CNLPA-MVS: Coarse-Hypotheses Guided Non-Local PAtchMatch Multi-View Stereo. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY

CNLPA-MVS: Coarse-Hypotheses Guided Non-Local PAtchMatch Multi-View Stereo

Abstract In multi-view stereo, unreliable matching in low-textured regions has a negative impact on the completeness of reconstructed models. Since the photometric consistency of low-textured regions is not discriminative under a local window, non-local information provided by the Markov Random Field (MRF) model can alleviate the matching ambiguity but limited in continuous space with high computational complexity. Owing to its sampling and propagation strategy, PatchMatch multi-view stereo methods have advantages in terms of optimizing continuous labeling problem. In this paper, we propose a novel method to address this problem, namely the Coarse-hypotheses guided Non-Local PAtchMatch Multi-View Stereo (CNLPA-MVS), which takes the advantages of both MRF-based non-local methods and PatchMatch multi-view stereo and compensates for their defects mutually. First, we combined dynamic programing (DP) and sequential propagation along scanlines in parallel to perform CNLPA-MVS, thereby obtaining the optimal depth and normal hypotheses. Second, we introduced coarse inence within a universal window provided by winner-takes-all to eliminate the stripe artifacts caused by DP and improve completeness. Third, we added a local consistency strategy based on the hypotheses of similar color pixels sharing approximate values into CNLPA-MVS for further improving completeness. CNLPA-MVS was validated on public benchmarks and achieved state-of-the-art performance with high completeness.

Keywords 3D reconstruction, multi-view stereo, PatchMatch, dynamic programming

1 Introduction

In recent years, multi-view stereo (MVS) has become a hot research topic that is widely used in image classification [1], SLAM [2], and image-based rendering [3], etc. The goal of MVS is to obtain a dense 3D presentation from a set of calibrated scene images. In the last decades, the high-resolution datasets [4, 5, 6] that are publically available have promoted MVS research and resulted in the development of numerous MVS methods. However, reconstructing a high-quality and a complete 3D model is still a challenge due to inaccurate and missed depth inference in low-textured regions.

The main difficulty in depth inference in low-textured regions is the matching ambiguity due to similar colors. Although learning-based methods [7, 8, 9] address the issue by introducing semantic information, the performance is highly dependent on the training datasets, and accuracy improvement is instable. Traditional non-local methods [10, 11] utilize global information. They regard the depth-map estimation as a pixel-labeling problem with the Markov Random Field (MRF) model, which can reduce matching ambiguity. However, discretizing depth values as a label set is unsuitable for reconstructing slanted surfaces, and the optimization in continuous space is complicated. Recently, PatchMatch MVS approaches have gained attention because of the high-accuracy reconstructed models generated and high-efficiency solutions for continuous labelling problems. GPU compatible methods [12, 13, 14] with modified propagation strategies were presented to further improve the efficiency. However, the original PatchMatch Stereo method [15] with only one data term suffers from the matching uncertainty in low-textured regions. Some global PatchMatch methods [16, 17, 18] in stereo matching that incorporate global MRF solution into the PatchMatch algorithms were proposed. Although these approaches can robustly solve the optimization in continuous space to some extent, efficient processing of high-resolution image pairs still remains intractable.

In this paper, we combined the advantages of non-local methods using MRF and PatchMatch MVS algorithms to improve the completeness of dense matching in low-textured regions. Unlike the previous global PatchMatch Stereo methods, our solution integrates sequential propaga-

Regular Paper

tion along scanlines in parallel into non-local optimization. We propose a coarse-hypotheses guided non-local PatchMatch MVS (CNLPA-MVS) method to improve the completeness of 3D reconstruction. We carefully analyzed the common point in sequential propagation (Figure 1(a)) and dynamic programming (DP) (Figure 1(b)) and found that they both sweep and update the status of each pixel along the scanline. Hence, based on this observation, we elaborately combined the two methods (sequential propagation and dynamic programming) to obtain a parallel non-local PatchMatch MVS framework (Figure 1(c)). This framework contributes to the completeness of depth estimation in low-textured regions. Moreover, we propose a coarse-hypotheses guidance strategy that integrates the winner-takes-all (WTA) results using coarser scale images into the hypotheses candidates for the DP process (Figure 1(d)). Within a universal window, the coarser scale images that contain more texture information can improve the ability of matching in low-textured regions, and the WTA estimation results help eliminate stripe artifacts caused by the simple DP algorithm. Similar to the previous work [19], we introduce local consistency to assume that neighboring pixels with similar colors share approximate depth values and adopt it as a smooth constraint. Rather than adopting local consistency constraint directly, we only consider the previous neighborhood along the sweep direction with the latest state instead of the 4-neighborhoods considered in [19], which neglects the state of the pixels.

Experiments show the great power of our method to reconstruct low-textured regions with strong occlusions, which are common in indoor and outdoor scenarios. The main contributions of this paper are summarized as follows:

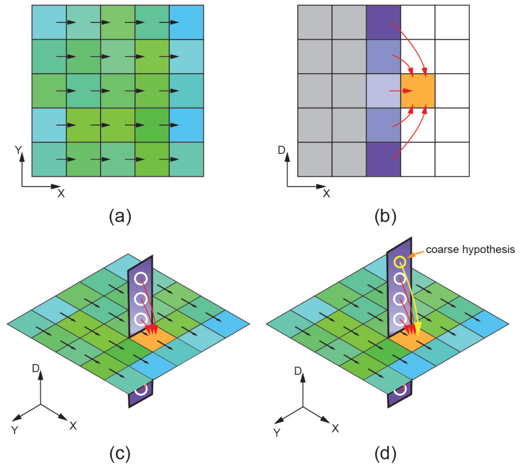


Fig.1. CNLPA-MVS derivation: (a) Sequential propagation. (b) Dynamic programming in discrete depth space. (c) Non-local PatchMatch MVS method in continuous depth space. (d) CNLPA-MVS. Note that the above images only show methods in the depth space. The black arrows are current propagation direction. The red arrows mean the penalty computed between the previous pixel with different depths and the processing pixel (the orange pixel). The white circles in (c) and (d) show the hypotheses representations in continuous depth space, and the yellow circle show the introduced depth hypothesis at a coarser scale using WTA.

- A novel parallel non-local PatchMatch MVS framework is proposed by combining dynamic programming and PatchMatch sequential propagation to alleviate the matching ambiguity in low-textured regions.
- A coarse-hypotheses guidance strategy is presented to propagate reliable estimations in low-textured regions obtained at a coarser scale to a finer scale and to reduce stripe artifacts caused by simple DP.
- A new local consistency constraint is introduced to the global PatchMatch energy function utilizing the latest status of the adjacent pixel with similar colors along the sweep direction.

2 Related Work

On the basis of the output scene representation, MVS algorithms can be categorized into four types: (1) depth-map based methods [10, 12, 20], (2) patch based methods [21, 22], (3) voxel based methods [23, 24], and (4) surface evolution based methods [25]. More details can be found in literature [26, 27]. The proposed algorithm belongs to the first type of methods, in which per-view depth map is estimated and fused into a whole point cloud. In this section, we review the most relevant methods that are widely used in multi-view stereo and two-view stereo applications.

