Pixelwise View Selection for Unstructured Multi-View Stereo

Johannes L. Schönberger[1], Enliang Zheng[2],
Marc Pollefeys[1,3], Jan-Michael Frahm[2]

[1]ETH Zürich, [2]UNC Chapel Hill, [3]Microsoft
{jsch,pomarc}@inf.ethz.ch, {ezheng,jmf}@cs.unc.edu

Abstract. This work presents a Multi-View Stereo system for robust and efficient dense modeling from unstructured image collections. Our core contributions are the joint estimation of depth and normal information, pixelwise view selection using photometric and geometric priors, and a multi-view geometric consistency term for the simultaneous refinement and image-based depth and normal fusion. Experiments on benchmarks and large-scale Internet photo collections demonstrate state-of-the-art performance in terms of accuracy, completeness, and efficiency.

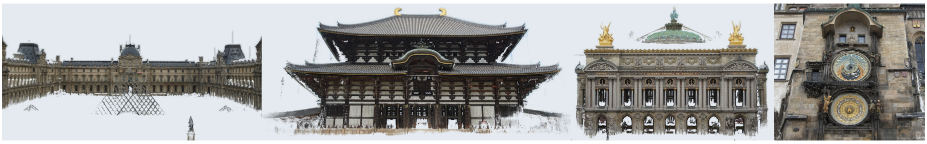

Fig. 1. Reconstructions for Louvre, Todai-ji, Paris Opera, and Astronomical Clock.

1 Introduction

Large-scale 3D reconstruction from Internet photos has seen a tremendous evolution in sparse modeling using Structure-from-Motion (SfM) [1–8] and in dense modeling using Multi-View Stereo (MVS) [9–15]. Many applications benefit from a dense scene representation, *e.g.*, classification [16], image-based rendering [17], localization [18], *etc.* Despite the widespread use of MVS, the efficient and robust estimation of accurate, complete, and aesthetically pleasing dense models in uncontrolled environments remains a challenging task. Dense pixelwise correspondence search is the core problem of stereo methods. Recovering correct correspondence is challenging even in controlled environments with known viewing geometry and illumination. In uncontrolled settings, *e.g.*, where the input consists of crowd-sourced images, it is crucial to account for various factors, such as heterogeneous resolution and illumination, scene variability, unstructured viewing geometry, and mis-registered views.

Our proposed approach improves the state of the art in dense reconstruction for unstructured images. This work leverages the optimization framework by Zheng *et al.* [14] to propose the following core contributions: (1) Pixelwise *normal estimation* embedded into an improved PatchMatch sampling scheme. (2) Pixelwise view selection using triangulation angle, incident angle, and image resolution-based *geometric priors*. (3) Integration of a "temporal" *view selection smoothness* term. (4) Adaptive window support through bilateral *photometric consistency* for improved occlusion boundary behavior. (5) Introduction of a multi-view *geometric consistency* term for simultaneous depth/normal estimation and image-based fusion. (6) Reliable depth/normal *filtering* and *fusion*. Outlier-free and accurate depth/normal estimates further allow for direct meshing of the resulting point cloud. We achieve state-of-the-art results on benchmarks (Middlebury [19], Strecha [20]). To demonstrate the advantages of our method in a more challenging setting, we process SfM models of a world-scale Internet dataset [5]. The entire algorithm is released to the public as an open-source implementation as part of [8] at https://github.com/colmap/colmap.

2 Related Work

Stereo methods have advanced in terms of accuracy, completeness, scalability, and benchmarking – from the minimal stereo setup with two views [21–24] to multi-view methods [9,10,14,15,25–28]. Furthermore, the joint estimation of semantics [29], dynamic scene reconstruction [30–34], and benchmarking [12,19,20, 23]. Our method performs MVS with pixelwise view selection for depth/normal estimation and fusion. Here, we only review the most related approaches, within the large body of research in multi-view and two-view stereo.

MVS leverages multiple views to overcome the inherent occlusion problems of two-view approaches [35–37]. Accordingly, view selection plays a crucial role in the effectiveness of MVS. Kang *et al.* [38] heuristically select the best views with minimal cost (usually 50%) for computing the depth of each pixel. Strecha *et al.* [39,40] probabilistically model scene visibility combined with a local depth smoothness assumption [39] in a Markov Random Field for pixelwise view selection. Different from our approach, their method is prohibitive in memory usage and does neither include normal estimation nor photometric and geometric priors for view selection. Gallup *et al.* [41] select different views and resolutions on a per-pixel basis to achieve a constant depth error. In contrast, our method simultaneously considers a variety of photometric and geometric priors improving upon the robustness and accuracy of the recently proposed depth estimation framework by Zheng *et al.* [14]. Their method is most closely related to our approach and is reviewed in more detail in Section 3.

MVS methods commonly use a fronto-parallel scene structure assumption. Gallup *et al.* [42] observed the distortion of the cost function caused by structure that deviates from this prior and combats it by using multiple sweeping directions deduced from the sparse reconstruction. Earlier approaches [43–45] similarly account for the surface normal in stereo matching. Recently, Bleyer *et al.* [46] use PatchMatch to estimate per-pixel normals to compensate for the distortion of the cost function. In contrast to these approaches, we propose to estimate normals not in isolation but also considering the photometric and geometric constraints guiding the matchabilty of surface texture and its accuracy. By probabilistically modeling the contribution of individual viewing rays towards reliable surface recovery, we achieve significantly improved depth and normal estimates.

Depth map fusion integrates multiple depth maps into a unified and augmented scene representation while mitigating any inconsistencies among individual estimates. Jancoseck *et al.* [28] fuses multiple depth estimates into a surface and, by evaluating visibility in 3D space, they also attempt to reconstruct parts that are not directly supported by depth measurements. In contrast, our method aims at directly maximizing the estimated surface support in the depth maps and achieves higher completeness and accuracy (see Section 5). Goesele *et al.* [47] propose a method that explicitly targets at the reconstruction from crowd-sourced images. They first select camera clusters for each surface and adjust their resolution to the smallest common resolution. For depth estimation, they then use the four most suitable images for each pixel. As already noted in Zheng *et al.* [14], this early pre-selection of reduced camera clusters may lead to less complete results and is sensitive to noise. Our method avoids this restrictive selection scheme by allowing dataset-wide, pixelwise sampling for view selection. Zach *et al.* [48] proposed a variational depth map formulation that enabled parallelized computation on the GPU. However, their volumetric approach imposes substantial memory requirements and is prohibitive for the large-scale scenes targeted by our method. Beyond these methods, there are several large-scale dense reconstruction and fusion methods for crowd-sourced images, *e.g.*, Furukawa *et al.* [10] and Gallup *et al.* [49,50], who all perform heuristic pre-selection of views, which leads to reduced completeness and accuracy as compared to our method.

