Abstract—Accurate recovery of 3D geometrical surfaces from calibrated 2D multi-view images is a fundamental yet active research area in computer vision. Despite the steady progress in multi-view stereo reconstruction, many existing methods are still limited in recovering fine-scale details and sharp features while suppressing noises, and may fail in reconstructing regions with less textures. To address these limitations, this paper presents a Detail-preserving and Content-aware Variational (DCV) multi-view stereo method, which reconstructs the 3D surface by alternating between reprojection error minimization and mesh denoising. In reprojection error minimization, we propose a novel inter-image similarity measure, which is effective to preserve fine-scale details of the reconstructed surface and builds a connection between guided image filtering and image registration. In mesh denoising, we propose a content-aware \( \ell_p \)-minimization algorithm by adaptively estimating the \( p \) value and regularization parameters. Compared with conventional isotropic mesh smoothing approaches, the proposed method is much more promising in suppressing noise while preserving sharp features. Experimental results on benchmark datasets demonstrate that our DCV method is capable of recovering more surface details, and obtains cleaner and more accurate reconstructions than state-of-the-art methods. In particular, our method achieves the best results among all published methods on the Middlebury dino ring and dino sparse datasets in terms of both completeness and accuracy.

Index Terms—Multi-view stereo, reprojection error, feature-preserving, \( \ell_p \) minimization, mesh denoising.

I. INTRODUCTION

MULTI-VIEW stereo (MVS), which aims at reconstructing the 3D geometric surface from a set of calibrated 2D images, is an active yet challenging task in computer vision. With the ubiquitous use of modern digital cameras, drones and smartphones, a large number of images could be easily captured in our daily life. MVS provides a promising way to reconstruct both indoor and outdoor scenes from multiple view images, and have attracted considerable interests in urban reconstruction, entertainment, augmented and mixed reality, and so on [1]–[3]. In the last decade, driven by the release of several benchmark datasets [4], [5], various MVS algorithms have been proposed, which significantly boost the reconstruction accuracy and completeness [6]–[9].

Z. Li is with School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, and Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong. Email:cszli@hotmail.com

K. Wang and W. Zuo are with School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. Email:wangkq@hit.edu.cn, cszmzuo@gmail.com

D. Meng is with Institute for Information and System Sciences, Xi’an Jiaotong University, Xi’an, China. Email:dymeng@mail.xjtu.edu.cn

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

According to how the 3D model is represented, the existing MVS algorithms can be grouped into three categories, i.e., point-cloud-based [9], [10], volumetric-based [6], [11], [12], and mesh-based [2], [3], [7], [13]–[16]. Among them, mesh-based method is much less memory demanding for representing high-resolution surface, and is more suitable for recovering small structures in large-scale reconstruction. In this work, we investigate the variational mesh-based MVS, which performs 3D reconstruction by minimizing an energy functional that consists of a fidelity term and a regularization term. However, due to the limitations of the existing fidelity or regularization terms, many mesh-based methods [7], [8], [14], [15] often fail in accurately recovering fine-scale details and sharp features of surface:

(1) The fidelity term models the cost for the correspondence between multiple images. In [2], [3], [7], [14], [15], [17], a functional is proposed to minimize the reprojection error measured between observed and predicted image, and the consistency of multiple images is assessed based on some isotropic measures, e.g., Zero-mean Normalized Cross-correlation (ZNCC) [2], [7], [14] and Sum of Square Difference (SSD) [3], [15], [17]. These isotropic measures, however, may smooth the fine-scale details, degrading the accuracy and visual quality of the reconstruction results. Anisotropic measures, which have been adopted in volumetric-based MVS [12] and binocular stereo [18]–[20], may be a remedy to this. To the best of our knowledge, by far no study has been done on anisotropic measure for mesh-based MVS.

(2) The regularization term imposes smooth and sparse priors on parametric surface for denoising. One conventional approach is to employ some first-order mesh smoothing operators [21], [22] and theirs approximations [13], [14], [23]. Several high-order derivatives have also been investigated in [7], [8], [17], and the anisotropic methods, e.g., two-step normal filtering [25], bilateral normal filtering [26], and \( \ell_0 \) mesh denoising [27], have been developed for mesh denoising. Most mesh denoising methods impose the same regularization method and parameters to all surface meshes. However, given the form of a specific regularizer, its parameter may vary for different surface meshes. Therefore, better performance can be achieved by joint estimation of the denoised mesh and the regularization parameters.

Other cues, e.g., silhouettes and shading, can be incorporated into the variational mesh-based MVS to improve the reconstruction accuracy. When the background is relatively clean, the silhouettes of an object can be easily segmented and fused to ensure that the reconstructed surface preserves the protrusions [13], [23]. Nevertheless, for the outdoor sce-
nario where the background is difficult to segmented, the incorporation of silhouettes cannot guarantee to be effective in recovering fine-scale surface details and sharp features. For the objects with smooth shaded surfaces and uniform albedo, the high-frequency surface details can be estimated via normal fields to refine the output of MVS algorithms [28], [29]. However, it is difficult to uniquely determine the geometry from the normal fields. Therefore, some surface constraints, e.g., visual hull, are essential to accurately recover the surface shape.