PatchMatch Multi-View Stereo. PatchMatch algorithm was first proposed by Barnes et al. [28]. It randomly initializes each label of pixels, and then it propagates the good estimation to the neighboring areas. Consequently, the best matches spread over the whole image. Based on this kernel idea, PatchMatch MVS algorithms were proposed to rapidly estimate approximate depth and normal hypotheses of each image. For efficiency, various propagation schemes tailored for parallel computation were presented, including 1) sequential propagation [5, 20] that traverses pixels along scanlines in parallel; and 2) diffusion-like propagation [13, 14] which process half of the pixels in an image at a time. For accuracy, Schönberger et al. [12] and Zheng et al. [20] jointly modeled pixelwise view selection and depth inference into a Hidden Markov Chain. However, the above methods suffer from the matching uncertainty in low-textured regions, resulting in reconstruction incompleteness due to the lack of a smoothness term in the original PatchMatch Stereo algorithm [15]. To overcome this problem, Wei et al. [29] and Xu et al. [14] constructed a multi-scale framework to increase completeness of reconstruction in low-textured regions, but this framework may reduce the accuracy of high-textured regions. Other methods address the problem by utilizing planar priors, such as super pixels [24] and coarse triangulations [31]. These solutions assume that the 3D points of pixels within a small area are in the same plane. However, they cannot be extended to the curved surfaces. To solve the overall matching uncertainty problem, Liao et al. [19] proposed the concept of local consistency, which assumes that the adjacent pixels with similar colors share approximate depth and normal hypotheses. However, their work ignored the status of 4-neighborhoods and hindered the propagation of the latest depth and normal hypotheses. Hence, non-optimal results were obtained. Xu et al. [32] proposed a new pixelwise window-size selection strategy to decrease the matching uncertainty of the low-textured regions. However, the method neglected the fact that the pixelwise view selection is important for accuracy. In contrast, our proposed algorithm utilizes a MRF model with a smoothness term to improve the reconstruction completeness while maintaining the high fidelity of the 3D reconstruction.

Non-local Methods Using MRF. Since matching with only one data term in low-textured regions leads to ambiguity, many non-local methods with a smoothness term that assume adjacent pixels share similar depth values were proposed. In binocular stereo vision, the non-local methods could use the MRF model to perform disparity estimation as a pixel-labelling problem. Since finding the best label assignment that minimizes the global energy function is NP-hard, various approximate approaches, such as graph cuts [33, 34] and belief propagation [35, 36], were proposed. For the label space, in some early approaches [37, 38], the disparity space was discretized into a finite set, and then, the best disparity for each pixel was determined. The discrete disparity space prefers front-parallel surfaces, where all pixels are assigned to the same depth value. Thus, the above methods were unsuitable for slanted or curved surfaces. Woodford et al. [11] introduced a second-order smoothness term into a global energy function over triple cliques, that was adapted to piecewise planar surfaces. However, this approach is computationally complex and hard for optimization. For large-scale scenes, the depth space with a large range further dramatically increases the optimization complexity. Campbell et al. [10] proposed a strategy for estimating true depth from multiple depth hypotheses rather than using discrete depth values as a label set for an entire image. They first selected several local optimal depths of each pixel from the photo-consistency curves of neighboring images and then utilized the MRF formulation to assign the depth of one of such local optimums to each pixel. Our method integrates the PatchMatch MVS algorithm into the global energy function, assigns different label sets to each

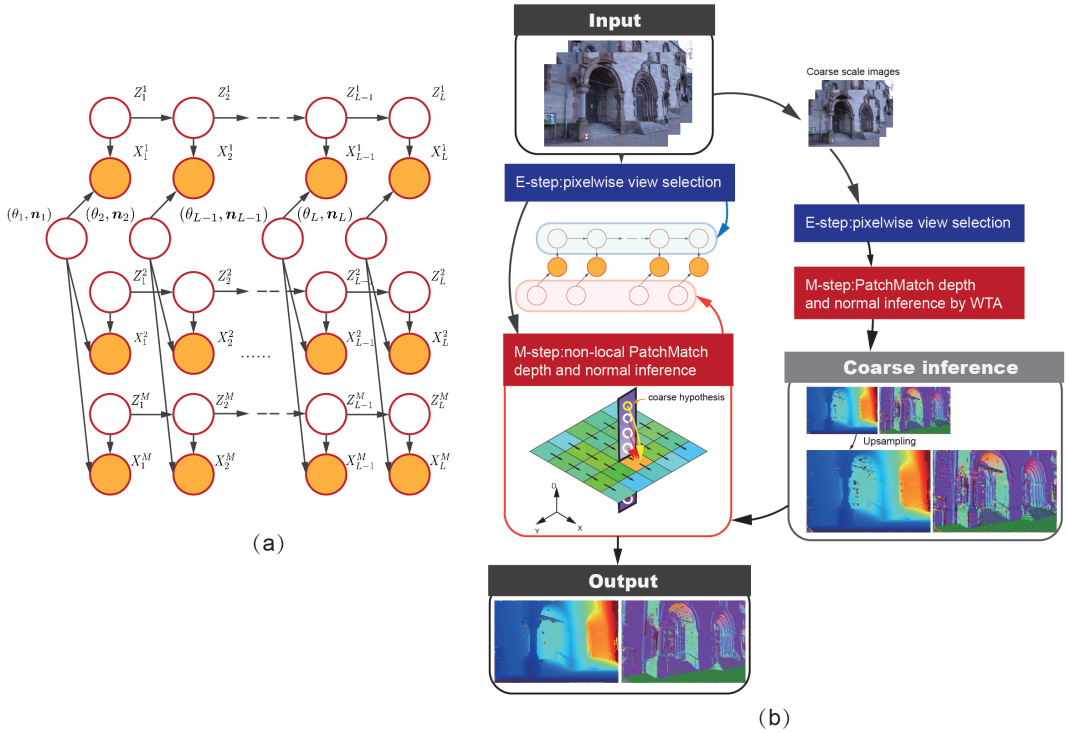

Z_1^1
Z_2^1
Z_{L-1}^1
Z_L^1
X_1^1
X_2^1
X_{L-1}^1
X_L^1
$(\theta_1, \boldsymbol{n}_1)$
$(\theta_2, \boldsymbol{n}_2)$
$(\theta_{L-1}, \boldsymbol{n}_{L-1})$
$(\theta_L, \boldsymbol{n}_L)$
Z_1^2
Z_2^2
Z_{L-1}^2
Z_L^2
X_1^2
X_2^2
X_{L-1}^2
X_L^2
......
Z_1^M
Z_2^M
Z_{L-1}^M
Z_L^M
X_1^M
X_2^M
X_{L-1}^M
X_L^M
(a)
Input
Coarse scale images
E-step:pixelwise view selection
E-step:pixelwise view selection
M-step:PatchMatch depth and normal inference by WTA
Coarse inference
Upsampling
M-step:non-local PatchMatch depth and normal inference
coarse hypothesis
Output
(b)

Fig.2. (a) Graphical model of [12, 20]. (b) The framework of our approach in one iteration. Here, θ_l and $\boldsymbol{n}_l$ are the depth and normal of pixel l, respectively. Z_l^m denotes the selection of image m at pixel l. The colored circle X_l^m is the observation on the source image m given depth θ_l and normal $\boldsymbol{n}_l$.

pixel utilizing the PatchMatch sampling scheme, and then solves the depth and normal estimation efficiently.

Non-local PatchMatch Methods. In stereo vision, the PatchMatch algorithm was introduced in some approaches such as PMBP [16], PM-Huber [17] and PMSC [18] for different elaborate optimizers to minimize the global energy function and then achieve better performance. From two-view to multi-view, the overall efficiency of non-local PatchMatch methods decreases because of the high optimization complexity and serial computations. Thus, considering the common merits of sequential propagation and dynamic programing, we combined these two methods into a novel parallel method.

3 Overview

Given a set of images with known camera parameters, our goal is to obtain the per-view depth map and normal map and then combine them into a complete 3D point cloud of the scene in the input images. The framework of our approach in one iteration is illustrated in Figure 2(b).

We first obtained the coarse inference by using the basic graphical model (Figure 2(a)), which jointly models pixelwise view selection and depth-normal estimation. After upsampling the estimation at the coarse scale, the coarse hypotheses were utilized to guide our parallel non-local PatchMatch MVS method. This method combines sequential propagation and dynamic programming. After several iterations, the optimal depth map and normal map were obtained.

In following sections, we describe the basic graphical model of the CNLPA-MVS in Section 4 and the detailed CNLPA-MVS algorithms in Section 5. The effectiveness of each individual part and the overall method are described in Section 6.