3 Review of Joint View Selection and Depth Estimation

This section reviews the framework by Zheng *et al.* [14] to introduce notation and context for our contributions. Since their method processes each row/column independently, we limit the description to a single image row with l as the column index. Their method estimates the depth θ_l for a pixel in the reference image X^{ref} from a set of unstructured source images $\boldsymbol{X}^{\text{src}} = \{X^m \mid m = 1 \dots M\}$. The estimate θ_l maximizes the color similarity between a patch X_l^{ref} in the reference image and homography-warped patches X_l^m in non-occluded source images. The binary indicator variable $Z_l^m \in \{0,1\}$ defines the set of non-occluded source images as $\bar{\boldsymbol{X}}_l^{\text{src}} = \{X^m \mid Z_l^m = 1\}$. To sample $\bar{\boldsymbol{X}}_l^{\text{src}}$, they infer the probability that the reference patch X_l^{ref} at depth θ_l is visible at the source patch X_l^m using

$$P(X_l^m | Z_l^m, \theta_l) = \begin{cases} \frac{1}{NA} \exp\left(-\frac{(1-\rho_l^m(\theta_l))^2}{2\sigma_\rho^2}\right) & \text{if } Z_l^m = 1 \\ \frac{1}{N}\mathcal{U} & \text{if } Z_l^m = 0, \end{cases} \tag{1}$$

where $A = \int_{-1}^{1} exp\{-\frac{(1-\rho)^2}{2\sigma_\rho^2}\}d\rho$ and N is a constant canceling out in the inference. In the case of occlusion, the color distributions of the two patches are unrelated and follow the uniform distribution $\mathcal{U}$ in the range $[-1, 1]$ with probability density 0.5. Otherwise, ρ_l^m describes the color similarity between the reference and source patch based on normalized cross-correlation (NCC) using fronto-parallel homography warping. The variable σ_ρ determines a soft threshold for ρ_l^m on the reference patch being visible in the source image. The state-transition matrix from the preceding pixel $l-1$ to the current pixel l is $P(Z_l^m | Z_{l-1}^m) = \left(\begin{smallmatrix} \gamma & 1-\gamma \\ 1-\gamma & \gamma \end{smallmatrix}\right)$ and encourages spatially smooth occlusion indicators, where a larger γ enforces neighboring pixels to have more similar indicators. Given reference and source images $\boldsymbol{X} = \{X^{\text{ref}}, \boldsymbol{X}^{\text{src}}\}$, the inference problem then boils down to recover, for all L pixels in the reference image, the depths $\boldsymbol{\theta} = \{\theta_l \mid l = 1 \ldots L\}$ and the occlusion indicators $\boldsymbol{Z} = \{Z_l^m \mid l = 1 \ldots L, m = 1 \ldots M\}$ from the posterior distribution $P(\boldsymbol{Z}, \boldsymbol{\theta} | \boldsymbol{X})$ with a uniform prior $P(\boldsymbol{\theta})$. To solve the computationally infeasible Bayesian approach of first computing the joint probability

$$P(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}) = \prod_{l=1}^{L} \prod_{m=1}^{M} [P(Z_l^m | Z_{l-1}^m) P(X_l^m | Z_l^m, \theta_l)] \tag{2}$$

and then normalizing over $P(\boldsymbol{X})$, Zheng *et al.* use variational inference theory to develop a framework that is a variant of the generalized expectation-maximization (GEM) algorithm [51]. For the inference of $\boldsymbol{Z}$ in the hidden Markov-Chain, the forward-backward algorithm is used in the E step of GEM. PatchMatch-inspired [46] sampling serves as an efficient scheme for the inference of $\boldsymbol{\theta}$ in the M step of GEM. Their method iteratively solves for $\boldsymbol{Z}$ with fixed $\boldsymbol{\theta}$ and *vice versa* using interleaved row-/columnwise propagation. Full depth inference

$$\theta_l^{\text{opt}} = \underset{\theta_l^*}{\text{argmin}} \sum\nolimits_{m=1}^{M} P_l(m)(1 - \rho_l^m(\theta_l^*)) \tag{3}$$

has high computational cost if M is large as PatchMatch requires the NCC to be computed many times. The value $P_l(m) = \frac{q(Z_l^m=1)}{\sum_{m=1}^{M} q(Z_l^m=1)}$ denotes the probability of the patch in source image m being similar to the reference patch, while $q(\boldsymbol{Z})$ is an approximation of the real posterior $P(\boldsymbol{Z})$. Source images with small $P_l(m)$ are non-informative for the depth inference, hence Zheng *et al.* propose a Monte Carlo based approximation of θ_l^{opt} for view selection

$$\hat{\theta}_l^{\text{opt}} = \underset{\theta_l^*}{\text{argmin}} \frac{1}{|S|} \sum\nolimits_{m \in S} (1 - \rho_l^m(\theta_l^*)) \tag{4}$$

by sampling a subset of images $S \subset \{1 \ldots M\}$ from the distribution $P_l(m)$ and hence only computing the NCC for the most similar source images.

4 Algorithm

In this section, we describe our novel algorithm that leverages the optimization framework reviewed in the previous section. We first present the individual terms of the proposed likelihood function, while Section 4.6 explains their integration into the overall optimization framework.

4.1 Normal Estimation

Zheng *et al.* [14] map between the reference and source images using fronto-parallel homographies leading to artifacts for oblique structures [42]. In contrast, we estimate per-pixel depth θ_l and normals $\boldsymbol{n}_l \in \mathbb{R}^3, \|\boldsymbol{n}_l\| = 1$. A patch at $\boldsymbol{x}_l \in \mathbb{P}^2$ in the reference image warps to a source patch at $\boldsymbol{x}_l^m \in \mathbb{P}^2$ using $\boldsymbol{x}_l^m = \boldsymbol{H}_l \boldsymbol{x}_l$ with $\boldsymbol{H}_l = \boldsymbol{K}^m(\boldsymbol{R}^m - d_l^{-1}\boldsymbol{t}^m \boldsymbol{n}_l^T)\boldsymbol{K}^{-1}$. Here, $\boldsymbol{R}^m \in SO(3)$ and $\boldsymbol{t}^m \in \mathbb{R}^3$ define the relative transformation from the reference to the source camera frame. $\boldsymbol{K}$ and $\boldsymbol{K}^m$ denote the calibration of the reference and source images, respectively, and $d_l = \boldsymbol{n}_l^T \boldsymbol{p}_l$ is the orthogonal distance from the reference image to the plane at the point $\boldsymbol{p}_l = \theta_l \boldsymbol{K}^{-1}\boldsymbol{x}_l$.