For better reconstructing fine-scale details and sharp features, we propose a novel variational mesh-based MVS method by developing a novel fidelity term and a novel regularization term. In image filtering and binocular stereo, guided filtering [30] has been widely adopted as a detail-preserving smoothing operator. Motivated by this, we analyze the connection between ZNCC and guided filtering, and indicate that guided filtering can be interpreted as one step of ZNCC-based registration with equality constraint on local variance. Based on their connection, we improve ZNCC by enforcing the closeness between local variance, and propose a novel inter-image similarity measure in the reprojection error functional framework. For the regularization term, we deploy a family of hyper-Laplacian distributions to model the mesh sharpness and adaptively estimate its specific regularization parameters and $p$ value for each 3D model, resulting in the proposed content-aware mesh denoising model. We then propose an alternating minimization algorithm to adaptively estimate the denoised mesh, $p$ value, and regularization parameters.

With our fidelity and regularization terms, a detail-preserving and content-aware variational (DCV) MVS model is presented in this paper. As shown in Fig. 1, to reconstruct the 3D model, the proposed method alternatively performs two steps: (i) reprojection error minimization with detail-preserved similarity measure and (ii) content-aware mesh denoising. The contribution of this work is two-fold:

- An inter-image similarity measure is proposed to preserve fine-scale details of the reconstructed surface. The proposed similarity measure builds a connection between guided image filtering and image registration, and gains promising edge-preserving performance.
- A content-aware $\ell_p$-minimization algorithm is proposed for mesh denoising based on the observation that surface gradient generally fits well to a hyper-Laplacian distribution. By adaptively estimating a suitable $p$ value and regularization parameters, our algorithm works very well in mesh smoothing while preserving sharp features.

Extensive experimental results on benchmark datasets validate the superiority of our DCV method in accurate 3D reconstruction. Moreover, DCV achieves the best results among all published methods on the Middlebury *dino ring* and *dino sparse* datasets in terms of both completeness and accuracy.

The paper is organized as follows. Section II introduces the related work. Section III briefly reviews the concept of reprojection error and its minimization. Section IV presents the pipeline of our method and its two major components, i.e., detail-preserving similarity measure and content-aware mesh denoising, respectively. Section V presents the experimental results. Finally, the paper is concluded in Section VI.

### II. Related Work

A comprehensive survey on MVS can be found in [4], [5]. In this work, we focus on variational mesh-based MVS methods where the energy functional is generally formulated as follows:

$$E(S) = E_{im}(S) + \lambda E_{reg}(S)$$

where $S$ denotes the reconstructed surface, $E_{im}$ is the data fidelity term, $E_{reg}$ is the regularization term, and $\lambda$ is the tradeoff parameter. In the following, we briefly review the most related work on the design of fidelity and regularization terms for variational mesh-based MVS.

**Data fidelity.** In some early works [8], [13], photometric consistency is measured by comparing projections of 3D surface points (or planar patches nearby surface point) with the corresponding neighboring images. $E_{im}$ is then represented as a summation of photometric consistency over all the mesh vertices independent to the observed images. Nevertheless, this measure is sensitive to the projective distortion occurred in the high curvature regions of objects. Recent progress on the reprojection error minimization framework [3], [7], [14], [17] attempts to compare the observed and predicted values of pixels generated from the reconstructed surface. Since reprojection error is evaluated in image space, the projective distortion can be well addressed and the reprojection error is more coherent to the image data. These methods are originally designed for the level-set based surface construction [24], [31] and then extended to the mesh-based methods [2], [3], [7], [14], [17]. In [3], [15], [17], for each reference image, a single predicted image is estimated based on the estimated surface (including both surface geometry and textures) by averaging the color from the visible views. Another line of methods [2], [7], [14] estimate multiple predicted images for each reference image by reprojecting the neighboring images to reference image via the estimated surface. Based on multiple predictions from neighboring images, the reprojection errors are more robust to outliers. In particular, Vu et al. [7] proposed a high-accuracy MVS pipeline consisting of two stages, i.e. depth maps fusion and mesh-based surface refinement, and achieved state-of-the-art performance.

Similarity measure is vital under the reprojection error minimization framework. In previous works, differentiable and isotropic similarity measures are widely used, such as SSD [15], [17], [31] and ZNCC [2], [7], [14], [24]. SSD is simple and efficient, but is sensitive to illumination variation. To enforce robustness to illumination affine variation, ZNCC-based similarity measure is developed and has been widely used in MVS [2], [7], [8], [13], [14], [24]. Unfortunately, ZNCC tends to over-smooth the sharp features of surface, and is limited in recovering fine-scale details. One may replace the isotropic measures with edge-aware anisotropic ones. Anisotropic methods have been successfully adopted in binocular stereo vision [18]–[20]; however, to the best our knowledge they have not been applied in variational mesh-based MVS. One possible reason is that, in stereo vision
the anisotropic methods, e.g., guided filtering [18], [19], are used to filter discrete disparity space images (DSI) and cannot be directly adopted in the reprojection error minimization framework where variational measure is necessary. In this work, we will investigate the connection between guided filtering and ZNCC-based variational registration, and improve ZNCC to reconstruct fine-scale details of the 3D model.

