selection method for elaborately handling occlusions. Since we have used occlusion-robust cost aggregation function in Eq. 2, we follow the global view selection method of [32] to efficiently select a set of neighboring view K for a reference view. We denote the method introduced above as baseline PatchMatch stereo.

In particular, accurate estimate of normals is important for recovering detailed surface and slanted geometric structure. However, since there is one more degree of freedom than depth, correct normal hypothesis is more difficult to be found. In most variants of PatchMatch stereo [9], [42], normal and depth are independently estimated. To further improve accuracy of normals, Schönberger et al. [10] used additional hypotheses to refine normals via random guess and local perturbation, which could be difficult to converge. Based on the observation that accuracy of estimated normal is always lower than accuracy of the corresponding depth in the optimization procedure, we generate a set of high-quality normal hypotheses based on depth estimates in each iteration, as shown in Fig. 3(a). Firstly, we filter depth map via a 3×3 median filter, then estimate normal hypotheses via cross-product of filtered depth maps. Formally, we generate new normal hypothesis $\bar{\mathbf{n}}_p$ for $\mathbf{p}$ by cross product of four neighbors of $\mathbf{p}$:

$$\bar{\mathbf{n}}_p = \frac{(\bar{\mathbf{x}}_l - \bar{\mathbf{x}}_r) \otimes (\bar{\mathbf{x}}_u - \bar{\mathbf{x}}_d)}{\|(\bar{\mathbf{x}}_l - \bar{\mathbf{x}}_r) \otimes (\bar{\mathbf{x}}_u - \bar{\mathbf{x}}_d)\|_2} \quad (3)$$

where $\otimes$ is cross-product operator, $\mathbf{p}_l$, $\mathbf{p}_r$, $\mathbf{p}_u$ and $\mathbf{p}_d$ are 4-connected pixels of $\mathbf{p}$ in the horizontal and vertical directions, respectively. $\bar{\mathbf{x}}_l = \mathrm{K}^{-1}(x_{p_l}, y_{p_l}, 1)z_{p_l}$, $\bar{\mathbf{x}}_r = \mathrm{K}^{-1}(x_{p_r}, y_{p_r}, 1)z_{p_r}$, $\bar{\mathbf{x}}_u = \mathrm{K}^{-1}(x_{p_u}, y_{p_u}, 1)z_{p_u}$, $\bar{\mathbf{x}}_d = \mathrm{K}^{-1}(x_{p_d}, y_{p_d}, 1)z_{p_d}$ are four 3D points in the camera coordinate system, where $z_{p_l}, z_{p_r}, z_{p_u}, z_{p_d}$ are the corresponding depth hypotheses and K is the camera intrinsic matrix. Note that $\bar{\mathbf{n}}_p$ in Eq. 3 is defined in camera space, which can be transformed into the global 3D space by $-\mathrm{R}^T\bar{\mathbf{n}}_p$, where R is a rotation matrix from global 3D space to camera space.

By using this simple procedure, we can significantly improve accuracy of normals with only slight increase of computational cost. Fig. 3(b) shows a visual comparison of estimated normal maps with and without the proposed normal refinement strategy. The quantitative evaluation are conducted in Section IV-A.

Finally, the outputs of MVDE are a set of dense depth map $\mathcal{D}^* = \{D^*\}$ and normal map $\mathcal{N}^* = \{N^*\}$ for all input views $\mathcal{I}$.

B. Confidence Prediction for Multi-view Depth Estimation

Due to matching ambiguities and occlusions, coarse depth maps by PatchMatch stereo are inevitably contaminated by noise and gross outliers. Instead of giving a hard classification between *correct* and *wrong*, we set confidence value to each pixel for its depth and normal assignment. In particular, a good depth hypothesis in MVS should be geometrically stable with support both from other views and its spatial neighbors in the same depth map. We thus use two complementary confidence measures to evaluate correctness of depth estimation based on cross-view geometric consistency and spatial consistency, respectively. In particular, we propose a self-supervised learning method for predicting spatial consistency.

The cross-view confidence measure $f_v(z_p)$ is based on cross-view consistency (visibility). It utilizes the redundancy of MVS and assumes that a correct depth hypothesis should be consistent to a number of neighboring depth maps. Let $z_p \in D^*$ and $\mathbf{n}_p \in N^*$ be depth and normal in pixel $\mathbf{p} = (x_p, y_p)$ of reference view and D_i^* be a depth map of neighboring view i, where $\mathbf{n}_p$ defined in the global space. Let $(\mathbf{q}_i, 1)^T =$

$\pi \circ \left(\Pi_i \circ \left(\Pi^{-1} \circ (x_p, y_p, 1)^T z_p\right)\right)$ be the correspondence of $\mathbf{p}$ in D_i^*, where Π^{-1} and Π_i represents camera inverse projection and projection operator in reference view and neighboring view i, respectively, and $\pi = (x/z, y/z, z/z)$ is a normalization operator. The reprojection position $\hat{\mathbf{p}}_i$ of $\mathbf{p}$ is defined by projecting $\mathbf{q}_i$ back to reference view. Depth hypothesis $z_p \in D^*$ is regarded to be consistent with neighboring view i if (1) distance between reprojection position $\hat{\mathbf{p}}_i$ and $\mathbf{p}$ is less than ϵ_1 and (2) angle between normal $\mathbf{n}_p$ in $\mathbf{p}$ and normal $\mathbf{n}_{q_i}$ in $\mathbf{q}_i$ is less than ϵ_2,

$$v_{z_p,i} = \begin{cases} 1 & \left\|\mathbf{p} - \hat{\mathbf{p}}_i\right\|_2 < \epsilon_1 \cap \arccos\left(\mathbf{n}_p \cdot \mathbf{n}_{q_i}\right) < \epsilon_2 \\ 0 & otherwise \end{cases} \tag{4}$$

where the normal $\mathbf{n}_p$ and normal $\mathbf{n}_{q_i}$ are all defined in the global 3D space, and ϵ_1 and ϵ_2 are two thresholds for reprojection error and normal consistency, respectively.

The more reliable depth estimation will be voted by more neighboring views. We use summation of consistency over all neighboring views as a confidence for z_p:

$$f_v(z_p) = \sum_{i \in \mathcal{I}} v_{z_p,i} \tag{5}$$

If the scene is densely covered by images, the cross-view confidence measure can provide a strong confidence for detecting inliers and outliers. However, if input views are sparse (e.g., some parts of scene are observed by only two views), this confidence measure becomes less reliable.

The spatial confidence measure $f_s(z_p)$ is based on spatial consistency. This is based on the assumption that a correct depth hypotheses should be consistent with its spatial neighbors in the same depth map. Traditional methods for detecting spatial confidence measures use some heuristic features, such as dissimilarity from all surrounding points [23], difference with median [3], [34] and total variation [2]. To adaptively explore the spatial consistency, we train a deep convolutional neural network (DCNN) $\mathcal{M}(R_p)$ that inputs a depth image patch R_p centering on pixel $\mathbf{p}$ and outputs a confidence measure $f_s(z_p)$ for the depth z_p. The network structure is inspired by an effective DCNN-based method [36] for stereo matching. Poggi et al. [36] proposed to directly learn spatial consistency from scratch by DCNN, which uses a fully convolutional network that inputs the small depth patches and returns the confidence of the center pixel. According to the quantitative evaluation conducted by [47], [36] is the most accurate data-driven confidence prediction method for stereo matching. Moreover, in contrast to other data-driven methods [34], [35], the only inputs of [36] are depth maps, so it can learn spatial confidence independent to other features, e.g., photometric costs. Based on these works, we use the similar network structure of [36] to learn the spatial confidence for depth maps estimated by MVS. Different from [36], we use a 15×15 depth image patch R_p as input and a deeper convolutional network (Fig. 4(a)) which has shown better performance in our experiments (Fig. 10). As shown in Fig. 4(a), the network $\mathcal{M}$ consists of seven 3×3 convolution layers and three 1×1 convolution layers. Each of convolution layers are followed by a ReLU (Rectified Linear Units) nonlinearity except the last convolution layer where softmax is adopted.

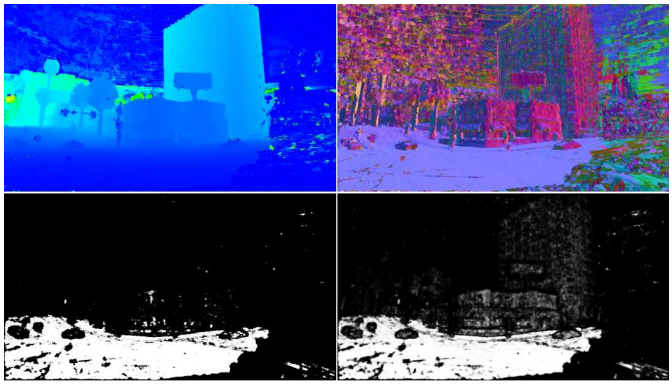


Fig. 5: Illustration of confidence prediction by hard classification and soft classification. First row: one of estimated depth maps and the corresponding normal map. Second row, from left to right: confidence map based on hard classification and soft classification, respectively.