Given no knowledge of the scene, we assume a uniform prior $P(\boldsymbol{N})$ in the inference of the normals $\boldsymbol{N} = \{\boldsymbol{n}_l \mid l = 1 \ldots L\}$. Estimating $\boldsymbol{N}$ requires to change the terms $P(X_l^m | Z_l^m, \theta_l)$ and $P_l(m)$ from Eqs. (1) and (4) to also depend on $\boldsymbol{N}$, as the color similarity ρ_l^m is now based on slanted rather than fronto-parallel homographies. Consequently, the optimal depth and normal are chosen as

$$(\hat{\theta}_l^{\mathrm{opt}}, \hat{\boldsymbol{n}}_l^{\mathrm{opt}}) = \underset{\theta_l^*, \boldsymbol{n}_l^*}{\mathrm{argmin}} \frac{1}{|S|} \sum\nolimits_{m \in S} (1 - \rho_l^m(\theta_l^*, \boldsymbol{n}_l^*)). \tag{5}$$

To sample unbiased random normals in PatchMatch, we follow the approach by Galliani *et al.* [15]. With the additional two unknown normal parameters, the number of unknowns per pixel in the M step of GEM increases from one to three. While this in theory requires PatchMatch to generate many more samples, we propose an efficient propagation scheme that maintains the convergence rate of depth-only inference. Since depth θ_l and normal $\boldsymbol{n}_l$ define a local planar surface in 3D, we propagate the depth $\theta_{l-1}^{\mathrm{prp}}$ of the intersection of the ray of the current pixel $\boldsymbol{x}_l$ with the local surface of the previous pixel $(\theta_{l-1}, \boldsymbol{n}_{l-1})$. This exploits first-order smoothness of the surface (*cf.* [52]) and thereby drastically speeds up the optimization since correct depths propagate more quickly along the surface. Moreover, different from the typical iterative refinement of normals using bisection as an intermediate step between full sweeps of propagations (*cf.* [15, 46]), we generate a small set of additional plane hypotheses at each propagation step. We observe that the current best depth and normal parameters can have the following states: neither of them, one of them, or both of them have the optimal solution or are close to it. By combining random and perturbed depths with current best normals and vice versa, we increase the chance of sampling the correct solution. More formally, at each step in PatchMatch, we choose the current best estimate for pixel l according to Eq. (4) from the set of hypotheses

$$\{(\theta_l, \boldsymbol{n}_l), (\theta_{l-1}^{\mathrm{prp}}, \boldsymbol{n}_{l-1}), (\theta_l^{\mathrm{rnd}}, \boldsymbol{n}_l), (\theta_l, \boldsymbol{n}_l^{\mathrm{rnd}}), (\theta_l^{\mathrm{rnd}}, \boldsymbol{n}_l^{\mathrm{rnd}}), (\theta_l^{\mathrm{prt}}, \boldsymbol{n}_l), (\theta_l, \boldsymbol{n}_l^{\mathrm{prt}})\}, \tag{6}$$

where θ_l^{rnd} and $\boldsymbol{n}_l^{\mathrm{rnd}}$ denote randomly generated samples. To refine the current parameters when they are close to the optimal solution, we perturb the current

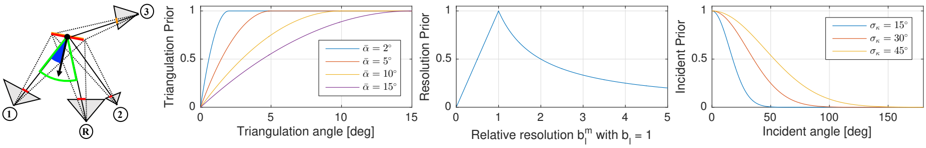

Fig. 2. Left: Illustration of geometric priors for reference view (R) and three source views (1-3). View 1 has similar resolution (red), and good triangulation (green) and incident angle (blue), while view 2 is oblique and has lower resolution. View 3 cannot see the patch. Right: Geometric prior likelihood functions with different parameters.

estimate as $\theta_l^{\text{prt}} = (1 \pm \epsilon)\theta_l$ and $\boldsymbol{n}_l^{\text{prt}} = \boldsymbol{R}_\epsilon \boldsymbol{n}_l$. The variable ϵ describes a small depth perturbation, and the rotation matrix $\boldsymbol{R}_\epsilon \in SO(3)$ perturbs the normal direction by a small angle subject to $\boldsymbol{p}_l^T \boldsymbol{n}_l^{\text{prt}} < 0$. Normal estimation improves both the reconstruction completeness and accuracy, while the new sampling scheme leads to both fast convergence and more accurate estimates (Section 5).

4.2 Geometric Priors for View Selection

This section describes how to incorporate geometric priors in the pixelwise view selection for improved robustness in particular for unstructured imagery. On a high level, the proposed priors encourage the sampling of source images with sufficient baseline (*Triangulation Prior*), similar resolution (*Resolution Prior*), and non-oblique viewing direction (*Incident Prior*). In contrast to prior work (*e.g.* [10, 47, 49]), which decouples inference and per-image geometric priors by pre-selecting source images, we integrate geometric priors on a per-pixel basis into the inference. The motivation for per-pixel geometric priors is similar to inferring per-pixel occlusion indicators $\boldsymbol{Z}$. Since the pre-selection of source images is based on a sparse and therefore incomplete scene representation, the selected source views are often sub-optimal. Occlusion boundaries, triangulation angles, relative image resolution, and incident angle can vary significantly between a single pair of reference and source images (Fig. 2). Incorporating geometric priors in addition to the photometric occlusion indicators $\boldsymbol{Z}$ leads to a more comprehensive and robust pixelwise view selection. In the following, we detail the proposed priors and explain their integration into the optimization framework.

Triangulation Prior. Zheng *et al.* [14] sample source images purely based on color similarity. Consequently, the more similar the reference patch is to the source patch, the higher the selection probability in the view sampling. Naturally, image pairs with small viewpoint change, which coincides with small baseline, have high color similarity. However, image pairs with zero baseline do not carry information for depth inference, because reconstructed points can arbitrarily move along the viewing ray without changing the color similarity. Pure photometric view selection favors to sample these uninformative views. To eliminate this degenerate case, we calculate the triangulation angle $\alpha_l^m = \cos^{-1} \frac{(\boldsymbol{p}_l - \boldsymbol{c}^m)^T \boldsymbol{p}_l}{\|\boldsymbol{p}_l - \boldsymbol{c}^m\| \|\boldsymbol{p}_l\|}$ with $\boldsymbol{c}^m = -(\boldsymbol{R}^m)^T \boldsymbol{t}^m$ and $\alpha_l^m \in [0, \pi)$ between two intersecting viewing rays as a measure of the stability of the reconstructed point $\boldsymbol{p}_l$. Empirically, we choose the following likelihood function $P(\alpha_l^m) = 1 - \frac{(\min(\bar{\alpha}, \alpha_l^m) - \bar{\alpha})^2}{\bar{\alpha}^2}$ to describe how informative a source image is for reconstructing the correct point. Intuitively, this function assigns low likelihood to source images for which the triangulation angle is below an *a priori* threshold $\bar{\alpha}$. Otherwise, no additional view selection preference is imposed (see Fig. 2).

Resolution Prior. Unstructured datasets usually contain images captured by a multitude of camera types under diverse viewing geometry. As a consequence, images capture scene objects in a wide range of resolutions. To avoid under- and oversampling in computing ρ_l^m, the patches in the reference and source image should have similar size and shape [47]. Similar size is favorable as it avoids comparing images captured at vastly different resolutions, *e.g.*, due to different zoom factors or distance to the object. Similar shape avoids significantly distorted source patches caused by different viewing directions. In the case of different shape, areas within the same source patch have different sampling rates. An approximate measure of the relative size and shape between the reference and source patch is $\beta_l^m = \frac{b_l}{b_l^m} \in \mathbb{R}^+$, where b_l and b_l^m denote the areas covered by the corresponding patches. In our implementation, the reference patch is always square. If the size and shape of the patches is similar, β_l^m is close to the value 1. To quantify the similarity in resolution between two images, we propose the likelihood function $P(\beta_l^m) = \min(\beta_l^m, (\beta_l^m)^{-1})$ and integrate it into $P_l(m)$. Note that, at increased computational cost, undersampling could alternatively be handled by adaptive resampling of the source image patch.