4 Basic Model of the CNLPA-MVS

The basic model [12, 20] of joint pixelwise view selection and depth-normal estimation is introduced in this section. Note that the framework processes each row and each column independently and alternately for the sake of parallel computing. We focus on a single line to describe this method.

Given a reference image $\boldsymbol{X}^{ref}$ and a set of source images $\boldsymbol{X}^{src} = \{\boldsymbol{X}^m \mid m = 1...M\}$, the method models the sequential estimation problem of depth θ_l and normal $\boldsymbol{n}_l$ for each pixel l as a Markov process and optimizes it iteratively. The probabilistic model corresponds the hidden states to binary indicator variables $Z_l^m \in \{0, 1\}$, which define the set of the non-occluded source images as $\overline{\boldsymbol{X}_l^m} = \{\boldsymbol{X}^m \mid Z_l^m = 1\}$. Then, the joint pixelwise inferences are formulated as a maximum-a posteriori (MAP) estimation where the posterior probability is as follows:

$$\begin{aligned} P(\boldsymbol{Z}, \theta, \boldsymbol{N} \mid \boldsymbol{X}) &= \frac{P(\boldsymbol{Z}, \theta, \boldsymbol{N}, \boldsymbol{X})}{P(\boldsymbol{X})} \\ &= \frac{1}{P(\boldsymbol{X})} \prod_{l=1}^{L} \prod_{m=1}^{M} \left[P\left(\boldsymbol{X}_l^m \mid Z_l^m, \theta_l, \boldsymbol{n}_l\right) \right. \\ &\quad \left. P\left(Z_{l,t}^m \mid Z_{l-1,t}^m, Z_{l,t-1}^m\right) P\left(\theta_l, \boldsymbol{n}_l \mid \theta_l^m, \boldsymbol{n}_l^m\right) \right], \end{aligned} \tag{1}$$

where L is the number of pixels in the considered line sweep; M is the number of source images and $\boldsymbol{X} = \{\boldsymbol{X}^{ref}, \boldsymbol{X}^{src}\}$. The likelihood term,

$$P(\boldsymbol{X}_l^m | Z_l^m, \theta_l, \boldsymbol{n}_l) = \begin{cases} \frac{1}{NA} \exp\left(-\frac{(1-\rho_l^m(\theta_l, \boldsymbol{n}_l))^2}{2\sigma_\rho^2}\right) & \text{if } Z_l^m = 1 \\ \frac{1}{N} u & otherwise, \end{cases} \tag{2}$$

represents the probability of photometric consistency between a reference patch $\boldsymbol{X}_l^{ref}$ and the source patches $\boldsymbol{X}_l^m$ in non-occluded source images, where $A = \int_{-1}^{1} \exp\left\{-\frac{(1-\rho)^2}{2\sigma_\rho^2}\right\} d\rho$ and N is a constant. In the case of occlusion, uniform distribution u in the range $[-1, 1]$ with probability densify 0.5 indicates that the two patches are irrelevant. A bilaterally weighted NCC based on color and spatial distances is used to compute the patch similarity ρ_l^m as the photometric consistency with σ_ρ as a constant. The spatial and temporal state-transitions are jointly modeled as $P\left(Z_{l,t}^m \mid Z_{l-1,t}^m, Z_{l,t-1}^m\right)$, which ensures that the occlusion maps are smooth both between adjacent pixels and along successive iterations. Finally, the geometric consistency term $P\left(\theta_l, \boldsymbol{n}_l \mid \theta_l^m, \boldsymbol{n}_l^m\right)$ enforces the consistency between multi-view depth and normal estimations.

To solve Equation (1), Zheng et al. [20] utilized a variational inference to estimate the optimal member of the family of approximate posterior. Similar to this work, Schönberger et al. [12] approximated $P(Z, \theta, \boldsymbol{N} \mid \boldsymbol{X})$ as $q(Z, \theta, \boldsymbol{N}) = q(Z)q(\theta, \boldsymbol{N})$ and in the sense that the Kullback-Leibler divergence between the two functions is minimized. Then, they utilized a variant of the generalized expectation-maximization (GEM) algorithm [39] to solve the corresponding probalistic model. In the E-step, the forward-backward algorithm through the Hidden Markov Model was employed to infer $q(Z_{l,t}^m)$ in iteration t while keeping $q\left(\theta_l, \boldsymbol{n}_l\right)$ fixed. In the M-step alternately, $q(Z_{l,t}^m)$ was fixed while $q\left(\theta_l, \boldsymbol{n}_l\right)$ was constrained to the family of Kronecker delta functions $q\left(\theta_l, \boldsymbol{n}_l\right) = \delta\left(\theta_l = \theta_l^*, \boldsymbol{n}_l = \boldsymbol{n}_l^*\right)$, and calculated via the PatchMatch sequential propagation and sampling. The optimal $\hat{\theta}_l^{opt}$ and $\hat{\boldsymbol{n}}_l^{opt}$ are calculated as follows:

$$\left(\hat{\theta}_l^{opt}, \hat{\boldsymbol{n}}_l^{opt}\right) = \underset{\theta_l^*, n_l^*}{\operatorname{argmin}} \frac{1}{|S|} \sum_{m \in S} \xi_l^m\left(\theta_l^*, \boldsymbol{n}_l^*\right), \tag{3}$$

$$\xi_l^m\left(\theta_l^*, \boldsymbol{n}_l^*\right) = 1 - \rho_l^m\left(\theta_l, \boldsymbol{n}_l\right) + \eta \min\left(\psi_l^m, \psi_{\max}\right). \tag{4}$$

S is a subset of source images according to probability $P_l(m)$, which prefers the non-occluded images with three priors, i.e., sufficient baseline, similar resolution, and non-oblique viewing direction. The cost $\xi_l^m\left(\theta_l^*, \boldsymbol{n}_l^*\right)$ combines the photometric cost ρ_l^m and the geometric forward-backward reprojection error $\psi_l^m = \|x_l - H_l^m H_l x_l\|$, where H_l^m and H_l denote the homograph matrix mapping of the patch from the source to the reference image and from the reference to the source image, respectively. Further, η and ψ_{max} are two constants that are set to 0.5 and 3 px. Moreover, the pair $(\theta_l^*, \boldsymbol{n}_l^*)$ is chosen from the set of hypotheses:

$$\begin{gathered}\left\{(\theta_l, \mathbf{n}_l), \left(\theta_{l-1}^{\mathrm{prp}}, \mathbf{n}_{l-1}\right), \left(\theta_l^{\mathrm{rnd}}, \mathbf{n}_l\right), \left(\theta_l, \mathbf{n}_l^{\mathrm{rnd}}\right),\right. \\ \left.\left(\theta_l^{\mathrm{rnd}}, \mathbf{n}_l^{\mathrm{rnd}}\right), \left(\theta_l^{\mathrm{prt}}, \mathbf{n}_l\right), \left(\theta_l, \mathbf{n}_l^{\mathrm{prt}}\right)\right\},\end{gathered} \tag{5}$$

where θ_l^{rnd} and $\boldsymbol{n}_l^{rnd}$ are randomly generated samples; θ_l^{prt} and $\boldsymbol{n}_l^{prt}$ are two samples that are slightly disturbed in depth space and normal space, respectively; and θ_{l-1}^{prp} and $\boldsymbol{n}_{l-1}$ denote the propagations from the parameters of the previous pixel.

5 Detailed Algorithms in the CNLPA-MVS

In this section, we describe the proposed algorithms in detail. Our algorithm leverages the graphical model introduced in Section 4. In Section 5.1, a new parallel non-local PatchMatch based depth-normal estimation framework is presented. This framework can be used to ensure that the matching uncertainty in low-textured regions due to the missing of global information can be alleviated significantly. In Section 5.2, the coarse-hypotheses guidance mechanism is proposed. This mechanism can propagate the hypotheses estimation at a coarser scale in low-textured regions and reduce the over-smoothing caused by simple dynamic programming. In Section 5.3, a robust local consistency measurement is proposed by considering the sequential propagation scheme comprehensively.