Surface regularization. The regularization term is introduced to preserve the smoothness and sharpness of the reconstructed 3D surface. To this end, there are mainly two categories of approaches, i.e., surface smoothing and denoising. For mesh smoothing, classic discrete Laplace-Beltrami operator \( \Delta \) can be employed to smooth each mesh vertex \( v \) to the mean of its neighborhood [13], [14]. Although this operator can maintain good distribution of vertices along the surface, it prefers to shrink the surface area and collapse the small components. To avoid surface shrinkage, higher order surface information can be taken into account [7], [8], [17]. In [17], a normal field filtering is utilized to suppress the noise on normals and then update the positions of vertices. In [7], a thin-plate energy is employed to measure the total curvature of the surface for penalizing strong bending and handling artificial shrinkage of small components. In [8], the first and second-order Laplace operators are combined to prevent the shrinkage bias of the surface. These smoothing methods, however, tend to suppress mesh noise at the cost of smoothing sharp features.

Mesh denoising, which aims to remove the noise or spurious details while preserving sharp edge and corner features, can be classified into three sub-categories: the first one is based on bilateral filtering on vertices [32]; the second one combines normal filtering and vertex position updating [25], [26]; and the third one is based on optimizing an \( \ell_0 \) norm based non-convex energy functional [27]. Although these methods show superiority over isotropic smoothing, they generally adopt some generic prior and ignore the specificity of each individual mesh to be reconstructed.

Recently, regularization models in gradient and dictionary domains have been successfully applied to reconstruct the detailed surface [33]–[35]. In [33], the dictionary is learned on the vertices and vertex connectivity of mesh by minimizing the distance metric between the input points and the reconstructed mesh via \( \ell_2,n \) optimization. The dictionary-based sparse model is then used for surface reconstruction from point cloud. In [34], cross-derivative constraint is integrated in derivative compressed sensing for surface reconstruction from gradient field without increasing the sampling rate. Harker et al. [35] presented an efficient framework based on the Sylvester equation for efficient regularized surface reconstruction from noisy gradient fields. The purposes of these methods are to reconstruct 3D surface from point cloud or gradient field, while in this paper we aim to reconstruct 3D surface from RGB images captured from different views. Moreover, we model the mesh sharpness with a family of hyper-Laplacian distributions, and propose a content-aware mesh denoising model to adaptively estimate the denoised mesh, and model parameters during the denoising procedure.

III. PREREQUISITES: REPROJECTION ERROR AND ITS MINIMIZATION

Let \( S \subset \mathbb{R}^3 \) denote a reconstructed surface of interested object, \( B \subset \mathbb{R}^3 \) stand for its background, and \( I_i : \Omega_i \subset \mathbb{R}^2 \to \mathbb{R}^d \) denote the observed (input) image in camera \( i \) (\( d = 1 \) for grayscale images, and \( d = 3 \) for color images). Let \( \pi_i : \mathbb{R}^3 \to \Omega_i \) be the perspective projection which projects a 3D point \( x \) to a 2D pixel \( p \). In image formation, the observed image only records the visible (i.e., unoccluded) part of a real scene, which includes both the object of interest and the irrelevant background. As shown in Fig. 2(a), \( S_i \) is the visible part of object surface for camera \( i \). We define \( S_{i,j} \) as the shared visible surface of camera \( i \) and camera \( j \). Denote by \( \{I_{j,k} | k = 1, 2, \cdots, K_j \} \) the set of neighbouring images of \( I_i \), and by \( \hat{I}_{i,j,B,S} \) the predicted image of \( I_i \) from \( I_{j,B} \). Then \( \hat{I}_{i,S} = \{\hat{I}_{i,j,B,S} | k = 1, 2, \cdots, K_j \} \) is defined as the set of predicted images of \( I_i \) from its neighbouring images via the surface, and \( \hat{I}_{i,B} \) is the predicted image of \( I_i \) via the background part. With the desired 3D reconstruction of object \( S \) and background \( B \), it is natural to assume that the images predicted by 3D object and background models should be similar to the observed image. Therefore, the minimization framework of reprojection errors adopts the following functional [3], [15], [17], [31]:

\[
E_{im}(S) = \sum_{i} \left[ \int_{\Omega_i} g_F(I_i, \hat{I}_{i,S})(p)dp + \int_{\Omega_i \setminus \Omega_i} g_B(I_i, \hat{I}_{i,B})(p)dp \right],
\]

(2)

where \( \pi_i \circ S_i \) denotes the projection of surface \( S_i \) onto \( I_i \), the reprojection error \( g_F \) measures the similarity between the observed image and its predicted images from neighbouring images, and the reprojection error \( g_B \) measures the similarity between the observed image and predicted background image.