Incident Prior. The inferred per-pixel normals provide geometric constraints on the solution space that we encode in the form of a prior. The estimated plane restricts the possible space of source camera locations and orientations. By construction, the camera location can only lie in the positive half-space defined by the plane $(\theta_l, \boldsymbol{n}_l^m)$, while the camera viewing direction must face towards the opposite normal direction. Otherwise, it is geometrically impossible for the camera to observe the surface. To satisfy this geometric visibility constraint, the incident angle of the source camera $\kappa_l^m = \cos^{-1} \frac{(\boldsymbol{p}_l - \boldsymbol{c}^m)^T \boldsymbol{n}_l^m}{\|\boldsymbol{p}_l - \boldsymbol{c}^m\| \|\boldsymbol{n}_l^m\|}$ with $\kappa_l^m \in [0, \pi)$ must be in the interval $0 \leq \kappa_l^m < \frac{\pi}{2}$. In our method, the likelihood function $P(\kappa_l^m) = \exp(-\frac{\kappa_l^{m\,2}}{2\sigma_\kappa^2})$ encodes the belief in whether this geometric constraint is satisfied. This associates some belief with a view even in the case where $\kappa_l^m \geq \frac{\pi}{2}$. The reason for this is, that in the initial inference stage, the variables θ_l and $\boldsymbol{n}_l^m$ are unknown and hence the geometric constraints are likely not yet correct.

Integration. Fig. 2 visualizes the geometric priors, and Fig. 4 shows examples of specific priors over all reference image pixels. We integrate the priors into the inference as additional terms in the Monte-Carlo view sampling distribution

$$P_l(m) = \frac{q(Z_l^m = 1) q(\alpha_l^m) q(\beta_l^m) q(\kappa_l^m)}{\sum_{m=1}^{M} q(Z_l^m = 1) q(\alpha_l^m) q(\beta_l^m) q(\kappa_l^m)}, \tag{7}$$

where $q(\alpha_l^m), q(\beta_l^m), q(\kappa_l^m)$ are approximations during the variational inference, in the sense that they minimize the KL-divergence to the real posterior [53]. The distributions need no normalization in the inference because we solely use them as modulators for the sampling distribution $P_l(m)$. This formulation assumes statistical independence of the individual priors as a simplifying approximation, which makes the optimization feasible using relatively simple models for well-understood geometric relations. Intuitively, non-occluded images with sufficient baseline, similar resolution, and non-oblique viewing direction are favored in the view selection. Section 5 evaluates the priors in detail and shows how they improve the reconstruction robustness especially for unstructured datasets.

4.3 View Selection Smoothness

The graphical model associated with the likelihood function in Eq. (2) uses state-transition probabilities to model spatial view selection smoothness for neighboring pixels in the propagation direction. Due to the interleaved inference using alternating propagation directions, Z_l^m suffers from oscillation, leading to striping effects as shown in Figure 5. To reduce the oscillation effect of $Z_{l,t}^m$ in iteration t, we insert an additional "temporal" smoothness factor into the graphical model. In this new model, the state of $Z_{l,t}^m$ depends not only on the state of its neighboring pixel $l-1$ but also on its own state in the previous iteration $t-1$. The temporal state-transition is defined as $P(Z_{l,t}^m | Z_{l,t-1}^m) = \left(\begin{smallmatrix} \lambda_t & 1-\lambda_t \\ 1-\lambda_t & \lambda_t \end{smallmatrix}\right)$, where a larger λ_t enforces greater temporal smoothness during the optimization. In fact, as the optimization progresses from $t = 1 \ldots T$, the value of the estimated $Z_{l,t-1}^m$ should stabilize around the optimal solution. Therefore, we adaptively increase the state-transition probability as $\lambda_t = \frac{t}{2T} + 0.5$, *i.e.*, the inferred $Z_{l,t}^m$ in iterations $t = 1$ and $t = T-1$ have maximal and minimal influence on the final value $Z_{l,T}^m$, respectively. The two state-transitions are jointly modeled as

$$P(Z_{l,t}^m | Z_{l-1,t}^m, Z_{l,t-1}^m) = P(Z_{l,t}^m | Z_{l-1,t}^m) P(Z_{l,t}^m | Z_{l,t-1}^m). \tag{8}$$

Fig. 5 shows the evolution of $Z_{l,t}^m$ during the optimization and demonstrates the reduced oscillation, which effectively also leads to less noisy view sampling.

4.4 Photometric Consistency

Zheng *et al.* [14] employ NCC to compute the color similarity ρ_l^m. NCC is statistically optimal for Gaussian noise but is especially vulnerable to producing blurred depth discontinuities [54]. Inspired by [46, 55], we diminish these artifacts by using a bilaterally weighted adaption of NCC. We compute ρ_l^m between a reference patch $\boldsymbol{w}_l$ at $\boldsymbol{x}_l$ with a corresponding source patch $\boldsymbol{w}_l^m$ at $\boldsymbol{x}_l^m$ as

$$\rho_l^m = \frac{\mathrm{cov}_w(\boldsymbol{w}_l, \boldsymbol{w}_l^m)}{\sqrt{\mathrm{cov}_w(\boldsymbol{w}_l, \boldsymbol{w}_l) \ \mathrm{cov}_w(\boldsymbol{w}_l^m, \boldsymbol{w}_l^m)}} \tag{9}$$

where $\mathrm{cov}_w(\boldsymbol{x}, \boldsymbol{y}) = E_w(\boldsymbol{x} - E_w(\boldsymbol{x}))\ E_w(\boldsymbol{y} - E_w(\boldsymbol{y}))$ is the weighted covariance and $E_w(\boldsymbol{x}) = \sum_i w_i x_i / \sum_i w_i$ is the weighted average. The per-pixel weight $w_i = \exp(-\frac{\Delta g_i}{2\sigma_g^2} - \frac{\Delta x_i}{2\sigma_x^2})$ indicates the likelihood that a pixel i in the local patch belongs to the same plane as its center pixel at l. It is a function of the grayscale color distance $\Delta g_i = |g_i - g_l|$ and the spatial distance $\Delta x_i = \|\boldsymbol{x}_i - \boldsymbol{x}_l\|$, whose importance is relatively scaled by the Gaussian dispersion σ_g and σ_x. By integrating the bilaterally weighted NCC into the term $P(X_l^m|Z_l^m, \theta_l, \boldsymbol{n}_l)$, our method achieves more accurate results at occlusion boundaries, as shown in Section 5.

4.5 Geometric Consistency

MVS typically suffers from gross outliers due to noise, ambiguities, occlusions, *etc.* In these cases, the photometric consistency for different hypotheses is ambiguous as large depth variations induce only small cost changes. Spatial smoothness constraints can often reduce but not fully eliminate the resulting artifacts. A popular approach to filter these outliers is to enforce multi-view depth coherence through left-right consistency checks as a post-processing step [15,46].