The receptive field of this DCNN is 15×15. The loss for training the DCNN is cross-entropy:

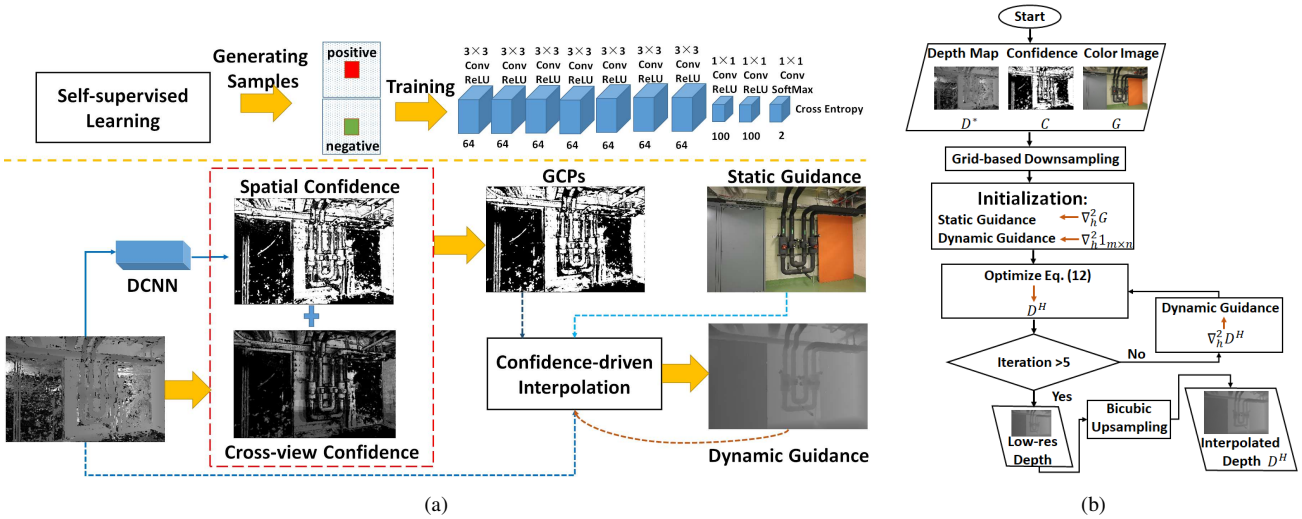
$$\mathcal{L} = -\frac{1}{n} \sum_p \left(gt_{z_p} \log\left(f_s(z_p)\right) + (1 - gt_{z_p}) \log\left(1 - f_s(z_p)\right)\right) \tag{6}$$

where gt_{z_p} is ground-truth label of spatial confidence, $f_s(z_p) = \mathcal{M}(R_p)$ is the predicted probability and n is the total number of samples of training data. The confidence map f_s is a probability map where each continuous value indicates the probability of the estimated depth being correct. As shown in Fig. 5, instead of directly removing the less reliable regions (e.g., distant buildings) as done by hard classification, the predicted confidence map sets a lower probability to these regions.

Learning confidence requires sufficient training data, especially for DCNNs. Moreover, to achieve reasonable performance, the characteristics of training data should be similar to the one to be reconstructed at present. However, it is difficult and expensive to build an abundance of ground-truth for different type of real-world scenes. Though synthetic data in the virtual world can be used, the characteristics of these synthetic data are different with real-world captured data in both appearance and geometric structures. In this work, we propose to use a self-supervised method based on cross-view consistency and truncated signed distance fields (TSDFs) to train DCNNs for confidence prediction of multi-view stereo. In comparison to binocular stereo matching, it is very likely for MVS to find a set of pseudo-positive and pseudo-negative depth samples based on redundant viewpoints. Since the labels of training samples are obtained based on the proposed self-supervised method, we use the prefix *pseudo* to emphasize that the labels are different with ground-truth labels obtained by directly comparing estimated depths with ground-truth depths. Our self-supervised method of constructing training samples for spatial consistency prediction is as follows:

To find a set of pseudo-negative samples, an intersection of $f_v(z_p) < 1$ and $\left|\frac{z_p - med(z_p)}{z_p}\right| > \epsilon_3$ are used, where $med(z_p)$ represents median depth of 3×3 window centering on $\mathbf{p}$.

To find a set of pseudo-positive samples, we select pixels whose depth are voted by a large number of neighboring



(a) (b)

Fig. 4: Illustration of the proposed confidence-driven and boundary-aware interpolation. (a) Overview of the method. (b) Algorithm flow chart of the interpolation method.

views, i.e. $f_v(z_p) > \eta_a$. Meanwhile, we use a confidence measure $f_t(z_p)$ based on truncated signed distance field to further eliminate false positive samples. Concretely, given a reference view, a 3D point $\mathbf{x} \in \mathcal{X}$ in the world coordinate space is calculated using depth z_p and camera inverse projection Π^{-1}: $\mathbf{x} = \Pi^{-1} \circ (x_p, y_p, 1)^T z_p$. The projected depth $z^x_{q_i}$ of $\mathbf{x}$ in neighboring view i is calculated by $(\mathbf{q}_i, 1)^T z^x_{q_i} = \Pi_i \circ \mathbf{x}$. Let z_{q_i} be bilinear depth interpolated by four neighboring pixels of $\mathbf{q}_i$ using depth map D^*_i, the difference between $z^x_{q_i}$ and z_{q_i} measures signed distance from 3D point $\mathbf{x}$ to the surface estimated by D^*_i [31]. We use truncated signed distance to measure geometric consistency between z_p and view i:

$$\text{tsdf}_{z_p,i} = \begin{cases} 1 & (z_{q_i} - z^x_{q_i}) > \delta \\ \frac{z_{q_i} - z^x_{q_i}}{\delta} & \left|z_{q_i} - z^x_{q_i}\right| \le \delta \\ -1 & (z_{q_i} - z^x_{q_i}) < -\delta \end{cases} \tag{7}$$

where δ is a parameter measuring uncertainty of near surface. Similar to volumetric fusion method in [11], we use a binary weight to consider uncertainty due to the occlusion,

$$w_{z_p,i} = \begin{cases} 0 & (z_{q_i} - z^x_{q_i}) < -\eta_b \\ 1 & (z_{q_i} - z^x_{q_i}) \ge -\eta_b \end{cases} \tag{8}$$

where $\eta_b = 3\delta$ controls the width of the occluded region behind the surface. $f_t(z_p)$ is a summation of weighted TSDFs on all neighboring views:

$$f_t(z_p) = \sum_{i \in \mathcal{J}} w_{z_p,i} \text{tsdf}_{z_p,i} \tag{9}$$

A depth estimation z_p is regarded to be consistent with 3D surface if $f_t(z_p)$ satisfies $0 < \left|f_t(z_p)\right| < \tau_t \delta$.

Finally, the pseudo-positive and pseudo-negative samples are extracted from coarse depth maps by:

$$gt_{z_p} = \begin{cases} 1 & f_v(z_p) \ge \eta_a \cap 0 < \left|f_t(z_p)\right| < \tau_t \delta \\ 0 & f_v(z_p) < 1 \cap \left|\frac{z_p - med(z_p)}{z_p}\right| > \epsilon_3 \end{cases} \tag{10}$$

In practice, we empirically set $\eta_a = 5$, $\tau_t = 0.1$, $\epsilon_3 = 0.01$ and δ be 0.1% of maximum depth.

To train the network $\mathcal{M}$, based on Eq. 10, we collect a set of 15×15 patches R_p and the corresponding labels as training samples from training set. We balance the ratio of pseudo-positive and pseudo-negative samples and extract about 5 million samples. Although the set of extracted samples is a subset of entire training set, it is enough to represent the characteristics of spatial consistency for both inliers and outliers. Before patches are fed to the network, we normalize them to zero mean and unit variance. The mean and standard deviation are calculated on entire training samples. Stochastic gradient descent (SGD) is used for training the network in 14 epochs. The batch size is fixed to 128. The learning rate is set to 0.003 for first 10 epochs and decreases to 0.0003 for next 4 epochs. After $\mathcal{M}$ is trained, for a depth map with zero padding in boundaries, the network can output a complete confidence map for each pixel of the depth map. By using the self-supervised method, we can train the spatial confidence prediction model on any type of multi-view images and scenes, and do not depend on whether ground-truths are available.

Cross-view consistency and spatial consistency are complementary. As shown in Section IV-A, by combining both of them, confidence measure can be more robust to scenes and results. We combine these two confidence measures to extract a set of ground control points (GCPs) which are used in the interpolation stage of Section III-C. The final confidence for a depth hypothesis z_p is defined by:

$$C(z_p) = \Lambda\left(f_v(z_p), \tau_v\right) \cdot f_v(z_p) + \lambda_1 \Lambda\left(f_v(z_p), 1\right) \cdot \Lambda\left(f_s(z_p), \tau_s\right) \tag{11}$$

where $\Lambda(f, \theta)$ is a function which equals to 1 for $f \ge \theta$ and 0 otherwise. The depth hypotheses with non-zero confidence are GCPs. Eq. 11 gives more weights to GCPs which are voted by a larger number of neighboring views. In the experiments, parameter τ_s is fixed to 0.6 and τ_v is fixed to 2 for all datasets.