5.1 Non-local Depth and Normal Inference

Dense stereo matching can be effectively solved by minimizing a MRF global energy function consisting of a data term and smoothness term:

$$E = \sum_l \varphi(l, \boldsymbol{u}_l) + \sum_l \sum_{r \epsilon N_l} \psi(l, r, \boldsymbol{u}_l, \boldsymbol{u}_r), \tag{6}$$

where N_l is the pairwise neighborhood set. The data term $\varphi(l, \boldsymbol{u}_l)$ computes the local cost for the hypotheses label $\boldsymbol{u}_l = (\theta_l, \boldsymbol{n}_l)$ of each pixel l. The smoothness term $\psi(l, r, \boldsymbol{u}_l, \boldsymbol{u}_r)$, provides a constraint that the planes defined by the hypotheses change smoothly except at the object boundaries.

After defining the energy function (6), an appropriate solver should be selected to minimize the energy function. The energy function is defined in terms of continuous variables $\boldsymbol{u}_l$. Hence, we first describe the label-set selection strategy in Section 5.1.1 and then provide a novel parallel optimization scheme in Section 5.1.2.

5.1.1 Hypotheses Set Generation

The smoothness term makes the minimization challenging. For a discrete space where $\boldsymbol{u}_l$ is in a finite set of size D, assuming the number of pixels in the reference image is n, the worst complexity of minimization is $O(D^n)$, which is extremely high. For a continuous case, the minimization problem will be more complicated because of the continuous feasible region.

The key to solve the global energy function for continuous state variables $\boldsymbol{u}_l$ is representation selection for each pixel hypotheses label. Similar to the previous work [16], we associate each pixel with a hypotheses label set $H_l = \left\{\left(\theta_l^{(i)}, \boldsymbol{n}_l^{(i)}\right)\right\}_{i=1}^K$ during each step of estimation. At the beginning, we random initialize K hypotheses as the plane set H_l for each pixel. Then after neighborhood propagation and resampling in each sweep, K best planes are selected as the new hypotheses set by minimizing (6). The candidate hypotheses at each propagation step in PatchMatch from (5) are modified as follows:

$$\begin{gathered}\left\{\left(\theta_l^{(i)}, \boldsymbol{n}_l^{(i)}\right), \left(\theta_{l-1}^{prp(i)}, \boldsymbol{n}_{l-1}^{(i)}\right), \left(\theta_l^{rnd(i)}, \boldsymbol{n}_l^{(i)}\right), \left(\theta_l^{(i)}, \boldsymbol{n}_l^{rnd(i)}\right),\right. \\ \left.\left(\theta_l^{rnd(i)}, \boldsymbol{n}_l^{rnd(i)}\right), \left(\theta_l^{prt(i)}, \boldsymbol{n}_l^{(i)}\right), \left(\theta_l^{(i)}, \boldsymbol{n}_l^{prt(i)}\right)\right\}_{i=1}^K.\end{gathered} \tag{7}$$

Note that this strategy still maintains optimization over a continuous $\boldsymbol{u}_l$ and is not limited to the current label set H_l. However, the computation of the smoothness term in (6) over the continuous $\boldsymbol{u}_r$ is changed into a calculation over a finite set H_r of size K.

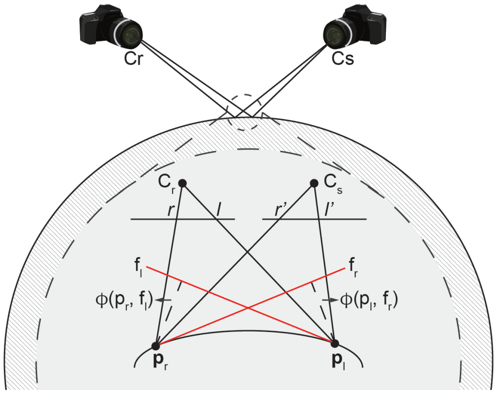


Fig.3. Plane similarity measurement. The illustration simplifies the multi-view problem into a two-view problem. Suppose C_r and C_s are the camera centers of the reference view and the source view, respectively, and p_l and p_r are the reconstructed 3D points of the pixel l and its neighborhood r, respectively. Red lines f_l and f_r show the tangent planes of p_l and p_r, respectively. The plane similarity measurement δ_{lr} can be computed using the two distances (dotted lines) from points to planes.

5.1.2 Optimization

As an early method in stereo matching, dynamic programming (DP) still remains one of the most popular optimizers due to its effective 1D optimization performance. DP solves a complex energy function by dividing it into several sub-problems that are minimized on a subset of the image, typically along a scanline. It can be discovered that the process flow of DP is similar to the sequential propagation scheme of PatchMatch, as illustrated in Figure 1.

To maintain the parallel solution framework reviewed in Section 4, the traditional DP optimizer is adopted to minimize the energy function. In this work, the data term is defined as shown in (8), which is directly obtained from (3) and (4), and the smoothness term is shown in (9):

$$\varphi\left(l, \boldsymbol{u}_l\right) = \frac{1}{|S|} \sum_{m \in S} \xi_l^m \left(\theta_l^*, \boldsymbol{n}_l^*\right), \tag{8}$$

$$\psi\left(l, r, \boldsymbol{u}_l, \boldsymbol{u}_r\right) = \lambda\left(1 - \delta_{lr}\right), \tag{9}$$

where λ is a constant regularizer. We denote $\delta_{lr} = \exp\left(\frac{\phi(p_l, f_r) + \phi(p_r, f_l)}{2\gamma_s}\right)$ as the plane similarity measurement. Considered the distance between the two tangent planes f_r and f_l should be small enough (approximately equal to 0), we set γ_s to an extremely small value (0.0005) to strictly estimate the plane similarity. As shown in Figure 3, $\phi(p_l, f_r)$ is defined as the distance from the reconstructed point p_l to the local plane f_r at the neighboring pixel, r and $\phi(p_r, f_l)$ has a similar definition. Note that the photometric consistency cost ${\rho_l}^m$ in (4) using bilaterally weighted NCC is replaced by the modified measurement from a previous study [19]. The photometric consistency cost ${\rho_l}^m$ is defined as follows:

$$\rho_l^m = \begin{cases} g & \text{if } h = 0 \\ \eta + 0.1 & \text{if } h = 2 \text{ and } |c_l^m - c_l| \leq 3\sigma_c \\ -1 & otherwise, \end{cases} \tag{10}$$

where g represents the original bilaterally weighted NCC in (4), σ_c is set to 0.05 that is the same as the previous work [19] and $\eta = 1 - \sigma_\rho\sqrt{-2\ln\left(\frac{A}{2}\right)}$. Then, $P\left(\boldsymbol{X}_l^m \mid Z_l^m, \theta_l, \boldsymbol{n}_l\right)$ becomes equal when $Z_l^m = 1$ and $Z_l^m = 0$. We denote h as the number of low-textured patches in $\boldsymbol{X}^{ref}$ and $\boldsymbol{X}_l^m$. For the case that both patches are textured, a bilaterally weighted NCC g is used for evaluation. In contrast, if colors c_l^m and c_l of the two low-textured patches are similar, the metric $\eta + 0.1$ slightly favors that $\boldsymbol{X}_l^m$ is non-occluded. For other cases, the two patches are regarded as unrelated and set to -1. The piecewise photometric consistency cost ρ_l^m is more suitable for the areas with homogeneous color.