In contrast to most approaches, we integrate multi-view geometric consistency constraints into the inference to increase both the completeness and the accuracy. Similar to Zhang *et al.* [56], we infer the best depth and normal based on both photometric and geometric consistency in multiple views. Since photometric ambiguities are usually unique to individual views (except textureless surfaces), exploiting the information from multiple views can often help to pinpoint the right solution. We compute the geometric consistency between two views as the forward-backward reprojection error $\psi_l^m = \|\boldsymbol{x}_l - \boldsymbol{H}_l^m \boldsymbol{H}_l \boldsymbol{x}_l\|$, where $\boldsymbol{H}_l^m$ denotes the projective backward transformation from the source to the reference image. It is composed from the source image estimates $(\theta_l^m, \boldsymbol{n}_l^m)$ interpolated at the forward projection $\boldsymbol{x}_l^m = \boldsymbol{H}_l \boldsymbol{x}_l$. Intuitively, the estimated depths and normals are consistent if the reprojection error ψ_l^m is small. Due to computational constraints, we cannot consider the occlusion indicators in the source image for the backward projection. Hence, to handle occlusion in the source image, we employ a robustified geometric cost in $\xi_l^m = 1 - \rho_l^m + \eta \min(\psi_l^m, \psi_{\max})$ using $\eta = 0.5$ as a constant regularizer and $\psi_{\max} = 3\text{px}$ as the maximum forward-backward reprojection error. Then, the optimal depth and normal is chosen as

$$(\hat{\theta}_l^{\text{opt}}, \hat{\boldsymbol{n}}_l^{\text{opt}}) = \underset{\theta_l^*, \boldsymbol{n}_l^*}{\operatorname{argmin}} \frac{1}{|S|} \sum\nolimits_{m \in S} \xi_l^m(\theta_l^*, \boldsymbol{n}_l^*). \tag{10}$$

The geometric consistency term is modeled as $P(\theta_l, \boldsymbol{n}_l|\theta_l^m, \boldsymbol{n}_l^m)$ in the likelihood function, and Section 4.6 shows how to integrate its inference into the overall optimization framework. Experiments in Section 5 demonstrate how this formulation improves both the accuracy and the completeness of the results.

4.6 Integration

This section contextualizes the individual terms of the proposed algorithm by explaining their integration into the overall optimization framework [14]. The joint likelihood function $P(\boldsymbol{X}, \boldsymbol{Z}, \ \boldsymbol{\theta}, \boldsymbol{N})$ of our proposed algorithm is defined as

$$\prod_{l=1}^{L} \prod_{m=1}^{M} [P(Z_{l,t}^m|Z_{l-1,t}^m, Z_{l,t-1}^m) P(X_l^m|Z_l^m, \theta_l, \boldsymbol{n}_l) P(\theta_l, \boldsymbol{n}_l|\theta_l^m, \boldsymbol{n}_l^m)]$$

over the input images $\boldsymbol{X}$, the occlusion indicators $\boldsymbol{Z}$, the depths $\boldsymbol{\theta}$, the normals $\boldsymbol{N}$, and is composed of several individual terms. First, the spatial and temporal smoothness term $P(Z_{l,t}^m|Z_{l-1,t}^m, Z_{l,t-1}^m)$ (Section 4.3) enforces spatially smooth occlusion maps with reduced temporal oscillation during the optimization. Second, the photometric consistency term $P(X_l^m|Z_l^m, \theta_l, \boldsymbol{n}_l)$ uses bilateral NCC (Section 4.4) and a slanted plane-induced homography (Section 4.1) to compute the color similarity ρ_l^m between the reference and source images. Third, the geometric consistency term $P(\theta_l, \boldsymbol{n}_l|\theta_l^m, \boldsymbol{n}_l^m)$ to enforce multi-view consistent depth and normal estimates. The photometric and geometric consistency terms are computed using Monte-Carlo view sampling from the distribution $P_l(m)$ in Eq. (7). The distribution encourages the sampling of non-occluded source images with informative and non-degenerate viewing geometry (Section 4.2).

Analog to Zheng *et al.* [14], we factorize the real posterior $P(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{N}|\boldsymbol{X})$ in its approximation $q(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{N}) = q(\boldsymbol{Z})q(\boldsymbol{\theta}, \boldsymbol{N})$ [53]. Furthermore, for tractability, we constrain $q(\boldsymbol{\theta}, \boldsymbol{N})$ to the family of Kronecker delta functions $q(\theta_l, \boldsymbol{n}_l) = \delta(\theta_l{=}\theta_l^*, \boldsymbol{n}_l{=}\boldsymbol{n}_l^*)$. Variational inference then aims to infer the optimal member of the family of approximate posteriors to find the optimal $\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{N}$. The validity of using GEM for this type of problem has already been shown in [14, 51]. To infer $q(Z_{l,t}^m)$ in iteration t of the E step of GEM, we employ the forward-backward algorithm as

$$q(Z_{l,t}^m) = \frac{1}{A} \overrightarrow{m}(Z_{l,t}^m) \overleftarrow{m}(Z_{l,t}^m) \tag{11}$$

with $\overrightarrow{m}(Z_{l,t}^m)$ and $\overleftarrow{m}(Z_{l,t}^m)$ being the recursive forward and backward messages

$$\overrightarrow{m}(Z_l^m) = P(X_l^m|Z_l^m, \theta_l, \boldsymbol{n}_l) \sum\nolimits_{Z_{l-1}^m} \overrightarrow{m}(Z_{l-1}^m) P(Z_{l,t}^m|Z_{l-1,t}^m, Z_{l,t-1}^m) \tag{12}$$

$$\overleftarrow{m}(Z_l^m) = \sum\nolimits_{Z_{l+1}^m} \overleftarrow{m}(Z_{l+1}^m) P(X_{l+1}^m|Z_{l+1}^m, \theta_{l+1}, \boldsymbol{n}_{l+1}) P(Z_{l,t}^m|Z_{l+1,t}^m, Z_{l,t-1}^m) \tag{13}$$

using an uninformative prior $\overrightarrow{m}(Z_0^m) = \overrightarrow{m}(Z_{L+1}^m) = 0.5$. The variable $q(Z_{l,t}^m)$ together with $q(\alpha_l^m), q(\beta_l^m), q(\kappa_l^m)$ determine the view sampling distribution $P_l(m)$ used in the M step of GEM as defined in Eq. (7). The M step uses PatchMatch propagation and sampling (Section 4.1) for choosing the optimal depth and normal parameters over $q(\theta_l, \boldsymbol{n}_l)$. Since geometrically consistent depth and normal inference is not feasible for all images simultaneously due to memory constraints, we decompose the inference in two stages. In the first stage, we estimate initial depths and normals for each image in the input set $\boldsymbol{X}$ according to Eq. (5). In the second stage, we use coordinate descent optimization to infer geometrically consistent depths and normals according to Eq. (10) by keeping all images but the current reference image as constant. We interleave the E and M step in both stages using row- and column-wise propagation. Four propagations in all directions denote a sweep. In the second stage, a single sweep defines a coordinate descent step, *i.e.*, we alternate between different reference images after propagating through the four directions. Typically, the first stage converges after $I_1 = 3$ sweeps, while the second stage requires another $I_2 = 2$ sweeps through the entire image collection to reach a stable state. We refer the reader to the supplementary material for an overview of the steps of our algorithm.