C. Confidence-driven and Boundary-aware Interpolation

Now we have a set of noisy depth maps with corresponding confidence maps $\{\mathcal{D}^*, \mathcal{N}^*, C\}$. Our goal is to integrate these data into a unified energy function so as to estimate smooth and complete depth maps. On the one hand, since some pixels

do not have reliable depth hypotheses, we should faithfully interpolate the missing depth value using large context information; on the other hand, the hypotheses with high confidence will have higher weights and can be allowed to propagate to low confidence regions. The proposed interpolation method adopts the weight least square (WLS) framework [20]. The optimization of WLS-based objective function is highly efficient for the high-resolution images. We also attempt to use noise-robust regularization terms, e.g., total variation (TV) or total generalized variation (TGV). However, these regularization terms showed lower convergence speed and required much longer running time. Moreover, compared with filtering-based methods which can only fill small holes, WLS-based global optimization method can leverage global context information to interpolate large missing surface. Since color and depth map are naturally aligned per view in the multi-view stereo, it is ideal to use color image as guidance for the depth map interpolation based on the assumption that edges and depth discontinuities are co-occupancy. However, it is not necessary that the image edges in the color image are corresponding to the true depth discontinuities. As a result, artifacts occur in the estimated depth maps. Ham et al. [22] proposed a WLS-based method that combines the first-order static and dynamic guidance to improve the robustness to the texture-copy problem for the image restoration task. To increase the robustness to noise and inconsistency of boundaries between image and depth, we follow [22] and further propose to use a combination of second-order static and dynamic guidance which enforces the second-order regularization to respect slanted structures in the real-world scenes. Moreover, instead of using binary weights for the GCPs, the soft weights defined in Eq. 11 have been used to give more weights for the GCPs which are voted by a larger number of neighboring views:

$$E(z_p) = \sum_{\mathbf{p}\in\Omega'} C(z_p)\left(z_p - D^*(\mathbf{p})\right)^2 + \lambda_2 \sum_{\mathbf{p}\in\Omega'} \sum_{h\in\{x,y\}} \phi_\mu\left(\nabla_h^2 G(\mathbf{p})\right)\psi_v\left(\nabla_h^2 D(\mathbf{p})\right) \tag{12}$$

where $\nabla_h^2 G$ and $\nabla_h^2 D$ are the second order difference on guidance color image and depth, respectively, and $h \in \{x, y\}$ indicates the direction of derivatives. $\Omega' \subset \Omega$ is a set of vertices of regular grids that overlay on image domain Ω with grid size of $m_g \times m_g$, and $\mathbf{p} \in \Omega'$ is a vertex of grid. The fidelity term constrains the estimated depth map z_p to fit coarse depth according to confidence measure. Confidence $C(z_p)$ gives a higher penalty for difference between estimated depth z_p and input depth $D^*(\mathbf{p})$ with high confidence, and eliminates influence of input depth with zero confidence. The second term is a regularization term that smooths the solution z_p, and makes its structures similar to static and/or dynamic guidance.

The static guidance is defined as $\phi_\mu(x) = \exp(-\mu x^2)$, and dynamic guidance is defined as $\psi_v(x) = (1 - \phi_v(x))/v$. ψ_v is Welsch function that is robust to outliers. By combining static and dynamic guidance, depth recovery is more robust to noise and less sensitive to inconsistency of boundaries between image and depth. An example of interpolated depth map is shown in Fig. 6.

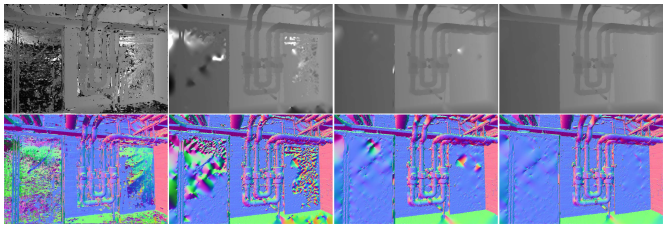


Fig. 6: Different interpolation strategies for prior maps construction. From left to right, coarse depth and normal maps, results by interpolation w/o confidence measure, confidence-driven interpolation w/o dynamic guidance, and the proposed method, respectively

Since the aim of the proposed interpolation is to provide a plausible depth map, it is not necessary to directly perform interpolation on pixel level. In our experiments, the width m_g of each grid is set to 4 for all datasets. The interpolation results for pixels inside grids are obtained by bicubic interpolation of vertices of grids. Fig. 4(b) illustrates the pipeline of the proposed interpolation procedure.

After iterative optimization of Eq. 12 via majorize-minimization algorithm, we obtain a dense and smooth depth map D^H. The corresponding normal map N^H can be calculated directly from D^H via cross product. For outdoor scenes, we remove interpolation results on textureless sky since these are usually not reasonable. We use a simple threshold segmentation based on its effectiveness and simplicity. We first build a sky mask based on image gradient $\|g\| < \tau_g$ based on the assumption that the color of textureless sky varies smoothly. Then we further smooth this mask image using 7×7 median filters. Finally, mask image is eroded with structuring element of 11×11. τ_g is empirically set to 7.5 pixels. Though the high-textured sky, e.g., clouds, could not be removed via above operations, these regions prefer to generate inconsistent depth for different views in the following refinement stage and can be finally removed in depth fusion via cross-view consistency checking. We further remove physically impossible interpolation surfaces: let $\mathbf{c}$ be the optical center of camera, and $\mathbf{x}$ be 3D point converted from $z_p^H \in D^H$ via camera parameters, when the angle between vector $\mathbf{c} - \mathbf{x}$ and $\mathbf{n}_p^H \in N^H$ is higher than a certain threshold τ_n, the estimated depth and normal are discarded from the D^H and N^H. Since $\mathbf{c} - \mathbf{x}$ represents a ray from 3D point to camera center and $\mathbf{n}_p^H$ is the corresponding normal vector, this post-processing operation attempts to remove the interpolated regions that are under oblique observations and usually are unreliable.

By using the confidence-driven and boundary-aware interpolation, we can obtain a set of clean and smooth depth maps and normal maps which can be used as prior to guide the further refinement in Section III-D.

D. Pixel-wise Depth and Normal Refinement

The results by the proposed interpolation can fill in holes while diminishing noise. However, some regions recovered via interpolation are not cross-view consistent and fine-scale geometric details are not kept. Hedman et al. [3] used interpolation results for calculating a near envelope cost then integrated it into MRF for global optimization of discrete

labels to discard erroneous near depth hypotheses. However, this refinement method needs huge computational efforts. In this work, by utilizing interpolation results as prior maps, the proposed pixel-wise refinement stage is conducted via PatchMatch stereo to improve surface details and correct interpolation artifacts. The PatchMatch stereo is initialized by prior depth and normal maps. And photometric matching costs are also constrained by prior maps by introducing prior probability and likelihood.

For each pixel $\mathbf{p} \in \Omega$, the interpolating depth $z_p^H \in D^H$ and normal $\mathbf{n}_p^H \in N^H$ can be used to construct a prior probability. The prior probability of a certain depth and normal hypothesis is formulated as Gaussian distribution:

$$p(z_p, \mathbf{n}_p) \propto \exp\left(-\frac{\left\|z_p - z_p^H\right\|_2^2 + \kappa\left\|\mathbf{n}_p - \mathbf{n}_p^H\right\|_2^2}{\sigma_1^2}\right) \quad (13)$$

The likelihood is formulated as Laplace distribution based on its robustness against noise:

$$p(z_p, \mathbf{n}_p | z_p^H, \mathbf{n}_p^H) \propto \exp\left\{-\frac{\|\Phi(\mathbf{p}, \mathbf{u}(\mathbf{p}))\|_1}{\sigma_2}\right\} \quad (14)$$

where $\Phi(\mathbf{p}, \mathbf{u}(\mathbf{p}))$ denotes photometric matching cost in pixel $\mathbf{p}$ with label assignment $\mathbf{u}(\mathbf{p}) = (z_p, \mathbf{n}_p)$. The depth refinement can be formulated by minimizing the following energy function:

$$E(\mathbf{p}, z_p, \mathbf{n}_p) = \left\|z_p - z_p^H\right\|^2 + \kappa\left\|\mathbf{n}_p - \mathbf{n}_p^H\right\|^2 + \alpha\Phi(\mathbf{p}, z_p, \mathbf{n}_p) \quad (15)$$

where $\alpha = \frac{\sigma_1^2}{\sigma_2}$ is a tradeoff parameter between likelihood and prior, and κ is a tradeoff parameter controlling penalty of depth difference and normal difference. Note that we use the pixel-wise constraint to efficiently conduct the optimization of depth and normal on GPU.