The energy function can be minimized in parallel along each row/column alternatively as follows:

$$\left\{\left(\theta_l^{(i)}, \boldsymbol{n}_l^{(i)}\right)\right\}_{i=1}^{K} = \arg\min_{K} M\left(l, \theta_l^*, \boldsymbol{n}_l^*\right), \tag{11}$$

$$\begin{aligned} M\left(l, \theta_l^*, \boldsymbol{n}_l^*\right) = \tfrac{1}{|S|} \sum_{m \in S} \xi_l^m\left(\theta_l^*, \boldsymbol{n}_l^*\right) + \min_{\left(\theta_{l-1}^{(i)}, \boldsymbol{n}_{l-1}^{(i)}\right) \in H_{l-1}} \\ \left[M\left(l-1, \theta_{l-1}^{(i)}, \boldsymbol{n}_{l-1}^{(i)}\right) + \lambda\left(1 - \delta_{l(l-1)}\right)\right]. \end{aligned} \tag{12}$$

During each sweep, the hypotheses set is selected as the K best planes by minimizing the energy, and simultaneously, the corresponding energies $M\left(l, \theta_l^{(i)}, \boldsymbol{n}_l^{(i)}\right)$ are saved to compute the plane set of the next pixel. The optimal $\hat{\theta}_l^{opt}$ and $\hat{\boldsymbol{n}}_l^{opt}$ are calculated by optimizing the following function:

$$\left(\hat{\theta}_l^{opt}, \hat{\boldsymbol{n}}_l^{opt}\right) = \underset{\left(\theta_l^{(i)}, \boldsymbol{n}_l^{(i)}\right) \in H_l}{\operatorname{argmin}} M\left(l, \theta_l^{(i)}, \boldsymbol{n}_l^{(i)}\right). \tag{13}$$

This process is performed for each scanline to obtain all the optimal hypotheses of the entire image.

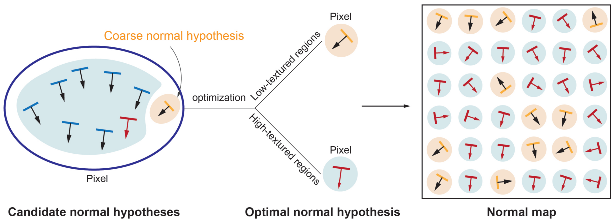


Fig.4. Optimal normal hypothesis selection. Candidate hypotheses set of one pixel is shown as a blue elliptic circle on the left. Hypotheses in blue regions are the candidate set at the original scale, and the optimal hypothesis is indicated in red. The coarse normal hypothesis (indicated in orange) is introduced to the candidate hypotheses set. After propagation and sampling of all pixels, a normal map can be obtained.

5.2 Coarse-hypotheses Guidance

Although acquiring the best hypotheses through DP with non-local information can effectively alleviate the matching problems in low-textured regions, two problems persist. First, finding the accurate hypotheses using only the DP is difficult in the case of high-resolution images that contain numerous pixels of homogeneous colors as the method only converges to the optimal solution along the scanline. Second, the depths estimated by the simplified DP algorithm change slowly in the depth discontinuity areas resulting in oversmoothing and stripe artifacts.

We analyzed the selection strategy of optimal hypotheses and found that the coarse hypotheses, that is, optimal depth and normal estimations acquired at a coarser scale, can help to improve the completeness of 3D reconstruction in low-textured areas. For these regions, it will be more discriminative to match under identical window sizes when an image is down-sampled [14, 29]. In this case, coarse hypotheses are more reliable than hypotheses at the original scale and may be selected as the optimum. However, in high-textured regions, the optimal hypotheses at the coarse scale are inadequate for the fine scale as details may be blurred under the universal window size. The coarse hypotheses can be excluded as the optimal hypotheses since the photometric consistency cost of coarse hypotheses may be higher than that of some fine hypotheses. Instead of using the coarse inference for all pixels as the initial hypotheses [14, 29], we added the optimal hypotheses at the coarse scale $({\theta_l}^0, \boldsymbol{n_l}^0)$ as an additional candidate into the candidate hypotheses set for each pixel. Then, (7) is modified as follows:

$$\begin{aligned}&\Big\{ \Big\{ \Big(\theta_l^{(i)},\ n_l^{(i)}\Big), \Big(\theta_{l-1}^{prp(i)}, n_{l-1}^{(i)}\Big), \Big(\theta_l^{rnd(i)}, n_l^{(i)}\Big), \Big(\theta_l^{(i)}, n_l^{rnd(i)}\Big),\\ &\Big(\theta_l^{rnd(i)}, n_l^{rnd(i)}\Big), \Big(\theta_l^{prt(i)}, n_l^{(i)}\Big), \Big(\theta_l^{(i)}, n_l^{prt(i)}\Big) \Big\}_{i=1}^{K}, \left(\theta_l^0, n_l^0\right) \Big\}.\end{aligned} \tag{14}$$

Finally, the coarse hypotheses are likely to be chosen as the optimal hypotheses in low-textured regions while retaining the high-frequency details. The process of optimal normal hypothesis selection is shown in Figure 4.

To alleviate the oversmoothing caused by DP, we integrated the WTA results into our method. We first applied the WTA strategy to estimate the approximate inference at a coarser resolution in each sweep. The down-sampling factor is denoted as $\frac{1}{2^s}$. Images are not downscaled when $s = 0$. We used the coarse depth maps and normal maps estimated by the WTA method to prevent stripe artifacts in DP results and propagated the downscaled texture information to the fine scale through a joint bilateral up-sampler [40]. The core idea is to add the coarse hypotheses of the previous pixel $(\theta_{l-1}^0, \boldsymbol{n}_{l-1}^0)$ to the candidate hypotheses set $H_{l-1}^{'} = \left\{H_{l-1}, (\theta_{l-1}^0, \boldsymbol{n}_{l-1}^0)\right\}$ for the DP process. Then, (12) is modified as follows:

$$\begin{aligned}M\left(l, \theta_l^*, \boldsymbol{n}_l^*\right) = \tfrac{1}{|S|} \sum_{m \in S} \xi_l^m \left(\theta_l^*, \boldsymbol{n}_l^*\right) + \min_{\left(\theta_{l-1}^{'}, \boldsymbol{n}_{l-1}^{'}\right) \in H_{l-1}^{'}}\\ \left[M\left(l-1, \theta_{l-1}^{'}, \boldsymbol{n}_{l-1}^{'}\right) + \lambda\left(1 - \delta_{l(l-1)}\right)\right].\end{aligned} \tag{15}$$

Thus, we implicitly transferred more reliable estimations from a coarser resolution to a finer resolution via a supplemental plane label. This helped avoid error propagation from the wrong hypotheses in areas with abundant details. Guided by the coarser information, a better trade-off between smoothness and details can be made.

5.3 Local Consistency

It is hard to estimate the depth and normal in low-textured regions only constrained by the photometric and geometric consistency. To alleviate this difficulty, we adopted a new local consistency strategy that sufficiently accounts for the characteristics of sequential propagation in PatchMatch. We assumed that adjacent pixels with similar colors are likely to belong to the same plane, such that the distance from the point of one pixel to the ones of its neighboring pixels is extremely close. We gave a larger penalty to constrain these pixels with homogeneous colors. Therefore, the smoothness term is changed as follows:

$$\psi\left(l, r, \boldsymbol{u}_l, \boldsymbol{u}_r\right) = \lambda\left(1+\varepsilon_{lr}\right)\left(1-\delta_{lr}\right). \tag{16}$$

We denote the color similarity measurement as $\varepsilon_{lr} = \exp\left(-\frac{|c_l-c_r|}{\gamma_c}\right)$, where $|c_l - c_r|$ denotes the color distances between pixel l and its neighborhood r. The constant γ_c is preset to 1.0, giving a slightly larger constraint to make the distance between the two reconstructed points of l and r with similar colors closer but avoiding over smooth. Hence, the minimization function along the scanline is defined as follows:

$$\begin{gathered} M\left(l, \theta_l^*, \boldsymbol{n}_l^*\right) = \tfrac{1}{|S|} \sum_{m \in S} \xi_l^m\left(\theta_l^*, \boldsymbol{n}_l^*\right) + \min_{\left(\theta_{l-1}^{\prime}, \boldsymbol{n}_{l-1}^{\prime}\right) \in H_{l-1}^{\prime}} \\ \left[M\left(l-1, \theta_{l-1}^{\prime}, \boldsymbol{n}_{l-1}^{\prime}\right) + \lambda\left(1+\varepsilon_{lr}\right)\left(1-\delta_{l(l-1)}\right)\right]. \end{gathered} \tag{17}$$

Unlike the method adopted by Liao et al. [19], who considered all 4-neighborhoods and were unsure of the latest status, our method only accounts for the previous pixel to avoid the impact of wrong hypotheses (Figure 5). Moreover, similar to the previous work [19], we also applied the local consistency to estimate view selection probability $P(Z_l^m)$. The transition probability of Z is formulated as follows:

$$P\left(Z_l^m \mid Z_{l-1}^m\right) = \begin{pmatrix} \gamma & 1-\gamma \\ 1-\gamma & \gamma \end{pmatrix}, \tag{18}$$

where the transition probability of view selection denotes $\gamma = \mu\varepsilon_{l(l-1)}$. Therefore, the view selection estimations between the adjacent pixels with similar colors are likely to be smooth. To avoid oversmoothness regardless of photometric consistency, we preset $\mu = 0.999$ to prevent $\gamma = 1$.