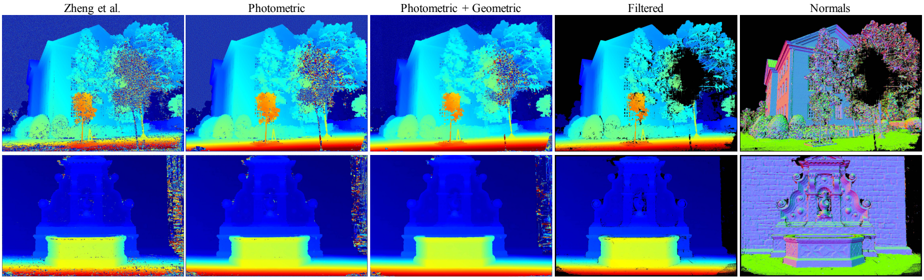

Fig. 3. Reconstruction results for *South Building* [29] and *Fountain* [20]. From left to right: Depth map by Zheng *et al.* [14], then ours only with the photometric term, with the photometric and geometric terms, and the final filtered depth and normal maps.

4.7 Filtering and Fusion

After describing the depth and normal inference, this section proposes a robust method to filter any remaining outliers, *e.g.*, in textureless sky regions. In addition to the benefits described previously, the photometric and geometric consistency terms provide us with measures to robustly detect outliers at negligible computational cost. An inlier observation should be both photometrically and geometrically stable with support from multiple views. The sets

$$\mathcal{S}_l^{\text{pho}} = \{\boldsymbol{x}_l^m \mid q(Z_l^m) > \bar{q}_Z\} \tag{14}$$

$$\mathcal{S}_l^{\text{geo}} = \{\boldsymbol{x}_l^m \mid q(\alpha_l^m) \geq \bar{q}_\alpha, q(\beta_l^m) \geq \bar{q}_\beta, q(\kappa_l^m) > \bar{q}_\kappa, \psi_l^m < \psi_{\max}\} \tag{15}$$

determine the photometric and geometric support of a reference image pixel $\boldsymbol{x}_l$. To satisfy both constraints, we define the effective support of an observation as $\mathcal{S}_l = \{\boldsymbol{x}_l^m \mid \boldsymbol{x}_l^m \in \mathcal{S}_l^{\text{pho}}, \boldsymbol{x}_l^m \in \mathcal{S}_l^{\text{geo}}\}$ and filter any observations with $|\mathcal{S}_l| < s$. In all our experiments, we set $s = 3$, $\bar{q}_Z = 0.5$, $\bar{q}_\alpha = 1$, $\bar{q}_\beta = 0.5$, and $\bar{q}_\kappa = P(\kappa = 90°)$. Figs. 3 and 6 show examples of filtered depth and normal maps.

The collection of support sets $\mathcal{S}$ over the observations in all input images defines a directed graph of consistent pixels. In this graph, pixels with sufficient support are nodes, and directed edges point from a reference to a source image pixel. Nodes are associated with depth and normal estimates and, together with the intrinsic and extrinsic calibration, edges define a projective transformation from the reference to the source pixel. Our fusion finds clusters of consistent pixels in this graph by initializing a new cluster using the node with maximum support $|\mathcal{S}|$ and recursively collecting connected nodes that satisfy three constraints. Towards this goal, we project the first node into 3D to obtain the location $\boldsymbol{p}_0$ and normal $\boldsymbol{n}_0$. For the first constraint, the projected depth $\tilde{\theta}_0$ of the first node into the image of any other node in the cluster must be consistent with the estimated depth θ_i of the other node such that $\frac{|\tilde{\theta}_0 - \theta_i|}{\tilde{\theta}_0} < \epsilon_\theta$ (*cf.* [57]). Second, the normals of the two must be consistent such that $1 - \boldsymbol{n}_0^T \boldsymbol{n}_i < \epsilon_n$. Third, the reprojection error ψ_i of $\boldsymbol{p}_0$ w.r.t. the other node must be smaller than $\bar{\psi}$. Note

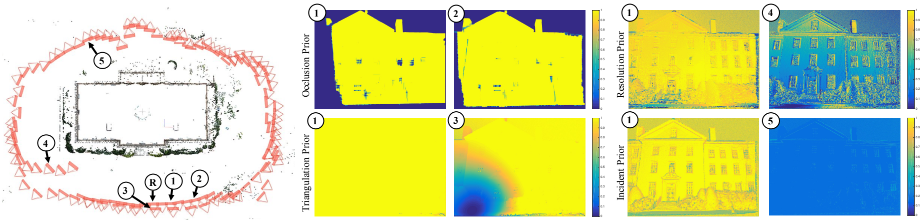

Fig. 4. Photometric and geometric priors for *South Building* dataset [29] between reference image (R) and each two selected source images (1-5).

that the graph can have loops, and therefore we only collect nodes once. In addition, multiple pixels in the same image can belong to the same cluster and, by choosing $\bar{\psi}$, we can control the resolution of the fused point cloud. When there is no remaining node that satisfies the three constraints, we fuse the cluster's elements, if it has at least three elements. The fused point has median location $\hat{\boldsymbol{p}}_j$ and mean normal $\boldsymbol{n}_j$ over all cluster elements. The median location is used to avoid artifacts when averaging over multiple neighboring pixels at large depth discontinuities. Finally, we remove the fused nodes from the graph and initialize a new cluster with maximum support $|\mathcal{S}|$ until the graph is empty. The resulting point cloud can then be colored (*e.g.* [58]) for visualization purposes and, since the points already have normals, we can directly apply meshing algorithms (*e.g.* Poisson reconstruction [59]) as an optional step.

5 Experiments

This section first demonstrates the benefits of the proposed contributions in isolation. Following that, we compare to other methods and show state-of-the-art results on both low- and high-resolution benchmark datasets. Finally, we evaluate the performance of our algorithm in the challenging setting of large-scale Internet photo collections. The algorithm lends itself for massive parallelization on the row- and column-wise propagation and the view level. In all our experiments, we use a CUDA implementation of our algorithm on a Nvidia Titan X GPU. We set $\gamma = 0.999$, leading to an average of one occlusion indicator state change per 1000 pixels. Empirically, we choose $\sigma_\rho = 0.6$, $\bar{\alpha} = 1°$, and $\sigma_k = 45°$.