The predicted images can be computed via rendering surface and background. As shown in Fig. 2(a), the elements of...
image set $\hat{I}_{i,S}$ are defined based on stereo pairs: $\hat{I}_{i,h,S}$ can be computed by first projecting one neighbouring image $I_h$ onto $S$ and then projecting into image space of camera $i$. The valid definition domain for $\hat{I}_{i,h,S}$ is the projection of shared visible surface $S_{i,h}$, i.e., $\pi_i \circ S_{i,h}$. By counting all the neighbouring images of $I_i$, $g_F$ is defined as:

$$g_F(I_i, \hat{I}_{i,h,S}) = \sum_m m(I_i, \hat{I}_{i,h,S}) = \sum_m m_{i,h},$$

where $m$ is a similarity measure of two pixels in a small squared window centered on $p$. The definition domain of predicted image $\hat{I}_B$ of $I_i$ via background is defined by $\Omega_i - \pi_i \circ S_i$, i.e., the supplementary set of $\pi_i \circ S_i$. To simplify the computation, we assume that the background is uniformly black, and this can be implemented by segmenting silhouettes from observed images. Let $d_i(x)$ be the vector pointing from the center of camera $i$ to 3D point $x$, $n(x)$ be the outward normal of surface $S$ on point $x$, $x_{i,z}$ be the depth of $x$ with respect to camera $i$. Based on the fact that $d \mathbf{p} = -d_i(x) \cdot n(x)/x_{i,z}^2 ds$ (see [17] for details), with simple algebra, Eq. (2) can be rewritten as an integral over the surface by counting only the visible points:

$$E_{\text{int}}(S) = \sum_i \int_{S} -\frac{d_i(x) \cdot n(x)}{x_{i,z}^2} (g_F(x) - g_B(x)) \Lambda_{i,S}(x) ds,$$

where $\Lambda_{i,S} : \mathbb{R}^3 \rightarrow [0,1]$ is the visibility function which equals to 1 if $x$ is visible from camera $i$ and 0 otherwise, $g_F(x) = \sum_k m(I_i(\pi_i \circ x), \hat{I}_{i,h,S}(\pi_i \circ x))$, and $g_B(x) = g_B(I_i(\pi_i \circ x), \hat{I}_B(\pi_i \circ x))$.

The functional of reprojection errors in Eq. (4) can be reformulated by mesh-based discrete representation. Let’s parametrize surface $S$ to a triangle mesh $M$ with a set of indexes $\mathcal{V} = \{v_1, v_2, \ldots, v_{\nu}\}$ and a set of triangular faces $\mathcal{F} = \{f_1, f_2, \ldots, f_{\nu}\}, f_i \in \mathcal{V} \times \mathcal{V} \times \mathcal{V}$. The geometric embedding of a triangle mesh into $\mathbb{R}^3$ is specified by associating each vertex to a 3D position. Let $x_i$ denote the position of a vertex $v_i$. As shown in Fig. 2(b), over each triangular face, points are parametrized using barycentric coordinates $\mathbf{x}(u) := \mathbf{u}(u, v) \in T = \{(u, v)|u \in [0, 1], v \in [0, 1 - u]\}$. The energy functional on the triangle mesh is formulated as follows:

$$E_{\text{int}}(M) = \sum_i \int_{f_i} G(\mathbf{x}) \Lambda_{i,S}(\mathbf{x}) d\mathbf{u},$$

where $\mathbf{x} = \mathbf{x}(u)$, $G(\mathbf{x}) = -(d_i(\mathbf{x})/x_{i,z}^2) \cdot [g_F(\mathbf{x}) - g_B(\mathbf{x})]$, $N_j$ and $A_j$ are the normal and area of the triangle $f_j$, respectively, and the term $d\mathbf{u} = 2A_j ds$ corresponds to the unit surface area element in the triangle mesh.

Denote by $V_k$ the velocity vector on the $k$-th vertex of the mesh. The energy functional in Eq. (5) can be optimized by using gradient descent over all the vertices of the mesh. According to [3], [15], [17], [31], the evolution equation for gradient descent flow is:

$$\begin{align*}
\frac{d\mathbf{x}_i}{dt} &= -\sum_j A_j \int_T \nabla \cdot G(\mathbf{x}) (1 - u) dudv, \tag{6}
\end{align*}$$

and $M_{k}^{\text{int}}$ is defined as:

$$\sum_i \sum_j 2A_j N_j \int_T \nabla \cdot G(\mathbf{x}) (1 - u - v) dudv,$$

and $M_{k}^{\text{horiz}}$ is defined as:

$$\sum_{H_{k,j}} \int_{u \in [0,1]} \int_{v \in [0,1]} \left[ G(T(\mathbf{d}(y))) - G(\mathbf{d}(y)) \right] \wedge H_{k,j} (1 - u) dudv,$$

where $A_k$ is the summation of the area of the triangles connecting vertex $v_k$, $\wedge$ is the cross product, $H_{k,j}$ is the vector such that $[x_{k}, x_{k} + H_{k,j}]$ is the edge of the triangular face $f_j$ generating the horizon, $y$ is defined as $y = x_k + \alpha H_{k,j}$, and $T(\mathbf{x})$ is the terminator of $x$, $\mathbf{d}(y)$ is the vector connecting camera center $i$ and $y$. The definitions of horizon and terminator are illustrated in Fig. 2(c). $M_{k}^{\text{int}}$ is the gradient for the vertex (interior point) that does not change its visibility state, and $M_{k}^{\text{horiz}}$ is the gradient for the vertex that exhibits strong changes in visibility during the evolution.