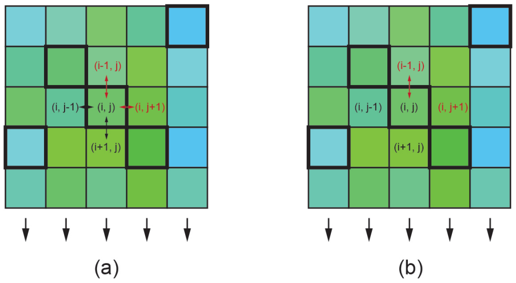


Fig.5. Difference in local consistency between [19] (a) and the proposed method (b). The black arrows show the sequential propagation direction. The bold pixels show the currently processing pixels in different scanlines, and the hypotheses of their previous pixels are the latest. We focus on pixel (i, j) and its adjacent pixels. A constraint based on the color similarity measurement (double-headed arrows) is given to assume coherent depths. The red fonts and arrows show data with the latest states. Note that in the previous work [19], 4-neighborhoods were used, while the proposed method only considers the previous one.

6 Experiments and Disccusion

The proposed method was implemented in C++ with CUDA and executed on a PC with Intel Core i7-6700K CPU, 64GB RAM and a couple of GeForce GTX 1080Ti GPUs. All the experiments were performed on ETH3D benchmark [5] and Strecha dataset [6]. Three criteria proposed in [5] were used for evaluation, including the evaluation of accuracy, completeness, and F_1 score which is defined as the harmonic mean of accuracy (precision) and completeness (recall).

Following the previous work [16], the number of hypotheses set K was set to 5. Balancing the details and completeness, we set $s = 2$ to obtain the coarse optimal hypotheses (θ_l^0, $\boldsymbol{n}_l^0$). All other parameters are the same as the default of values adopted in [12]. We adopted the fusion method implemented in [12] to obtain a whole point cloud.

We validated the effectiveness of three individual components with the proposed method, i.e., non-local depth and normal inference, coarse-hypotheses guidance, and local consistency. The overall qualitative and quantitative evaluations were subsequently done by comparing with the state-of-the-art MVS methods. Finally, we discussed the parameter analysis, versatility, and limitations of the proposed method.

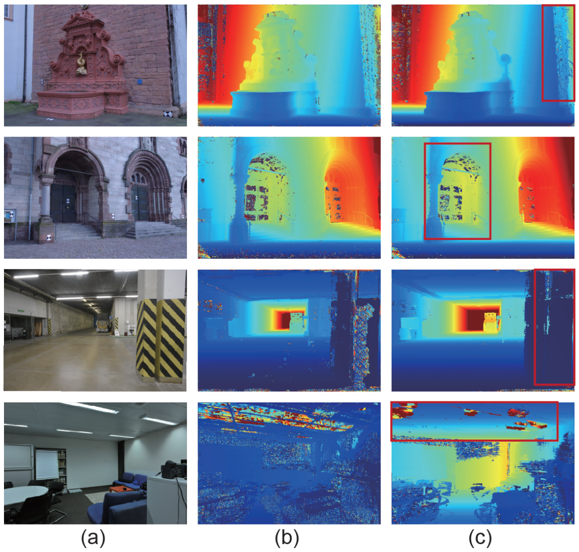


Fig.6. (a) Reference image. Depth maps obtained before (b) and after (c) using the non-local depth and normal inference method.

6.1 Validation of Three Components in CNLPA-MVS

Non-local Depth and Normal Inference. The method for non-local depth and normal inference proposed in Section 5.1 introduces non-local information and obtains the optimal depth and normal hypotheses through DP to alleviate the matching ambiguity in low-textured regions. To test the effectiveness of the non-local depth and normal inference, we first compared our result (Figure 6(c)) with the baseline local algorithm [12] (Figure 6(b)). The wrong depth estimation in low-textured regions was alleviated. A quantitative comparison is presented in Table 1. NL and COLMAP denote our method and the baseline algorithm [12], respectively. It can be seen that the completeness and F_1 score improved after utilizing the non-local depth and normal inference method.

Coarse-hypotheses Guidance. The introduction of coarse hypotheses to the candidate hypotheses set propagates more reliable estimations in low-textured regions obtained at a coarser scale to a finer scale, and decreases the oversmoothing. As shown in Figure 7(b), some details, particularly, edges, are missing owing to scanline optimization during DP. After integrating WTA results into DP process, the oversmoothing was alleviated, and more details were obtained (Figure 7(c)). We subsequently replaced the WTA result at the original scale ($s = 0$) by the result at the coarse scale ($s = 2$). The proposed coarse-hypotheses guidance strategy reduces noises in the low-textured regions and improves the completeness (Figure 7(d)). Quantitative results are provided in Table 1. Without coarse-hypotheses guidance (w/o CI), the accuracy and completeness both decreased.

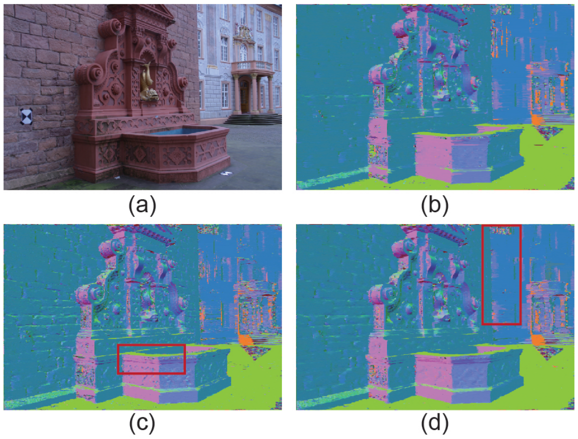


Fig.7. (a) Reference image. Normal maps obtained utilizing (b) non-local PatchMatch method with DP, (c) after introducing WTA result at the original scale, and (d) at a coarse scale with $s = 2$.

Local Consistency. The proposed local consistency further improves the quality of reconstruction in low-textured regions. A qualitative comparison is shown in Figure 8; many false depth values in low-textured areas are corrected and a cleaner depth map is obtained after introducing local consistency. Quantitative comparisons are summarized in Table 1. Our method without local consistency is denoted as w/o LC. It can be seen that F_1 score and completeness are reduced in this case.