Components. This paragraph shows the benefits of the individual components in isolation based on the *South Building* dataset [29], which consists of 128 unstructured images with a resolution of 7MP. We obtain sparse reconstructions using SfM [5]. For each reference view, we use all 127 images as source views with an average runtime of 50s per sweep. *Normal Estimation*: Fig. 3 shows depth maps using fronto-parallel homographies (1st column) and with normal estimation (2nd to 5th columns), which leads to increased completeness and accuracy for depth inference of oblique scene elements, such as the ground. In addition, our method estimates more accurate normals than standard PatchMatch (Fig. 5(b)). Due to the proposed PatchMatch sampling scheme, our algorithm requires the

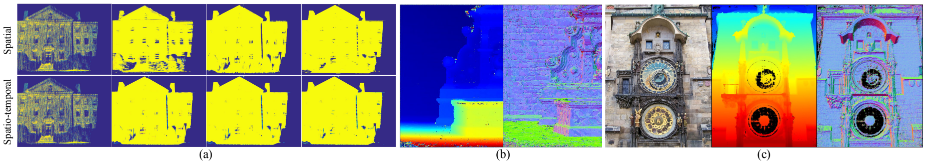

Fig. 5. (a) Comparison of spatial smoothness term [14] with our proposed spatial and temporal smoothness term for the occlusion variables $\boldsymbol{Z}$. Algorithm starts from the left with the first sweep and is followed by consecutive sweeps to the right. (b) Estimated depths and normals using standard PatchMatch propagation (*cf.* Fig. 3 for ours). (c) Reference image with filtered depths and normals for crowd-sourced images

same number sweeps to converge and only ≈ 25% more runtime due to more hypotheses as compared to Zheng *et al.* [14], who only estimate per-pixel depths. *Geometric Priors*: Fig. 4 demonstrates the benefit of each geometric prior. We show the likelihood functions for the reference view against one representative source image. For all priors, we observe varying likelihood within the same source image, underlining the benefit of pixel-wise view selection. The priors correctly downweigh the influence of source images with small triangulation angle, low resolution, or occluded views. *Selection Smoothness* Fig. 5(a) shows that our temporal smoothness term effectively mitigates the oscillation of the pure spatial smoothness term. While the occlusion variables in the formulation by Zheng *et al.* [14] oscillate depending on the propagation direction, in our method they quickly converge in a stable state leading to more stable view sampling. *Geometric Consistency*: Fig. 3 demonstrates improved completeness when incorporating the geometric consistency term, and it also allows to reliably detect outliers for practically outlier-free filtered results. To measure the quantitative impact of our contributions, we obtain benchmark results by omitting a single component or combinations of components from the formulation (Table 1). We observe that each component is important to achieve the overall accuracy and completeness of our method. For further evaluations and impressions of the benefits of our method, we strongly encourage the reader to view the supplementary material.

Benchmarks. The *Middlebury* benchmark [23] consists of the *Dino* and *Temple* models captured at 640x480 under varying settings (*Full*, *Ring*, *Sparse*). For each reference image, we use all views as source images at a runtime of ≈ 40s per view for the *Full* models with ≈ 300 images. We achieve excellent accuracy and completeness on both models[1]. Specifically, using the standard settings, we rank 1st for *Dino Full* (tied) and *Dino Sparse*, while achieving competitive scores for the *Temple* (4th for *Full*, 8th for *Ring*). Note that our method performs best for higher resolutions, as normal estimation needs large patch sizes. Also, we use basic Poisson meshing [59], underlining the highly accurate and outlier-free depth/normal estimates produced by our method. The *Strecha* benchmark [20] consists of high-resolution images with ground-truth, and we follow the evaluation protocol of Hu and Mordohai [60]. Fig. 3 shows outputs for the *Fountain*

[1] Full results online at http://vision.middlebury.edu/mview/eval/.

Table 1. Strecha benchmark [20] with reported values from [60]. Ratio of pixels with error less than 2cm and 10cm. Ours w/o normals (\N), geom. priors (\P), temp. smoothness (\S), geom. consistency (\G), bilateral NCC (\B), and with all components.

		[14]	[60]	[9]	[62]	[61]	[28]	[15]	\N	\P	\S	\B	\PSB	\G	Ours
Fountain	2cm	0.769	0.754	0.731	0.712	0.732	0.824	0.693	0.799	0.824	0.825	0.826	0.817	0.804	**0.827**
	10cm	0.929	0.930	0.838	0.832	0.822	0.973	0.838	0.937	0.972	0.973	0.973	0.965	0.949	**0.975**
Herzjesu	2cm	0.650	0.649	0.646	0.220	0.658	**0.739**	0.283	0.673	0.686	0.688	0.690	0.688	0.679	0.691
	10cm	0.844	0.848	0.836	0.501	0.852	0.923	0.455	0.901	0.928	0.927	0.929	0.921	0.907	**0.931**

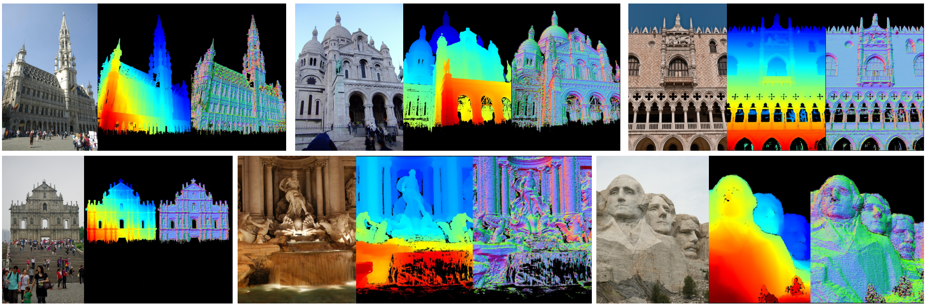

Fig. 6. Reference image with filtered depths and normals for crowd-sourced images.

dataset and, Table 1 lists the results quantifying both the accuracy and completeness. To maintain comparability against Zheng *et al.* [14], we evaluate our raw depth maps against the ground-truth. We produce significantly more accurate and complete results than Zheng *et al.*, and we outperform the other methods in 3 of 4 categories, even though the results of [28,60,61] are evaluated based on the projection of a 3D surface obtained through depth map fusion.

Internet Photos. We densely reconstruct models of 100M Internet photos released by Heinly *et al.* [5,8] using a single machine with 4 Nvidia Titan X. We process the 41K images at a rate of 70s per view using 2 threads per GPU and finish after 4.2 days in addition to the 6 days needed for sparse modeling using SfM. Whenever we reach the GPU memory limits, we select the most connected source images ranked by the number of shared sparse points. Usually, this limit is reached for ≈ 200 images, while image sizes vary from 0.01MP to 9MP. The fusion and filtering steps consume negligible runtime. Fig. 1 shows fused point clouds, Figs. 6 and 5(c) show depth/normal maps, and the supplementary material provides more results and comparisons against [9,10,47].

6 Conclusion

This work proposes a novel algorithm for robust and efficient dense reconstruction from unstructured image collections. Our method estimates accurate depth and normal information using photometric and geometric information for pixelwise view selection and for image-based fusion and filtering. We achieve state-of-the-art results on benchmarks and crowd-sourced data.