E. Fusion of Depth and Normal Maps

We uses a simple depth map fusion method proposed by [9] to efficiently fuse all depth and normal maps, which is based on cross-view consistency. According to Eq. 4 and Eq. 5, a depth estimate is filtered out if the consistent depth from all neighboring depth images are less than 2, otherwise, the positions and normals of 3D points from consistent depth estimates are averaged. The fused point clouds are the final reconstruction results. We select the fusion method of [9] based on following considerations: (1) it is GPU-friendly and efficient for the large-scale reconstruction, which is suitable for our aim in this work; (2) compared with other sophisticated fusion methods, [9] is a very simple method that only consists of cross-view checking and points averaging, so we can ensure the performance gains of the proposed method are mainly from the claimed contributions.

IV. Experimental Evaluation

To evaluate the proposed CLD-MVS method on multi-view stereo from high-resolution images, we conducted quantitative and qualitative evaluation on recent published ETH3D high-resolution MVS benchmark [6], as well as on DTU large-scale MVS benchmark [5]. In the current implementation, coarse depth and normal estimation and refinement were implemented using C++ with OpenCV and CUDA, DCNN was trained using MatConvNet, the depth interpolation was implemented using Matlab. The experiments were conducted on a computer equipped with a i7 CPU and a GTX 1070 GPU.

Unless otherwise stated, for experiments in ETH3D high-resolution MVS benchmark, the matching window diameter m, weight parameter γ and best K_b views used in PatchMatch stereo are fixed to 25, 20 and 3, respectively, and $\epsilon_1 = 1.0$, $\epsilon_2 = \pi/6$, $\lambda_1 = 4$, $\lambda_2 = 0.02$, $\mu = 200$, $\nu = 200$, $\tau_n = 16\pi/33$, $\kappa = 0.001$ and $\alpha = 2$. For all datasets used in DTU MVS benchmark, we set $m = 15$, $\gamma = 5$, $K_b = 3$, $\epsilon_1 = 0.5$, $\epsilon_2 = \pi/6$, $\lambda_1 = 4$, $\lambda_2 = 0.1$, $\mu = 200$, $\nu = 200$, $\tau_n = 16\pi/33$, $\kappa = 0.001$, and $\alpha = 2$, respectively.

A. Evaluation on ETH3D High-resolution MVS Benchmark

ETH3D high-resolution MVS benchmark consists of calibrated high-resolution photos of indoor and outdoor challenge scenes with the image size of 6,048× 4,032. There are training and testing branches. Training branch includes 13 datasets with ground-truth, and testing branch includes 12 datasets and ground-truth is not publicly available. This benchmark proposes to use a novel metric in terms of F_1 score which is the harmonic mean of accuracy and completeness of reconstructions. We downsample the undistorted images to $3,200 \times 2,130$ and evaluate our reconstruction results on both testing and training branches. Quantitative evaluation results for ours and competing methods[1] are shown in Table I. These results are public in the website of the benchmark. As shown in Table I, CLD-MVS achieves best results in terms of F_1 score for all scenes except the outdoor scene in training branch where our method ranked the second. The visual comparison among different methods are shown in Fig. 7. CLD-MVS achieves much higher completeness while preserving fine-scale geometric details for both indoor and outdoor scenes. The DCNN for spatial confidence estimation is trained using samples from training branch.

We further quantitatively evaluate the contribution of initialization, interpolation and refinement in CLD-MVS framework. To speed up experimental validation, we downsample undistorted images to approximately one fourth of original resolution (i.e., $1,600 \times 1,064$). The evaluation results in terms of F_1 score for different evaluation thresholds are listed in Table II. The *baseline PM* is a PatchMatch stereo method without the proposed normal refinement. The *normal-aware PM* is a PatchMatch stereo method with our normal refinement which has been used as initial reconstruction of depth and normal maps for the subsequent stages. The *interp w/o dynamic*, *interp w/o crossview* and *interp w/o spatial* are three variants of the proposed interpolation method by removing dynamic guidance, spatial confidence and cross-view confidence, respectively. The *interp* is the proposed confidence-driven and boundary-aware interpolation method and the *full pipeline* means refinement after interpolation, which represents full stages of our MVS pipeline. Note that the interpolations

[1]In Table I, we only show results whose publications are available. The full lists can be found in website of the benchmark

TABLE I: The quantitative evaluation on ETH3D high-resolution benchmark. Competing methods include MVE [27], PMVS [8], COLMAP [10], Gipuma [9], LTVRE [2], CMPMVS [43], ACMH [33], OpenMVS [46], ACMM [48] and TAPAMVS [49]. Reconstructions are measured in terms of F_1 score(%) and running time(minutes). The evaluation threshold is 0.02 which is default setting for the evaluation.

Indicator	MVE	PMVS	COLMAP	Gipuma	LTVRE	CMPMVS	ACMH	OpenMVS	ACMM	TAPAMVS	Ours
Training											
indoor	23.27	43.30	66.76	35.80	62.64	62.52	70.00	76.82	78.13	80.05	**81.23**
outdoor	17.20	49.28	68.70	23.27	60.87	62.46	71.54	75.37	**79.71**	74.94	77.16
all	20.47	46.06	67.66	36.38	61.82	62.49	70.71	76.15	78.86	77.69	**79.35**
time	221.31	13.94	44.84	**9.80**	827.60	34.73	15.80	31.60	17.46	46.07	116.45
Testing											
indoor	25.89	40.28	70.41	41.86	74.54	68.16	73.93	78.33	79.84	77.94	**81.65**
outdoor	43.81	55.82	80.81	55.16	81.41	76.28	81.77	84.09	83.58	82.79	**84.29**
all	30.37	44.16	73.01	41.86	76.25	70.19	75.89	79.77	80.78	79.15	**82.31**
time	175.84	15.95	27.64	**11.50**	389.63	33.05	16.13	37.72	19.42	56.25	122.30

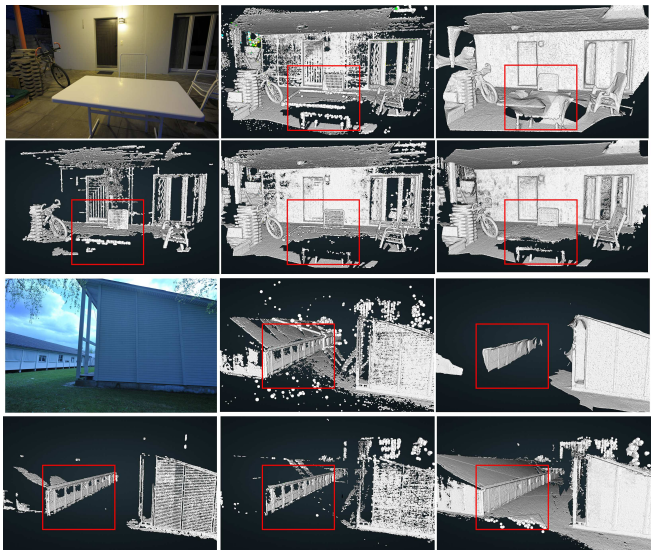


Fig. 7: Qualitative comparison on *terrace 2* and *meadow* of ETH3D high-resolution benchmark. For each dataset, from left to right, first row shows input image, results by [10], [43], respectively, and second row shows results by [8], [2] and our method, respectively. The point clouds are shaded using EyeDome lighting.

are conducted on regular grids with the grid size of 4 and the refinement is conducted pixel-wisely. According to Table II, each proposed contribution is helpful for improving reconstruction quality. In particular, our interpolation method can even improve reconstruction scores of initial reconstruction on a one fourth of input resolution. If we conduct interpolation on smaller grid size (e.g. 2), the reconstruction score can be further improved (e.g., F_1 scores are 59.14 74.61 85.07 89.46 92.67 95.88 for grid size being 2). Since the main aim of interpolation is to provide a plausible prior for refinement, it is not necessary to conduct interpolation on denser grids. Besides improving reconstruction quality, combination of spatial and cross-view confidence can improve robustness of interpolation, as shown in Fig. 8. For some parts of scenes, they may be only observed by sparse views, the cross-view confidence is not reliable since it can either introduce noise or remove true surface (e.g., in the middle left and middle right of Fig. 8). By combining both spatial and cross-view confidence, our method can recover more plausible depth map (in the right of Fig. 8).

We further visually illustrate the estimated normal map in initialization, interpolation and refinement stages of CLD-MVS on *pipes*, *office* and *kicker* in Fig. 11, respectively, as well as fused point clouds in Fig. 12. The resolution of input images is $3,200 \times 2,130$. We can clearly see that the depth maps and normal maps by coarse estimation stage contain a large amount of noise due to textureless parts dominated in these scenes, and the resultant fused point clouds contain lots of holes. The results by the proposed interpolation can fill holes while diminishing noise. However, some regions recovered via interpolation are not cross-view consistent and fine-scale geometric details are not kept. Finally, refinement stage improves accuracy and cross-view consistency, and results in both of highest completeness and detailed reconstruction. The quantitative evaluation is consistent to visual comparison. The results of *pipes* in terms of F_1 score (threshold 0.02) in three stages are 61.62, 66.21 and 73.21, respectively. The results of *office* in terms of F_1 score are 57.05, 67.02 and 72.96, respectively. The results of *kicker* in terms of F_1 score are 69.48, 71.74 and 80.07, respectively. The results demonstrate that both the proposed interpolation and pixel-wise refinement stages make contributions to high-quality reconstruction.