The term $M_{k}^{\text{horiz}}$ is used to confine the horizons of the surface in different cameras. Although its influence will be considerably decreased by the introduction of surface regularization, this term is very useful for persevering thin protruding structures on the border between object and background. This naturally corresponds to a silhouette constraint [13], [23]. To save the computational cost, the horizontal term can be simplified in the implementation. When the background has constant color and the silhouette can be easily segmented, the SSD can be used to measure the consistency of model and silhouettes in the form of $g_B$. When the silhouette is hard to be segmented, we simply remove the term and reconstruct the 3D model only based on the term $M_{k}^{\text{int}}$. 
The term $M^k$ is crucial to the reconstruction quality. To evolve the current surface $S$, we should estimate the derivative of $g_F(x)$. As shown in Eq. (3), for image $I_1$ and $I_{1,k,S}$, the reprojection error is measured by the similarity measure $m_{i,j}$. In the following, we remove the dependency on $k$ to simplify the notations. As shown in [24], the gradient of $m_{i,j}$ with respect to an infinitesimal vector displacement $\delta S$ of 3D surface point $x$ can be computed using the chain rule:

$$\lim_{\epsilon \to 0} \frac{\partial m_{i,j}(S + \epsilon \delta S)}{\partial \epsilon} = \lim_{\epsilon \to 0} \int_{S \cap \Omega_0} \frac{\partial m(i, \hat{I}_{i,j,S})}{\partial I_{i,j,S}} \times \frac{\partial I_{i,j,S}}{\partial \delta x} \times \frac{\partial \delta x}{\partial \epsilon} dp_i,$$

and we have

$$\nabla g_F(S)(x) = -\sum_{j \neq f} m(i, \hat{I}_{i,j,S}) \times \frac{\partial I_{i,j,S}}{\partial \delta x} \times \frac{\partial \delta x}{\partial \epsilon},$$

where $p_i$ and $p_j$ are the pixel positions in $I_i$ and $I_{1,k,j}$, respectively. $\eta_{i,j}: \Omega \to S \subset \mathbb{R}^3$ is the back-projection which projects pixels from camera $i$ onto the surface, $\eta$ is the Kronecker symbol which cancels the gradient computation in the region outside the shared visible surface of both cameras. When the surface moves, the predicted image tends to be changed. Hence, the variation of reprojection errors involves the derivative of the similarity measure with respect to its second argument $\hat{I}_{i,j,S}$, i.e., $\partial_2 m_{i,j}$, as shown in the first derivative term of the right part of Eq. (10).

IV. PROPOSED METHOD

Following the variational mesh-based MVS framework, our DCV model consists of two terms, i.e., the data fidelity term of the right part of Eq. (10) and the surface regularization term $E_S \pi^r$ of $g_{ff}$.

In the following, we remove the dependency on $k$ to simplify the notations. Following the variational mesh-based MVS framework, our model can be expressed as:

$$E_{\text{reg}} = \| S - S^{\text{k+0.5}} \|^2 + \lambda\eta E_{\text{reg}}(S),$$

where $\eta$ is the stepsize.

Step 1. Gradient Descent. Given the current estimate $S^k$, the gradient descent algorithm is adopted to minimize the data fidelity term $E_{\text{im}}$:

$$S^{k+0.5} = S^k - \eta \partial E_{\text{im}}(S)/\partial S,$$

where $\eta$ is the stepsize.

Step 2. Surface Denoising. Given $S^{k+0.5}$, the reconstructed surface $S$ is further refined by solving the following mesh denoising problem:

$$S^{k+1} = \arg \min_S \frac{1}{2} \| S - S^{k+0.5} \|^2 + \lambda \eta E_{\text{reg}}(S).$$

A. Detail-preserving Inter-image Similarity Measure

The similarity measure $m(I_1, \hat{I}_{i,j,S})$ is critical for the minimization of reprojection error between $I_i$ and its predicted image $\hat{I}_{i,j,S}$. In the variational framework, it is desirable that $m(I_1, \hat{I}_{i,j,S})$ is differentiable. Among the existing similarity measures [37]–[39], ZNCC is the most commonly used one due to the following advantages: (1) it is robust to inter-image affine illumination variation; and (2) its derivative can be efficiently computed. However, the isotropic ZNCC treats all pixels equally and tends to flatten the details of surface. In this section, we first review the derivative of ZNCC-based similarity measure and then propose a detail-preserving similarity measure based on the principle of guided filtering.

1) ZNCC-based Similarity Measure: The ZNCC measure is defined as follows:

$$m(I_1, I_2)(p) = v_{1,2}(p)/\sqrt{v_1(p)v_2(p)},$$

where $v_1$, $v_2$ and $v_{1,2}$ are given by

$$v_i(p) = G_\sigma \ast I_i^2(p)/\omega(p) - m_i^2(p) + \epsilon,$$

$$v_{1,2}(p) = G_\sigma \ast I_1I_2(p)/\omega(p) - m_1m_2(p),$$

$$\mu_i(p) = G_\sigma \ast I_i(p)/\omega(p).$$

where $G_\sigma$ is a Gaussian kernel with standard deviation $\sigma$. $\omega$ is a normalization coefficient accounting for the shape of support domain: $\omega(p) = \int_{Q_i} G_\sigma(p-q)dq$, and the small positive constant $\epsilon$ is introduced to prevent the denominator from being zero. The derivative of $m(I_1, I_2)$ with respect to any entry of $I_2$ at pixel position $p$ has the following form [24]:

$$\partial_2 m(p) = \alpha(p)I_1(p) + \beta(p)I_2(p) + \gamma(p),$$

$$\alpha(p) = G_\sigma \ast \frac{-1}{\omega \sqrt{v_1v_2}},$$

$$\beta(p) = G_\sigma \ast \frac{m}{\omega v_2}(p),$$

$$\gamma(p) = G_\sigma \ast (\frac{\mu}{\omega \sqrt{v_1v_2}} - \frac{\mu_1m_2}{\omega v_2})(p).$$

Note that the variation at $p$ also tends to affect the similarity measure of its neighboring positions. Actually, if we restrict ZNCC in a local square window of size $w$, the variation of pixel $p$ will affect its entire neighbouring pixels in the region of size $2w \times 2w$.

2) Connection Between ZNCC and Guided Image Filtering:

Let $I_1$ be the input image, and $I_2$ be the guidance image. The principle of guided image filtering is to assume a local linear transformation between filtering output $Q$ and a guidance image $I_2$ for any pixel $p$ belonging to a local window of size $w_k$ ($k$ is the center of window):

$$J(p) = a(p)I_2(p) + b(p).$$

By minimizing the difference between $J$ and $I_1$, we can obtain parameters $a(p)$ and $b(p)$:

$$a(p) = (v_{1,2}/v_2)(p),$$

$$b(p) = (\mu_1 - q_2)(p) = (\mu_1 - v_{1,2}\mu_2/v_2)(p).$$
Note that the tolerance $\epsilon$ in Eq. (14) can also be included to penalize large $\alpha(p)$ in (22) and (23). The role of $\epsilon$ in the guided filter is similar to the range variance $\sigma^2$ in the bilateral filter, which determines the edge patch that should be preserved. Finally, the filtering output $Q$ has the following form:

$$Q(p) = \frac{G_\phi \star a}{\omega}(p)I_2(p) + \frac{G_\phi \star b}{\omega}(p).$$

(24)

We can then have an interesting connection between the derivatives of reprojection errors and guided image filtering. Based on (13), (18)-(20), Eq. (17) can be reformulated as:

$$\partial_2 m(p) = (G_\phi \star -\frac{1}{\omega \sqrt{v_1^2 + v_2^2}}(p))I_1(p) + (G_\phi \star \frac{v_1}{\omega \sqrt{v_1^2 + v_2^2}}(p))I_2(p) + G_\phi \star \frac{(v_2 \mu_1 - v_1 \mu_2)}{\omega v_2}(p).$$

(25)

Based on (22)-(23), Eq. (24) can be rewritten as:

$$Q(p) = (G_\phi \star \frac{v_1}{\omega v_2}(p))I_2(p) + (G_\phi \star \frac{(v_2 \mu_1 - v_1 \mu_2)}{\omega v_2})(p).$$

(26)

Let $v_1(p) = v_2(p)$, Eq. (25) becomes:

$$\partial_2 m(p) + (G_\phi \star \frac{1}{\omega v_2}(p))I_1(p) = (G_\phi \star \frac{v_1}{\omega v_2}(p))I_2(p) + G_\phi \star \frac{(v_2 \mu_1 - v_1 \mu_2)}{\omega v_2}(p).$$

(27)

Suppose that $v_2(p)$ varies more slowly than $G_\phi$ and $I_1(p)$ in spatial domain. Then, based on the fact $(G_\phi \star \frac{1}{\omega}(p)) = 1$, Eq. (27) can be approximately rewritten as:

$$v_2 \partial_2 m(p) + I_1(p) \approx (G_\phi \star \frac{v_1}{\omega v_2}(p))I_2(p) + G_\phi \star \frac{(v_2 \mu_1 - v_1 \mu_2)}{\omega v_2}(p).$$

(28)

Note that the right sides of Eq. (26) and Eq. (28) are the same. Therefore, we have:

$$I_1(p) + v_2 \partial_2 m(p) = Q(p).$$

(29)

and guided image filtering can be approximately interpreted as one step of variational image registration of $I_1(p)$ and $I_2(p)$ with constraint $v_1(p) = v_2(p)$ and stepsize $v_2$.

Our assumption that $v_2(p)$ varies more slowly than $G_\phi$ and $I_1(p)$ is plausible. From the definition in Eq. (14) and Eq. (16), one can see that $v_2(p)$ is obtained based on the smoothed images of $I_2(p)$ and $I_2^2(p)$, and thus it is reasonable to assume that $v_2(p)$ varies more slowly than $I_1(p)$. The Gaussian smoothing operator $G_\phi$ is introduced to smooth the noise caused by image acquisition and mis-registration, while $v_2(p)$ is believed to carry salient structure and edge features of the image. Thus, it is reasonable to assume that $v_2(p)$ varies more slowly than $G_\phi$ (otherwise it is impossible to reconstruct salient structure based on the smoothed images $I_1(p)$ and $I_2(p)$). Finally, based on the definition of $\omega$: $\omega = G_\phi \star 1(p)$, we can directly have $(G_\phi \star \frac{1}{\omega}(p)) = 1$.