References

1. Schaffalitzky, F., Zisserman, A.: Multi-view matching for unordered image sets, or How do I organize my holiday snaps? In: ECCV. (2002)
2. Snavely, N., Seitz, S., Szeliski, R.: Photo tourism: exploring photo collections in 3d. ACM Trans. Graphics (2006)
3. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S., Szeliski, R.: Building rome in a day. In: ICCV. (2009)
4. Frahm, J.M., Fite-Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.H., Dunn, E., Clipp, B., Lazebnik, S., Pollefeys, M.: Building Rome on a Cloudless Day. In: ECCV. (2010)
5. Heinly, J., Schönberger, J.L., Dunn, E., Frahm, J.M.: Reconstructing the World* in Six Days *(As Captured by the Yahoo 100 Million Image Dataset). In: CVPR. (2015)
6. Zheng, E., Wu, C.: Structure from motion using structure-less resection. In: ICCV. (2015)
7. Schönberger, J.L., Radenović, F., Chum, O., Frahm, J.M.: From single image query to detailed 3d reconstruction. In: CVPR. (2015)
8. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR. (2016)
9. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. In: CVPR. (2007)
10. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Towards internet-scale multi-view stereo. In: CVPR. (2010)
11. Bailer, C., Finckh, M., Lensch, H.P.: Scale robust multi view stereo. In: ECCV. (2012)
12. Shan, Q., Adams, R., Curless, B., Furukawa, Y., Seitz, S.M.: The visual turing test for scene reconstruction. In: 3DV. (2013)
13. Shan, Q., Curless, B., Furukawa, Y., Hernandez, C., Seitz, S.M.: Occluding contours for multi-view stereo. In: CVPR. (2014)
14. Zheng, E., Dunn, E., Jojic, V., Frahm, J.M.: Patchmatch based joint view selection and depthmap estimation. In: CVPR. (2014)
15. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: ICCV. (2015)
16. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. In: Comm. ACM. (2013)
17. Chen, S.E., Williams, L.: View interpolation for image synthesis. In: Conference on Computer graphics and interactive techniques. (1993)
18. Forster, C., Pizzoli, M., Scaramuzza, D.: Air-ground localization and map augmentation using monocular dense reconstruction. In: IROS. (2013)
19. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: CVPR. (2006)
20. 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: CVPR. (2008)
21. Intille, S.S., Bobick, A.F.: Disparity-space images and large occlusion stereo. (1994)
22. Kanade, T., Okutomi, M.: A stereo matching algorithm with an adaptive window: Theory and experiment. PAMI (1994)

23. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In: IJCV. (2002)
24. Rhemann, C., Hosni, A., Bleyer, M., Rother, C., Gelautz, M.: Fast cost-volume filtering for visual correspondence and beyond. In: CVPR. (2011)
25. Campbell, N., Vogiatzis, G., Hernández, C., Cipolla, R.: Using multiple hypotheses to improve depth-maps for multi-view stereo. In: ECCV. (2008)
26. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Manhattan-world stereo. In: CVPR. (2009)
27. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Reconstructing building interiors from images. In: CVPR. (2009)
28. Jancosek, M., Pajdla, T.: Multi-view reconstruction preserving weakly-supported surfaces. In: CVPR. (2011)
29. Hane, C., Zach, C., Cohen, A., Angst, R., Pollefeys, M.: Joint 3d scene reconstruction and class segmentation. In: CVPR. (2013)
30. Tung, T., Nobuhara, S., Matsuyama, T.: Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In: ICCV. (2009)
31. Ji, D., Dunn, E., Frahm, J.M.: 3d reconstruction of dynamic textures in crowd sourced data. In: ECCV. (2014)
32. Oswald, M., Cremers, D.: A convex relaxation approach to space time multi-view 3d reconstruction. In: ICCV Workshops. (2013)
33. Martin-Brualla, R., Gallup, D., Seitz, S.M.: 3d time-lapse reconstruction from internet photos. In: ICCV. (2015)
34. Radenović, F., Schönberger, J.L., Ji, D., Frahm, J.M., Chum, O., Matas, J.: From dusk till dawn: Modeling in the dark. In: CVPR. (2016)
35. Yang, Q., Wang, L., Yang, R., Stewénius, H., Nistér, D.: Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling. PAMI (2009)
36. Sun, J., Li, Y., Kang, S.B., Shum, H.Y.: Symmetric stereo matching for occlusion handling. In: CVPR. (2005)
37. Zitnick, C.L., Kanade, T.: A cooperative algorithm for stereo matching and occlusion detection. PAMI (2000)
38. Kang, S.B., Szeliski, R., Chai, J.: Handling occlusions in dense multi-view stereo. In: CVPR. (2001)
39. Strecha, C., Fransens, R., Van Gool, L.: Wide-baseline stereo from multiple views: a probabilistic account. In: CVPR. (2004)
40. Strecha, C., Fransens, R., Van Gool, L.: Combined depth and outlier estimation in multi-view stereo. In: CVPR. (2006)
41. Gallup, D., Frahm, J.M., Mordohai, P., Pollefeys, M.: Variable baseline/resolution stereo. In: CVPR. (2008)
42. Gallup, D., Frahm, J.M., Mordohai, P., Yang, Q., Pollefeys, M.: Real-time plane-sweeping stereo with multiple sweeping directions. In: CVPR. (2007)
43. Burt, P., Wixson, L., Salgian, G.: Electronically directed focal stereo. In: ICCV. (1995)
44. Birchfield, S., Tomasi, C.: Multiway cut for stereo and motion with slanted surfaces. In: ICCV. (1999)
45. Zabulis, X., Daniilidis, K.: Multi-camera reconstruction based on surface normal estimation and best viewpoint selection. In: 3DPVT. (2004)
46. Bleyer, M., Rhemann, C., Rother, C.: Patchmatch stereo-stereo matching with slanted support windows. In: BMVC. (2011)

47. Goesele, M., Snavely, N., Curless, B., Hoppe, H., Seitz, S.M.: Multi-view stereo for community photo collections. In: CVPR. (2007)
48. Zach, C.: Fast and high quality fusion of depth maps. In: 3DPVT. (2008)
49. Gallup, D., Pollefeys, M., Frahm, J.M.: 3d reconstruction using an n-layer heightmap. In: Pattern Recognition. (2010)
50. Zheng, E., Dunn, E., Raguram, R., Frahm, J.M.: Efficient and scalable depthmap fusion. In: BMVC. (2012)
51. 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)
52. Heise, P., Jensen, B., Klose, S., Knoll, A.: Variational patchmatch multiview reconstruction and refinement. In: CVPR. (2015)
53. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer Verlag (2006)
54. Hirschmüller, H., Scharstein, D.: Evaluation of stereo matching costs on images with radiometric differences. PAMI (2009)
55. Yoon, K.J., Kweon, I.S.: Locally adaptive support-weight approach for visual correspondence search. In: CVPR. (2005)
56. Zhang, G., Jia, J., Wong, T.T., Bao, H.: Recovering consistent video depth maps via bundle optimization. In: CVPR. (2008)
57. Merrell, P., Akbarzadeh, A., Wang, L., Mordohai, P., Frahm, J.M., Yang, R., Nistér, D., Pollefeys, M.: Real-time visibility-based fusion of depth maps. In: CVPR. (2007)
58. Waechter, M., Moehrle, N., Goesele, M.: Let there be color! large-scale texturing of 3d reconstructions. In: ECCV. (2014)
59. Kazhdan, M., Hoppe, H.: Screened poisson surface reconstruction. ACM Trans. Graphics (2013)
60. Hu, X., Mordohai, P.: Least commitment, viewpoint-based, multi-view stereo. In: 3DIMPVT. (2012)
61. Tylecek, R., Sara, R.: Refinement of surface mesh for accurate multi-view reconstruction. IJVR (2010)
62. Zaharescu, A., Boyer, E., Horaud, R.: Topology-adaptive mesh deformation for surface evolution, morphing, and multiview reconstruction. PAMI (2011)