We also report the running time of the proposed method on ETH3D benchmark, as shown in Table I. Note that, different methods report running time based on their own platforms. For example, the platform for running COLMAP is Intel Xeon CPU E5-2697 and GTX 1080 GPU, and the platform for running LTVRE is a cluster with 96 Intel Xeon Cores. There is always a tradeoff between computational complexity and reconstruction quality. As shown in Table II, when input images are downsampled to $1,600 \times 1,064$, our F_1 scores are still better than most of competitors, where the running time decrease by a factor of four times. By considering that our method achieves much better reconstruction results than competitors, the computation complexity is reasonable. Moreover, there are some rooms for further improving efficiency, e.g., by replacing unoptimized Matlab code with C++ on GPU.

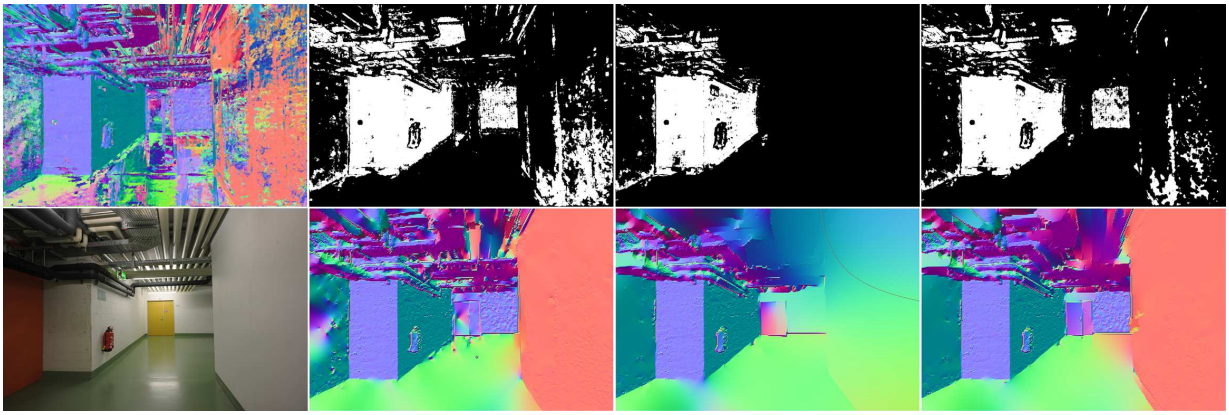


Fig. 8: Illustration of different confidence measures for depth interpolation. First row, from left to right: noisy depth map, GCPs by selecting pixels that are visible at least in one neighboring view and two neighboring views, respectively, and GCPs by combining both cross-view and spatial confidence. Second row, from left to right: the reference image, the corresponding interpolation results by using GCPs in the first row. To aid visualization, the results are visualized in normal map. Combining both of spatial and cross-view confidence shows best results.

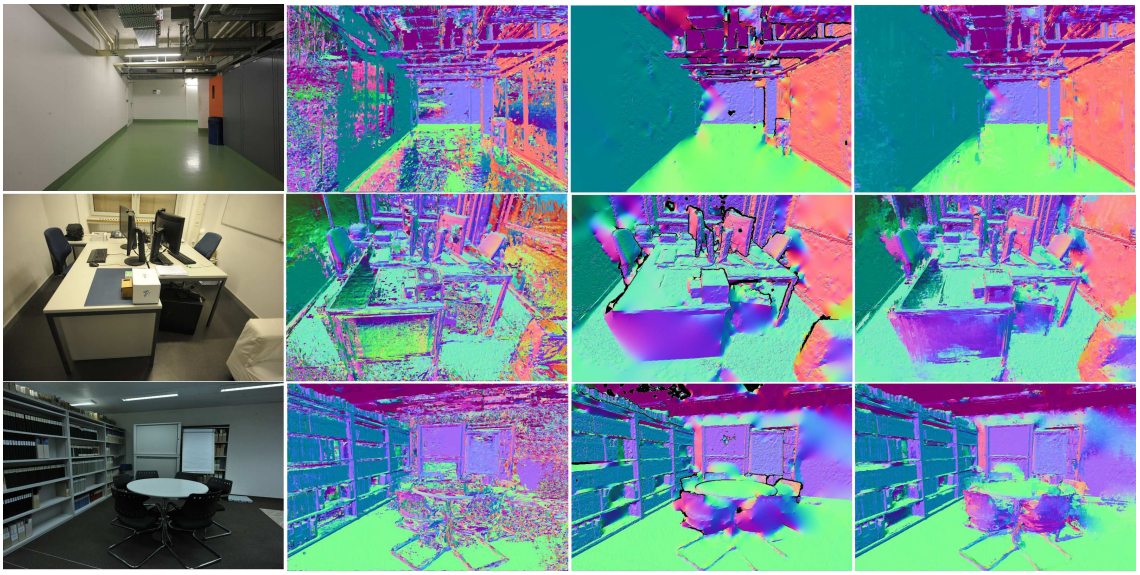


Fig. 11: The estimated normal maps by each stage of the proposed method on one reference view of *pipes*, *office* and *kicker* of ETH3D benchmark. From left to right, results by coarse estimation, confidence-driven and boundary-aware interpolation, and refinement, respectively.

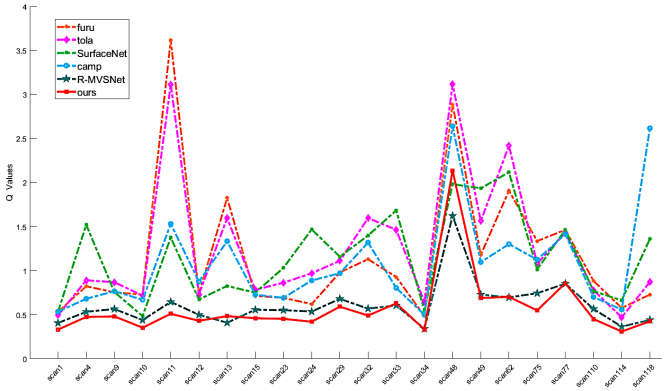


Fig. 9: Quantitative evaluation on *testing set* of DTU benchmark in terms of overall quality Q. Compared to the competing methods, our CLD-MVS achieves overall best performance.

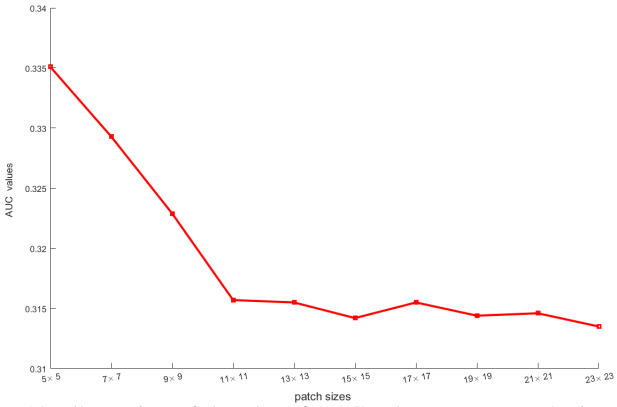


Fig. 10: Illustration of the plot of AUC values versus patch sizes on *GTAV* dataset. When patch size is gradually increased, the confidence prediction accuracy in terms of AUC is quickly improved then converged around patch size 15×15.

B. Evaluation on DTU Large-scale MVS Benchmark

We further evaluate our method on DTU large-scale MVS benchmark [5]. The DTU benchmark includes 124 different

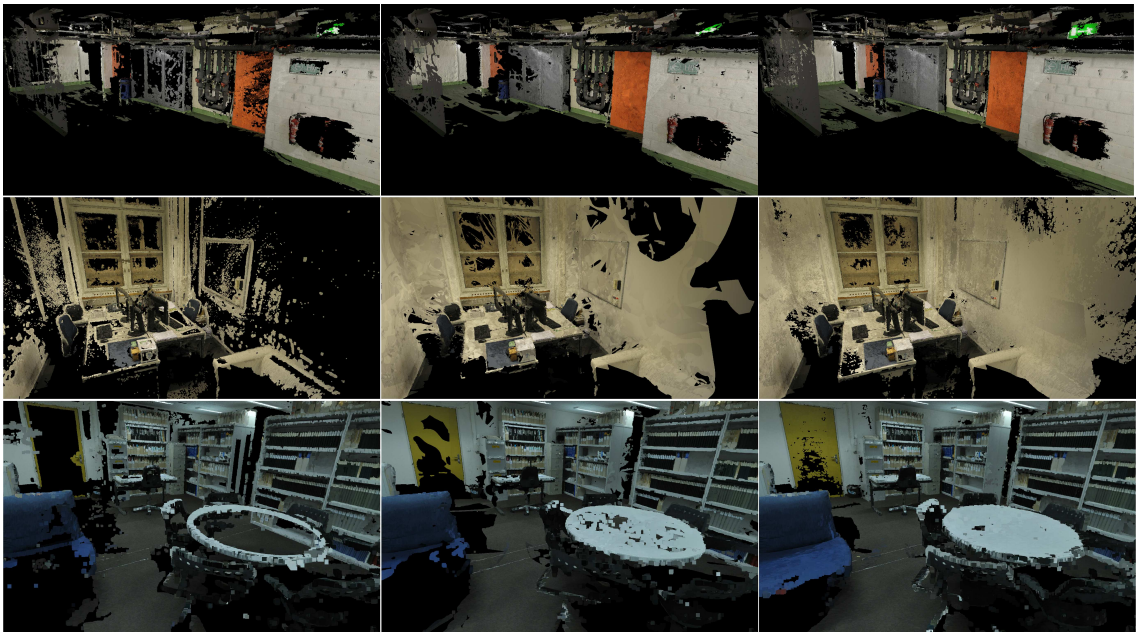


Fig. 12: Fused point clouds by each stage of the proposed method on *pipes*, *office* and *kicker* of ETH3D benchmark. From left to right, the textured point clouds by coarse estimation, confidence-driven and boundary-aware interpolation, and refinement, respectively.