3) Detail-preserving Similarity Measure: Motivated by the connection between guided image filtering and image registration, and to enhance the edge preservation ofZNCC-based similarity measure, under the warm assumption discussed in

Section IV-A2, we modify the ZNCC in Eq. 13 by adding a term to enforce the constraint $v_1(p) = v_2(p)$:

$$\hat{m}(p) = m(p) + \kappa \frac{v_1}{v_2} ||v_1 - v_2||^2.$$  

(30)

Then the derivative of Eq. (30) is:

$$\partial \hat{m}(p) = \alpha I_1(p) + \beta I_2(p) + \gamma(p) + \kappa G_\phi \star \left( \frac{(v_2 - v_1)(I_2(p) - \mu_2)}{v_1^2} \right)(p).$$

(31)

where the value of $\mu_2$ varies along with $G_\phi$ around the neighbours of $p$, $\kappa$ is a tradeoff parameter to adjust the influence of the variance constraint. In practice, we initialize $\kappa$ with a small value in the beginning of surface evolution, and gradually increase it until convergence. By Eq. (31), the predicted image $\hat{I}_{s,5}$ is implicitly set as the guidance image. As shown in Fig. 3 and Fig. 4, the proposed similarity measure can recover the fine-scale details and largely extend the edge preservation capability of the original isotropic measure.

B. Content-aware Mesh Denoising via $\ell_p$-norm Minimization

Let matrix $X_0 = (x_{0i})$ $\in \mathbb{R}^{n \times 3}$ store the positions of the $n$ vertices of the reconstructed noisy surface mesh $S^{k+0.5}$, and matrix $X = (x_i)$ $\in \mathbb{R}^{n \times 3}$ store the positions of the $n$ vertices of the noisy-free surface mesh $S$. The mesh denoising problem in Eq. (12) can be reformulated as:

$$\min_X \frac{1}{2} ||X - X_0||^2 + \lambda R(X).$$

(32)

In this section, we first present the proposed content-aware model based on the MAP framework, and then propose an
alternating minimization algorithm for content-aware mesh denoising.

1) MAP-based Mesh Denoising with Hyper-Laplacian Prior: Denote by \( q(X) \) the prior on the sharpness of noise-free mesh, and by \( q(X_0|X) \) the likelihood of noisy mesh. The MAP framework estimates \( X \) by maximizing a posterior probability \( q(X|X_0) \propto q(X_0|X)q(X) \). By assuming that the noise is additive white Gaussian noise with standard deviation \( \sigma \), the likelihood of noisy mesh can be modeled as:

\[
q(X_0|X, \sigma^2) = \prod_i \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(X_i - X_0)^2}{2\sigma^2}\right). \tag{33}
\]

For surface mesh, the edge-based discrete Laplacian operator \( D \in \mathbb{R}^{m \times n} \) proposed in He et al. [27] can be adopted for computing surface gradients (where \( m \) is the number of edges in the mesh). In image restoration, it has been empirically verified that the natural image gradients generally follow a heavy-tailed distribution and can be well described by hyper-Laplacian [41]. Therefore, we suggest using hyper-Laplacian to model surface gradients:

\[
q(X|\theta, p) = \prod_i \frac{p}{2} \left(\frac{\theta}{\Gamma(\frac{1}{p})}\right)^\frac{1}{p} \exp\left(-\frac{\theta}{2} |(DX)|^p\right), \tag{34}
\]

where \( \Gamma \) is the Gamma function, and \( p \) and \( \theta \) are the shape parameters. \( p \in [0, 1] \) determines the peakiness and \( \theta \) determines the width of a hyper-Laplacian distribution. \((DX)\) denotes the surface gradients of the \( i \)-th edge.

One concern is whether surface gradients of real 3D models follow the hyper-Laplacian distribution. Fig. 5 shows the empirical distributions and the corresponding hyper-Laplacian fits of the surface gradients of three real models. We also plot the fitted Laplacian profiles (i.e. the \( \ell_1 \)-norm regularizer) for comparison. It can be observed that, the fitted hyper-Laplacian profiles is much closer to the corresponding empirical profiles than the fitted Laplacian profiles, especially for red circular box and hand Olivier, which validates that the empirical distribution can be well modeled by hyper-Laplacian.

It should be noted that, for different 3D models the shape parameters \( p \) and \( \theta \) will vary. To illustrate this, we compute the shape parameters on more than twenty public models, including Armadillo, Bunny in the Stanford repository [42] and models in the AIM@SHAPE repository [43]. We provide the shape parameters of ten models in Fig. 6. One can see that the \( p \) values vary from 0.2 to 0.75 and the \( \theta \) values vary from 5.506 to 483.9. Therefore, instead of fixing shape parameters, \( p \) and \( \theta \) values should be adaptively estimated for different 3D models. We propose the content-aware mesh denoising model that jointly estimates the mesh \( X \), noise level \( \sigma^2 \), and the shape parameters \( \theta \) and \( p \) from the observation \( X_0 \). By minimizing the negative logarithm of a posterior, the MAP framework can be equivalently formulated as the following minimization model:

\[
(X, \sigma^2, \theta, p) = \arg \min_{X, \sigma^2, \theta, p} \left\{ \frac{1}{2} n \log(2\pi\sigma^2) + \frac{\|X - X_0\|^2}{2\sigma^2} + \frac{\theta}{2} \|DX\|^p \right\} + 3m \left( \log(\Gamma(\frac{1}{p})) - \log(\frac{p}{2}) - \frac{p}{2} \log(\theta^2) \right). \tag{35}
\]