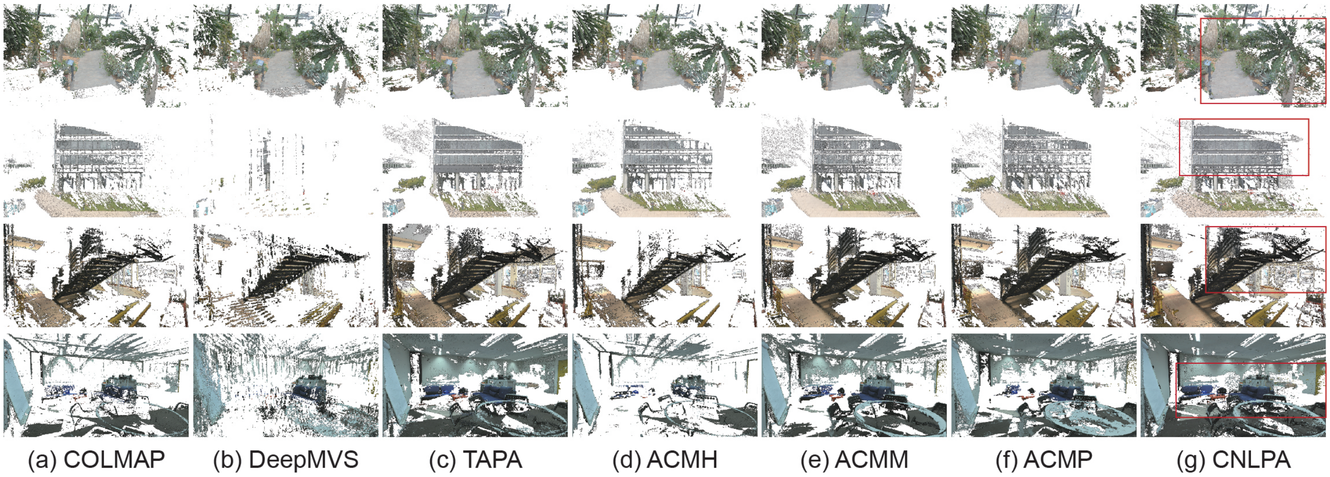


Fig.9. Qualitative point cloud comparison on some high-resolution datasets of ETH3D benchmark.

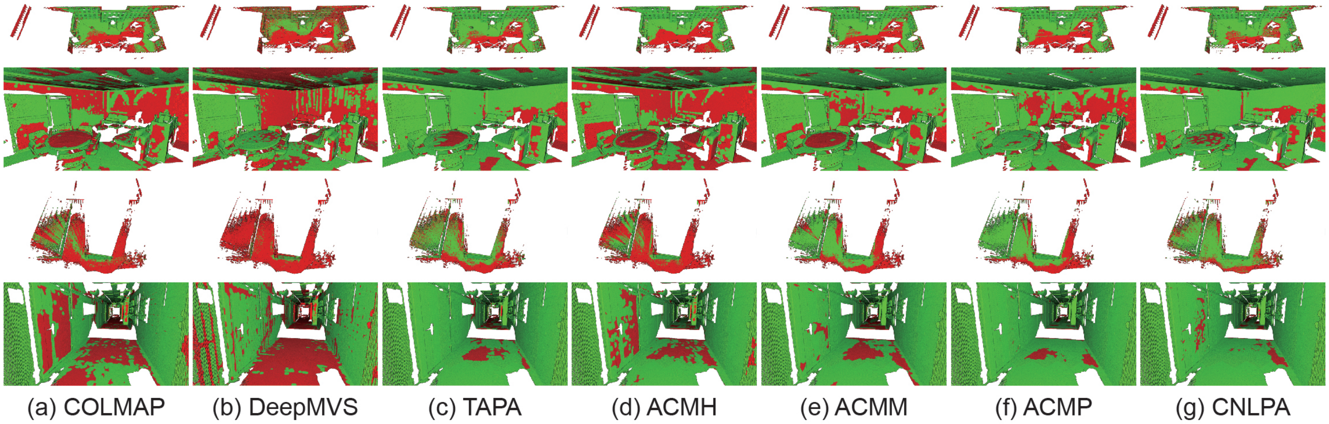


Fig.10. Completeness comparison on some high-resolution training datasets of ETH3D benchmark with threshold of 5 cm.

and λ was set to 7. We evaluated our method by comparing with the baseline method [12] (COLMAP), a leaning-based method [7] (DeepMVS), and other state-of-the-art methods for improving completeness (TAPA [30], ACMH [14], ACMM [14], and ACMP [31]). Qualitative point cloud comparisons and completeness comparison results are shown in Figure 9 and Figure 10, respectively. The reconstructed models by the proposed CNLPA-MVS are more complete than those obtained by other methods. Table 2 lists the quantitative comparisons on the test datasets with thresholds of 5 cm and 10 cm. In the term of completeness, CNLPA outperforms other methods in both indoor and outdoor scenarios. In the term of F_1 score, for a threshold of 5 cm, CNLPA ranks second over high-resolution test datasets with only 0.07 points less than the first. For a threshold of 10 cm, CNLPA ranks first over high-resolution test datasets.

Table 3. Results (F_1 score, accuracy, and completeness) of high-resolution multi-view training datasets of ETH3D benchmark with different thresholds and λ coefficients.

method	1 cm			2 cm			5 cm		
	F_1	A	C	F_1	A	C	F_1	A	C
$\lambda = 5$	62.63	**77.15**	53.49	77.08	**85.80**	70.91	**88.43**	**93.33**	84.54
$\lambda = 6$	62.72	76.97	53.74	77.12	85.68	71.10	88.41	93.28	84.47
$\lambda = 7$	**62.78**	76.67	54.00	**77.20**	85.50	71.37	88.40	93.14	**84.72**
$\lambda = 8$	62.72	76.34	54.11	77.16	85.27	71.49	88.24	93.00	84.59
$\lambda = 10$	62.47	75.43	**54.22**	76.95	84.61	**71.64**	88.08	92.58	84.70

Strecha Benchmark. We further tested our method on fountain-P11 and HerzJesu-P8 [6]. Note that fountain-P11 and HerzJesu-P8 have 11 and 8 images, respectively, with a resolution of 3072×2048. We set $\lambda = 1$ for reconstructing the two relatively well-textured scenes. Since the online service is not available, we compared our method with the two open-source algorithms (COLMAP [12] and DeepMVS [7]) among the above state-of-the-art methods. As shown in Figure 11, we calculated the ratio of pixels with error less than different thresholds from the ground truth. It can be observed that even for well-textured scenes, our method possess the best performance and still slightly improves the depth-map results of COLMAP.

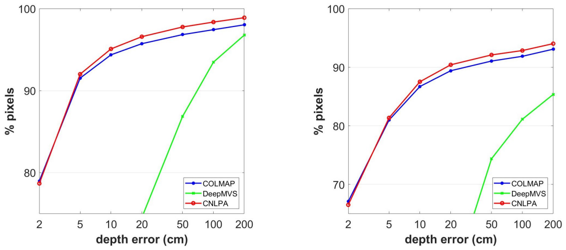


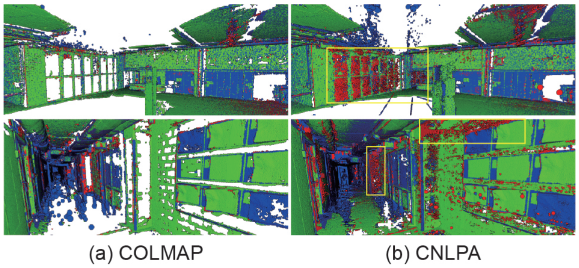
Fig.11. Depth error distributions with different thresholds on fountain-P11 and HerzJesu-P8 datasets.

6.3 Discussion

Parameter Analysis. We analyzed the performance of CNLPA-MVS with different λ coefficients. As shown in Table 3, the accuracy of the models reconstructed by CNLPA-MVS decreased with increasing λ as some details may have been smoothed by an over-weighted penalty. With increasing λ, the completeness increased. The reason is that the increasing penalty will give a stronger constraint between the adjacent pixels, which is conducive to the reconstruction in low-textured regions. Thus, balancing the accuracy and the completeness, the F_1 score first increased and then decreased with increasing λ.

Versatility. The above evaluations demonstrate that CNLPA-MVS is suitable for most planes, such as curved or slanted planes. Particularly, as shown in the first and last rows of Figure 9, CNLPA-MVS has a great power to alleviate the matching uncertainty problem under strong occlusions in low-textured regions. Theses problems are common in indoor and outdoor scenarios, such as libraries, game halls, gardens, and crowded streets. Therefore, CNLPA-MVS can be widely applied for reconstructing generic scenarios.