TABLE II: Quantitative evaluation of the components of the proposed CLD-MVS method on ETH3D training branch in terms of F_1 score.

Methods \ thresholds	0.01	0.02	0.05	0.1	0.2	0.5
Baseline PM	52.01	69.08	81.35	87.00	91.32	95.22
Normal-aware PM	54.36	71.55	83.64	89.00	92.73	95.94
Interp w/o dynamic	53.73	71.46	83.69	88.61	92.14	95.57
Interp w/o crossview	53.13	70.28	82.33	87.35	91.14	95.14
Interp w/o spatial	54.53	72.15	84.09	88.72	92.05	95.39
Interp	54.87	72.60	84.59	89.23	92.59	95.89
Full pipeline	**60.91**	**77.46**	**88.12**	**92.23**	**94.78**	**96.99**

TABLE III: The quantitative evaluation on DTU *testing set*. The metrics include accuracy (mean of Acc. / median of Acc.), completeness (mean of Comp. / median of Comp.) and overall quality (Q score and A score). [18] and [50] are denoted by surfnet and r-mvsnet.

Indicators	furu	camp	tola	surfnet	r-mvsnet	ours
Mean Acc.	0.6124	0.8360	0.3426	0.4496	0.3835	**0.3347**
Med Acc.	0.3240	0.4908	0.2101	0.2539	0.2223	**0.2083**
Mean Comp.	0.9386	0.5545	1.1900	1.0432	0.4520	**0.4307**
Med Comp.	0.4628	0.1923	0.4921	0.2854	0.2662	**0.1851**
Q Score	1.1608	1.0795	1.2534	1.1597	0.6070	**0.5709**
A Score	0.7755	0.6952	0.7663	0.7464	0.4178	**0.3827**

indoor datasets at a resolution of $1,600 \times 1,200$ pixels, where 80 of them have been used in the evaluation. The photos of scenes were captured using a robot arm at multiple calibrated viewpoints (49 or 64) under controlled light. We quantitatively compare our method to five state-of-the-art methods [1], [7], [8], [18], [50]. Furu [8] and tola [1] are two feature-based MVS methods, camp [7] is a depth map-based MVS method, and surfaceNet [18] and R-MVSNet [50] are two deep learning-based MVS methods. The reconstruction results of first three methods were provided by the DTU benchmark, and the authors of surfaceNet published 22 reconstructed point clouds which are reported in [18] as testing set, i.e., 1, 4, 9-13, 15, 23, 24, 29, 32-34, 48, 49, 62, 75, 77, 110, 114, 118. The authors of R-MVSNet [50] also published their reconstruction results on these datasets. To conduct faithfully comparison, we evaluate competing methods and ours in these 22 datasets, named as "*testing set*". For surfaceNet, we use the results for cubes with 64^3 voxels which are reported higher accuracy and completeness in [18]. The DTU benchmark uses four metrics to evaluate reconstruction, including mean of accuracy, mean of completeness, median of accuracy and median of completeness. For each of these metrics, smaller value is better. The mean values of four metrics [2] on *testing set* are shown in Table III. Our method achieves best results in terms of all metrics. Note that there is always a tradeoff between accuracy and completeness. For example, tola [1] prefers a high-accurate reconstruction while the completeness usually lowers than other methods, and camp [7] achieves a high-complete reconstruction with lowest accuracy among these methods. To measure the overall quality of the reconstruction, [44] uses $Q = \sqrt{(\text{mean acc.})^2 + (\text{mean comp.})^2}$ for each dataset as a comprehensive metric in the spirit of F_1 score. Yao et al. [50] proposed another overall score A that calculates average of mean accuracy and mean completeness on entire *testing set*. As shown in Table III, our method ranked the highest in terms of overall quality Q and A. In Fig. 9, the metric Q is plotted for the results by competing methods and ours for each of *testing set*. One can see that our method achieves better results than all competing methods for 17 out of 22 datasets. Overall, compared with these competing methods, our method can obtain best results in terms of completeness, accuracy

[2]We followed the standard evaluation protocol provided by DTU benchmark. [18] uses a slightly different way to compute statistics according to their published code. To keep faithful comparison, we also evaluate methods according to the way used by [18] in which our method is still top ranked for 5 out 6 metrics. The results can be found in the supplementary materials.

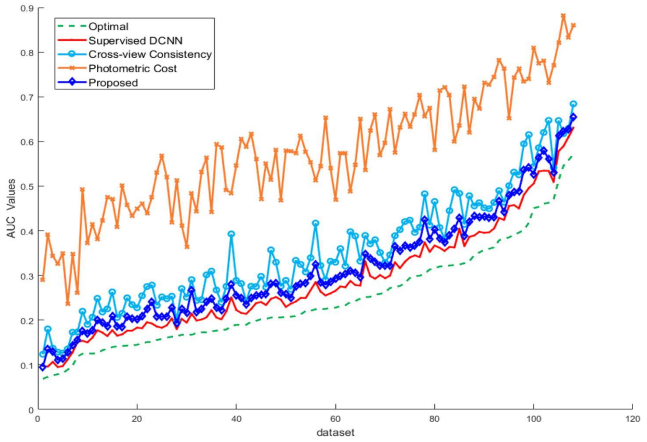


Fig. 13: Average AUC values for testing sequences of *GTAV* obtained by photometric cost, cross-view consistency, supervised DCNN, and our self-supervised DCNN, respectively. The optimal AUC values are also plotted in green dashed curve. Testing set has been sorted in order of increasing optimal AUC values for better visualization.

and overall quality. The samples for training the confidence prediction network are extracted from 10 training set of DTU via the proposed self-supervised method.

Several visual comparisons of reconstructions by different methods on representative datasets are shown in Fig. 16 and Fig. 17. *Scan11* contains a large amount of low-textured parts, and *scan77* contains both low-textured and highly-specular regions. Obviously, the proposed method can achieve significant better visual quality. Besides, we give a qualitative and quantitative evaluation on *scan102*. This dataset captures a pot plant including lots of thin-structures and heavy self-occlusions which pose a challenging to MVS methods. As shown in Fig. 18, our method shows higher geometric completeness while visually more pleasing. The quantitative evaluation results in terms of overall quality (i.e., Q score) for furu, camp, tola and ours are 1.2622, 1.0541, 1.8626, and 0.5522, respectively.

C. Evaluation of Confidence Prediction

We quantitatively validate the self-supervised confidence learning method on *GTAV* a synthetic multi-view dataset provided by [45]. This dataset consists of 120 image sequences, and each sequence includes 100 video frames with the corresponding ground-truth depth maps. In our experiments, we uniformly sample 25 images for each sequence and run depth estimation algorithm introduced in Section III to estimate depth maps of sampled 120 sequences (3000 depth maps in total). The first 12 sequences are used as training set (300 depth maps and more than 5 million training samples) and others are used as testing set. Four different confidence measures are used in evaluation: photometric cost, cross-view consistency, DCNN trained on ground-truth and DCNN trained based on the self-supervised learning method proposed in Section III-B. A depth estimation z_p is regarded as incorrect if $\left|\left(z_p - z_{gt}\right)/z_{gt}\right| < 0.01$. Following literature on learning stereo confidence [34], [36], we used area under the curve (AUC) to quantify the ability of a confidence measure to predict correct depth estimates. We first compute receiver operating characteristic (ROC) curves as a function of the depth map density as follows: all depth estimates are ranked in decreasing order according of confidence measure, and pixels are gradually selected from top to bottom to record error rate. Ties are handled by including pixels with same confidence in a sample. The AUC of a confidence measure is the area integral under its ROC curve. According to [34], the optimal AUC is computed by $\epsilon + (1-\epsilon)\ln(1-\epsilon)$, where ϵ is the error rate for a depth map at full density. We plot mean of AUC of each testing sequence for different confidence measures and optimal AUC, respectively, as shown in Fig. 13. We can see that the DCNN trained from ground-truth shows best performance, and the prediction ability of self-supervised confidence measure is the second best on each sequence and always better than cross-view consistency and photometric cost. The average AUC of all testing sequences by using optimal AUC, supervised DCNN, self-supervised DCNN, cross-view consistency and photometric cost, are: 0.2459, 0.2869, 0.3139, 0.3501 and 0.5727, respectively. In particular, the average AUC increases to 0.3315 if TSDFs are not used to generate pseudo-positive samples in Eq. 10, which demonstrates that TSDFs are helpful for improve quality of training samples. The visual comparison of GCPs detection via confidence prediction ($f_s > 0.6$) between the supervised and the self-supervised method is shown in Fig. 14. Though the self-supervised method shows lower prediction accuracy than the supervised method, the results are still reasonable.