Limitations. For accuracy, since the optimization strategy of our method is limited in a scanline, some reconstructed low-textured regions are inaccurate (e.g., the colored rectangles shown in Figure 12), such that the overall accuracy is decreased. A more global optimization strategy can alleviate this limitation and will be further considered for a better reconstructed result. For efficiency, although the consuming time of CNLPA-MVS was dramatically less than that of other global PatchMatch methods, the efficiency is still limited in that it is difficult to process real-time tasks by our method. We will account for a better algorithm parallelism strategy to further improve the performance.



(a) COLMAP (b) CNLPA

Fig.12. Accuracy comparisons on ETH3D benchmark with threshold of 2 cm. The ground-truth values in blue regions are missed. The green and red regions represent the accurate and inaccurate regions, respectively.

7 Conclusions

In this paper, we proposed a coarse-hypotheses guided non-local PatchMatch MVS method. This method can efficiently alleviate the matching uncertainty problem in low-textured regions. The proposed parallel non-local depth and normal inference algorithm optimized by DP introduces more information that helps achieve accurate matching for low-textured regions. The proposed coarse-hypotheses guidance strategy is conducive to reduce stripe artifacts caused by simple DP and to improve the completeness of low-textured regions by leveraging the coarse WTA inference. Finally, a new local consistency strategy was proposed to further improve the completeness by assuming adjacent pixels sharing approximate depth values. Experiments showed that CNLPA-MVS achieves state-of-the-art performance with high completeness and can be widely used for reconstructing low-textured regions under strong occlusions.

CNLPA-MVS offers a new solution to improve the efficiency of optimization in the global MRF methods. In future works, we will combine the diffusion-like propagation and a global optimizer to further improve the computational efficiency and the quality of reconstructed models.

References

[1] Xiao X, Xu C, Wang J, Xu M. Enhanced 3-d modeling for landmark image classification. *IEEE Trans. Multim.*, 2012, 14(4):1246-1258.

[2] Forster C, Pizzoli M, Scaramuzza D. Air-ground localization and map augmentation using monocular dense reconstruction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pages 3971-3978.

[3] Hedman P, Alsisan S, Szeliski R, Kopf J. Casual 3d photography. *ACM Trans. Graph.*, 2017, 36(6):234:1-234:15.

[4] Knapitsch A, Park j, Zhou Q Y, Koltun V. Tanks and temples: benchmarking large-scale scene reconstruction. *ACM Trans. Graph.*, 2017, 36(4):78:1-78:13.

[5] Schöps T, Schönberger J L, Galliani S, Sattler T, Schindler K, Pollefeys M, Geiger A. A multi-view stereo benchmark with high-resolution images and multicamera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pages 2538-2547.

[6] Strecha C, von Hansen W, Gool L V, Fua P, Thoennessen U. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pages 1-8.

[7] Huang P, Matzen K, Kopf J, Ahuja N, Huang J. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pages 2821-2830.

[8] Luo K, Guan T, Ju L, Huang H, Luo Y. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pages 10451-10460.

[9] Yao Y, Luo Z, Li S, Fang T, Quan L. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, 2018, volume 11212, pages 785-801.

[10] Campbell N D F, Vogiatzis G, Hernandez C, Cipolla R. Using multiple hypotheses to improve depthmaps for multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, 2008, volume 5302, pages 766-779.

[11] Woodford O J, Torr P H S, Reid I, Fitzgibbon A W. Global stereo reconstruction under second-order smoothness priors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009, 31(12):2115-2128.

[12] Schönberger J L, Zheng E, Frahm J, Pollefeys M. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, 2016, volume 9907, pages 501-518.

[13] Galliani S, Lasinger K, Schindler K. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pages 873-881.

[14] Xu Q, Tao W. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pages 5483-5492.

[15] Bleyer M, Rhemann C, Rother C. Patchmatch stereo - stereo matching with slanted support windows. In *Proceedings of the British Machine Vision Conference*, 2011, pages 1-11.

[16] Besse F, Rother C, Fitzgibbon A W, Kautz j. Pmbp: Patchmatch belief propagation for correspondence field estimation. *Int. J. Comput. Vis.*, 2014, 110(1):2-13.

[17] Heise P, Klose S, Jensen B, Knoll A C. Pm-huber: Patchmatch with huber regularization for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pages 2360-2367.

[18] Li L, Zhang S, Yu X, Zhang L. Pmsc: Patchmatchbased superpixel cut for accurate stereo matching. *IEEE Trans. Circuits Syst. Video Technol.*, 2018, 28(3):679-692.

[19] Liao J, Fu Y, Yan Q, Xiao C. Pyramid multiview stereo with local consistency. *Comput. Graph. Forum*, 2019, 38(7):335-346.

[20] Zheng E, Dunn E, Jojic V, Frahm J. Patchmatch based joint view selection and depthmap estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pages 1510-1517.

[21] Furukawa Y, Ponce J. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, 32(8):1362-1376.

[22] Locher A, Perdoch M, Gool L V. Progressive prioritized multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pages 3244-3252.

[23] Vogiatzis G, Esteban C H, Torr P H S, Cipolla R. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007, 29(12):2241-2246.

[24] Ulusoy A O, Geiger A, Black M J. Towards probabilistic volumetric reconstruction using ray potentials. In *Proceedings of the IEEE International Conference on 3D Vision*, 2015, pages 10-18.

[25] Vu H H, Labatut P, Pons J P, Keriven R. High accuracy and visibility-consistent dense multiview stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, 34(5):889-901.

[26] Seitz S M, Curless B, Diebel J, Scharstein D, Szeliski R. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pages 519-528.

[27] Furukawa Y, Hernandez C. Multi-view stereo: A tutorial. Found. *Trends Comput. Graph. Vis.*, 2015, 9(1-2):1-148.

[28] Barnes C, Shechtman E, Finkelstein A, Goldman D B. Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 2009, 28(3):24.

[29] Wei J, Resch B, Lensch H P A. Multi-view depth map estimation with cross-view consistency. In *Proceedings of the British Machine Vision Conference*, 2014.

[30] Romanoni A, Matteucci M. TAPA-MVS: texturelessaware patchmatch multi-view stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10412-10421, 2019.

[31] Xu Q, Tao W. Planar prior assisted patchmatch multiview stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pages 12516-12523.

[32] Xu Z, Liu Y, Shi X, Wang Y, Zheng Y. MARMVS: matching ambiguity reduced multiple view stereo for efficient large scale scene reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pages 5980-5989.

[33] Boykov Y, Veksler O, Zabih R. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2001, 23(11):1222-1239.

[34] Taniai T, Matsushita Y, Naemura T. Graph cut based continuous stereo matching using locally shared labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pages 1613-1620.

[35] Ogawara K. Approximate belief propagation by hierarchical averaging of outgoing messages. In *Proceedings of the IEEE International Conference on Pattern Recognition*, 2010, pages 1368-1372.

[36] Yu T, Lin R, Super B J, Tang B. Efficient message representations for belief propagation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2007, pages 1-8.

[37] Kolmogorov V, Zabih R. Computing visual correspondence with occlusions via graph cuts. In *Proceedings of the International Conference On Computer Vision*, 2001, pages 508-515.

[38] Klaus K, Sormann M, Karner K F. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *Proceedings of the IEEE International Conference on Pattern Recognition*, 2006, pages 15-18.

[39] Neal R M, Hinton G E. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, 1998, volume 89, pages 355-368.

[40] Kopf J, Cohen M F, Lischinski D, Uyttendaele M. Joint bilateral upsampling. *ACM Trans. Graph.*, 2007, 26(3):96.

[41] Li Y, Min D, Brown M S, Do M N, Lu J. SPMBP: sped-up patchmatch belief propagation for continuous mrfs. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pages 4006-4014.