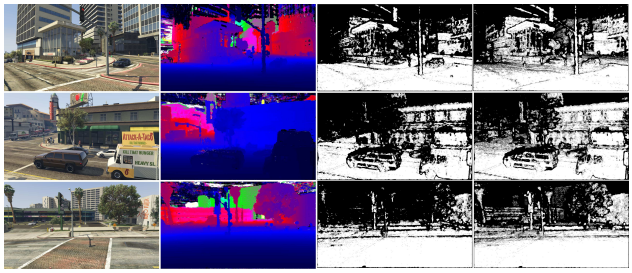


Fig. 14: Confidence prediction on *GTAV* dataset. From left to right, input images, depth maps, GCPs by the self-supervised method, GCPs by the supervised method.

Fig. 10 illustrates the plot of confidence prediction performance along with the increase of patch size. When patch size is gradually increased, the confidence prediction accuracy in terms of AUC is quickly improved then converged around patch size 15×15. Note that the time for inference also gradually increases when the patch size is increased. Our DCNN uses 15×15 patches as inputs to achieve good trade-off between confidence prediction accuracy and efficiency.

Fig. 15 shows GCPs detection results on DTU benchmark. The self-supervised method trains DCNN based on extracted samples from 10 training set of DTU benchmark. It clearly shows that GCPs detected by integrating both of spatial and cross-view confidence based on Eq. 11 are more complete while having less noise.

V. Conclusion

In this paper, we proposed a confidence-based large-scale dense multi-view stereo method, namely CLD-MVS, which

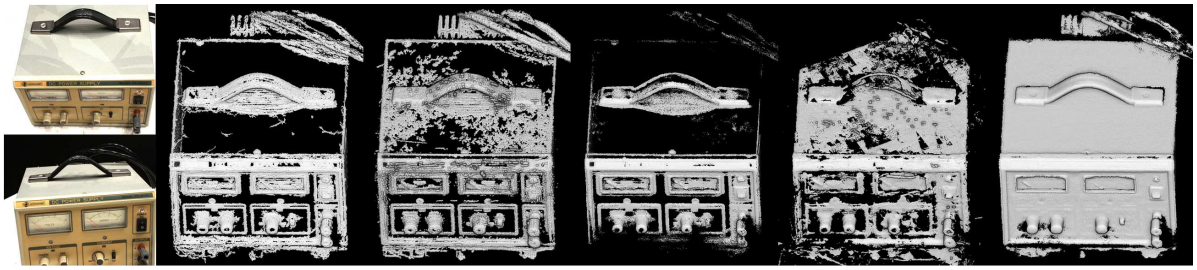


Fig. 16: Reconstruction results of *scan11* of DTU benchmark. From left to right, cropped input images, and results by Furukawa et al. [8], Campbell et al. [7], Tola et al. [1], Ji et al. [18] and our method, respectively. The point clouds are shaded using EyeDome lighting.

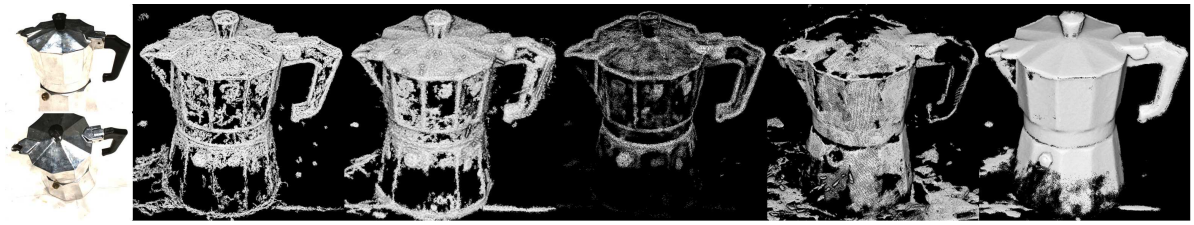


Fig. 17: Reconstruction results of *scan77* of DTU benchmark. From left to right, cropped input images, and results by Furukawa et al. [8], Campbell et al. [7], Tola et al. [1], Ji et al. [18] and our method, respectively. The point clouds are shaded using EyeDome lighting.

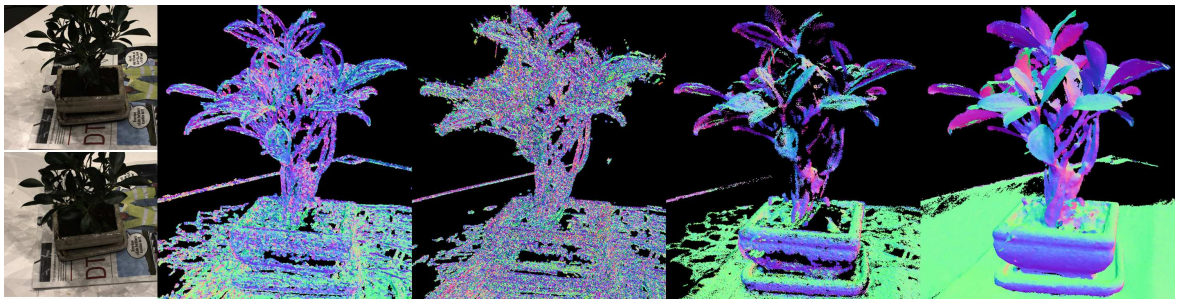


Fig. 18: Reconstruction results of *scan102* of DTU benchmark. From left to right, cropped input image, result by Furukawa, et al. [8], Campbell et al. [7], Tola et al. [1] and our method, respectively. The point clouds are shaded using normal map.

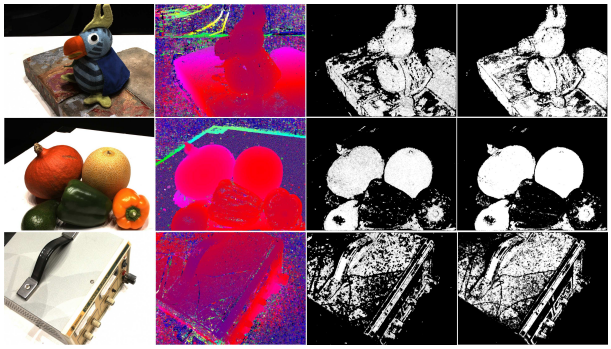


Fig. 15: GCPs detection on DTU. From left to right, input images, depth maps, GCPs by the self-supervised method ($f_s > 0.6$), GCPs by utilizing both of spatial and cross-view confidence based on Eq.11.

consists of four key stages: coarse depth and normal estimation, confidence prediction, confidence-driven interpolation and refinement. We leveraged the normal-aware PatchMatch stereo to improve the accuracy of normal map in the coarse estimation stage, and presented a self-supervised learning approach to predict spatial confidence for multi-view depth maps while combining spatial and cross-view confidence to detect high-quality GCPs. We then presented a confidence-driven and boundary-aware interpolator to construct high-quality depth and normal map priors, and refined pixel-wisely the geometric details to diminish the artifacts caused by interpolation errors. Quantitative and qualitative evaluations on large-scale MVS benchmarks demonstrated that our CLD-MVS achieves more promising reconstruction results than the state-of-the-art MVS methods, and the proposed self-supervised learning method is competitive for confidence prediction in real world scenes without ground-truth. In the future, we will investigate how to combine semantic segmentation cues with the proposed method for separating different objects from background (e.g., sky) and interpolating large missing regions.

Acknowledgement

The authors would like to thank the associate editor and the anonymous reviewers for their constructive suggestions. They would also like to thank the support of NVIDIA Corporation with the donation of the TITAN X GPU for the early stage of this research.

References

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

[2] A. Kuhn, H. Hirschmüller, D. Scharstein and H. Mayer, "A TV Prior for High-Quality Scalable Multi-View Stereo Reconstruction," *Int. J. Comput. Vis.*,vol. 124, no. 1, pp. 2-17, 2017.

[3] P. Hedman, S. Alsisan, R. Szeliski and J. Kopf, "Casual 3D Photography," *ACM Trans. on Graphics*, vol. 36, no. 6, pp. 234:1-234:15, 2017.

[4] S. Seitz, B. Curless, J. Diebel, D. Scharstein and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proc. CVPR*, pp. 519-526, 2006.

[5] R. Jenseny, A. Dahly, G. Vogiatzisz, E. Tolax and H. Aans, "Large Scale Multi-view Stereopsis Evaluation," in *Proc. CVPR*, pp. 1-8, 2014.

[6] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys and A. Geiger, "A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos," in *CVPR*, pp. 1-8, 2017.

[7] N. Campbell, G. Vogiatzis, C. Hernandez and R. Cipolla, "Using multiple hypotheses to improve depth-maps for multi-view stereo," in *Proc. ECCV*, pp. 1-14, 2008.

[8] Y. Furukawa and J. Ponce, "Accurate, Dense, and Robust Multi-View Stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362-1376, 2010.

[9] S. Galliani, K. Lasinger and K. Schindler, "Massively Parallel Multiview Stereopsis by Surface Normal Diffusion," in *ICCV*, pp. 873-881, 2015.

[10] J. L. Schönberger, E. Zheng, M. Pollefeys and J. M. Frahm, "Pixelwise View Selection for Unstructured Multi-View Stereo," in *Proc. ECCV*, pp. 1-17, 2016.

[11] C. Zach, T. Pock and H. Bischof, "A Globally Optimal Algorithm for Robust TV-L1 Range Image Integration," in *Proc. ICCV*, pp. 1-8, 2007.

[12] G. Graber, J. Balzer, S. Soatto and T. Pock, "Efficient minimal-surface regularization of perspective depth maps in variational stereo," in *Proc. CVPR*, pp. 511-520, 2015.

[13] I. Kostrikov, E. Horbert and B. Leibe, "Probabilistic Labeling Cost for High-Accuracy Multi-View Reconstruction," in *CVPR*, pp. 1-8, 2014.

[14] M. Jancosek and T. Pajdla, "Segmentation based Multi-View Stereo," in *Proc. CVWW*, pp. 1-6, 2009.

[15] S. Y. Bao, M. Chandraker, Y. Lin and S. Savarese, "Dense Object Reconstruction with Semantic Priors," in *CVPR*, pp. 1264-1271, 2016.

[16] C. Hane, N. Savinov and M. Pollefeys, "Class Specific 3D Object Shape Priors Using Surface Normals," in *Proc. CVPR*,pp. 652-659, 2014.

[17] C. Hane, C. Zach, A. Cohen and M. Pollefeys, "Dense Semantic 3D Reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*,vol. 39, no. 9, pp. 1730-1743, 2017.

[18] M. Ji, J. Gall, H. Zheng, Y. Liu and L. Fang, " SurfaceNet: an End-to-end 3D Neural Network for Multiview Stereopsis," in *Proc. ICCV*, pp. 2307-2315, 2017.

[19] Q. Shan, B. Curless, Y. Furukawa, C. Hernandez and S. M. Seitz, "Occluding Contours for Multi-View Stereo," in *Proc. CVPR*, 2014.

[20] Z. Farbman, R. Fattal, D. Lischinski and R. Szeliski, "Edge-Preserving Decompositions for Multi-Scale Tone and Detail Manipulation," *ACM Trans. on Graphics*, vol. 27, no. 3, pp. 1-10, 2008.

[21] A. Bódis-Szomor, H. Riemenschneider and L. V. Gool, "Superpixel Meshes for Fast Edge-Preserving Surface Reconstruction," in *Proc. CVPR*, pp. 2011-2020, 2015.

[22] B. Ham, M. Cho and J. Ponce, "Robust Image Filtering Using Joint Static and Dynamic Guidance," in *Proc. ICCV*, pp. 4823-4831, 2015.

[23] A. Geiger, M. Roser and R. Urtasun, "Efficient Large-Scale Stereo Matching," in *Proc. ACCV*, pp. 1-14, 2010.

[24] Z. Li, K. Wang, W. Zuo, D. Meng and L. Zhang, "Detail-Preserving and Content-Aware Variational Multi-View Stereo Reconstruction," *IEEE Trans. on Image Processing*, vol. 25, no. 2, pp. 864-877, 2016.

[25] M. Goesele, N. Snavely, B. Curless, H. Hoppe and S. M. Seitz, "Multi-View Stereo for Community Photo Collections," in *ICCV*, pp. 1-8, 2007.

[26] H. H. Vu, P. Labatut, J. P. Pons and R. Keriven, "High Accuracy and Visibility-Consistent Dense Multiview Stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 889-901, 2011.

[27] S. Fuhrmann, F. Langguth and M. Goesele, "MVE - A Multi-View Reconstruction Environment," in *EUROGRAPHICS Workshops on Graphics and Cultural Heritage*, pp. 1-8, 2014.

[28] S. M. Seitz and C. R. Dyer, "Photorealistic scene reconstruction by voxel coloring," in *Proc. CVPR*, 1997.

[29] D. Xu, Q. Duan, J. Zheng, J. Cai, and T.-J. Cham, "Recovering surface details under general unknown illumination using shading and coarse multi-view stereo," in *Proc. CVPR*, pp. 1526-1533, 2014.

[30] K. Kim, A. Torii and M. Okutomi, "Multi-view Inverse Rendering Under Arbitrary Illumination and Albedo," in *Proc. ECCV*, pp. 750-767, 2016.

[31] K. Wolff, C. Kim, H. Zimmer, C. Schroers, M. Botsch, O. Sorkine-Hornung and A. Sorkine-Hornung, "Point Cloud Noise and Outlier Removal for Image-Based 3D Reconstruction," in *Proc. 3DV*, 2016.

[32] C. Bailer, M. Finckh and H. P. A. Lensch, "Scale Robust Multi View Stereo," in *Proc. ECCV*, pp. 1-14, 2012.

[33] Q. Xu and W. Tao, "Multi-View Stereo with Asymmetric Checkerboard Propagation and Multi-Hypothesis Joint View Selection," in *arXiv:1805.07920v1*, 2018.

[34] A. Spyropoulos, N. Komodakis and P. Mordohai, " Learning to Detect Ground Control Points for Improving the Accuracy of Stereo Matching," in *Proc. CVPR*, pp. 1621-1628, 2014.

[35] M. Park and K. Yoon, "Leveraging Stereo Matching with Learning-based Confidence Measures," in *Proc. CVPR*, pp. 101-109, 2015.

[36] M. Poggi and S. Mattoccia, "Learning from scratch a confidence measure," in *Proc. BMVC*, pp. 1-13, 2016.

[37] S. Kim, D. Min, S. Kim and K. Sohn, "Unified Confidence Estimation Networks for Robust Stereo Matching," *IEEE Trans. on Image Processing*, vol. 28, no. 3, pp. 1299-1313, 2019.

[38] C. Mostegel, M. Rumpler, F. Fraundorfer and Horst Bischof, "Using Self-Contradiction to Learn Confidence Measures in Stereo Vision," in *Proc. CVPR*, pp. 4067-4076, 2016.

[39] F. Tosi, M. Poggi, A. Tonioni, L. D. Stefano and S. Mattoccia, "Learning confidence measures in the wild," in *Proc. BMVC*, pp. 1-13, 2017.

[40] X. Hu and P. Mordohai, "Least Commitment, Viewpoint-Based, Multi-view Stereo," in *Proc. 3DIMPVT*, pp. 531-538, 2012.

[41] Y. Heo, K. Lee and S. Lee, "Robust stereo matching using adaptive normalized cross-correlation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 807-822, 2011.

[42] M. Bleyer, C. Rhemann and C. Rother, "PatchMatch Stereo - Stereo Matching with Slanted Support Windows," in *BMVC*, pp. 1-11, 2011.

[43] M. Jancosek and T. Pajdla, "Multi-View Reconstruction Preserving Weakly-Supported Surfaces," in *Proc. CVPR*, pp. 3121-3128, 2011.

[44] S. Galliani and K. Schindler, " Just Look at the Image: Viewpoint-Specific Surface Normal Prediction for Improved Multi-View Reconstruction," in *Proc. CVPR*, pp. 5479-5487, 2016.

[45] P. Huang, K. Matzen, J. Kopf, N. Ahuja and J. Huang, " DeepMVS: Learning Multi-view Stereopsis," in *Proc. CVPR*, pp. 2821-2830, 2018.

[46] *OpenMVS*. [Online]. Available: http://cdcseacave.github.io/openMVS/, accessed data: 2019.

[47] M. Poggi, F. Tosi and S. Mattoccia, " Quantitative evaluation of confidence measures in a machine learning world," in *Proc. ICCV*, pp. 5228-5237, 2017.

[48] Q. Xu and W. Tao, " Multi-Scale Geometric Consistency Guided Multi-View Stereo," in *Proc. CVPR*, 2019.

[49] A. Romanoni and M. Matteucci, " TAPA-MVS: Textureless-Aware PAtchMatch Multi-View Stereo," in *arXiv:1903.10929v1*, 2019.

[50] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang and L. Quan, "Recurrent MVSNet for High-resolution Multi-view Stereo Depth Inference," in *Proc. CVPR*, 2019.