2) Alternating Minimization: As shown in Fig. 7, we propose an alternating minimization algorithm to minimize Eq. (35). We jointly estimate the mesh \( X \), noise standard deviation \( \sigma \), and hyper-Laplacian parameters \( \theta \) and \( p \) by iteratively solving the following two subproblems.

1) Given \( X \), the optimization problem w.r.t. \( \sigma, \theta, \) and \( p \) can be reformulated as

\[
\sigma^2 = \arg \min_\sigma \left\{ \frac{n}{2} \log(\sigma^2) + \frac{\|X - X_0\|^2}{2\sigma^2} \right\}, \tag{36}\]

\[
\theta, p = \arg \min_{\theta, p} \left\{ \frac{\theta}{2} \|DX\|^p + 3m \left( \log(\Gamma(\frac{1}{p})) - \log(\frac{p}{2}) - \frac{p}{2} \log(\theta^2) \right) \right\}. \tag{37}\]

The \( \sigma \)-subproblem has the closed-form solution \( \sigma^2 = \frac{\|X - X_0\|^2}{2m} \). The problem in Eq. (37) can be solved by: (a) finding the optimal \( \theta \) for given \( p \), which results in a closed-form solution \( \theta = \frac{p}{6m} \|DX\|^p \), and (b) using a simple 1D exhaustive searching strategy to obtain the estimation of \( p \) for given \( \theta \).

2) Given \( \sigma, \theta \) and \( p \), we define \( \lambda = \theta \sigma^2 / 2 \), and then we have

\[
\bar{X} = \arg \min_X \frac{1}{2} \|X - X_0\|^2 + \lambda \|DX\|^p, \tag{38}\]

where \( \lambda \) is the regularization parameter. Using the variable splitting approach, Eq. (38) can be reformulated as:

\[
\bar{X} = \arg \min_X \frac{1}{2} \|X - X_0\|^2 + \lambda \|\psi\|^p + \beta \|DX - \psi\|^2, \tag{39}\]

which can also be optimized with an alternating optimization.
Generally, the noise standard deviation \( \sigma \) the denoising procedure.

\[
\theta = \text{grad}_{x} f(x, \theta) \quad \text{grad}_{x} \text{has been applied to image denoising, which assumes gradually increasing it until convergence.}
\]

**Fig. 5:** Surface gradient distributions of three real 3D models: (a) the Red circular box, (b) the Hand Olivier, and (c) the Gargo. (d)-(f) show their empirical distributions of sharpness (red), the fitted hyper-Laplacian profiles (blue) and the fitted Laplacian profiles (green).

method. Fix \( \psi, X \) can be optimized by solving the following quadratic problem:

\[
\bar{X} = \arg \min_{X} \frac{1}{2} \| X_{0} - X \|^{2} + \beta DX - \psi^{2}.
\]  

(40)

Fix \( X, \psi \) can be optimized by solving the following subproblem:

\[
\tilde{\psi} = \arg \min_{\psi} \lambda \| \psi \|^{p} + \beta |D X - \psi|^{2}.
\]  

(41)

The above subproblem can be efficiently solved by using the generalized shrinkage/thresholding (GST) method [40]. Solution to each \( \psi \), can be written as:

\[
\psi_{i} = T_{p,\beta}^{GST} ((D X_{i}); \lambda),
\]  

(42)

where \( T_{p,\beta}^{GST} \) is the generalized shrinkage/thresholding operator [40]. When the penalty factor \( \beta \to \infty \), the solution to Eq. (39) converges to that of Eq. (38). In practice, we adopt the continuation technique by initializing \( \beta \) with a small value and gradually increasing it until convergence.

In [41], the MAP framework with hyper-Laplacian priors on gradients has been applied to image denoising, which assumes that \( \theta, \sigma, p \) are known and only the latent image \( X \) is unknown. Compared with [41], the novelty of our content-aware mesh denoising model lies in the joint estimation of clean mesh \( X \), noise standard deviation \( \sigma \), and the shape parameters \( p \) and \( \theta \). Generally, the noise standard deviation \( \sigma \) is unknown in MVS, and different 3D models have different shape parameters \( p \) and \( \theta \). Therefore, it is critical to adaptively estimate \( \theta, \sigma, p \) during the denoising procedure.

V. EXPERIMENTS

We implemented the proposed DCV method using C++ with OpenGL, CGAL, and TAUCS library. The predicted images are estimated using projective texture mapping, and OpenGL is adopted to generate the horizon and terminator of triangular surface. CGAL is used to manipulate the triangular mesh and TAUCS is used to manipulate the sparse matrix. We quantitatively and qualitatively evaluate the performance of DCV on multiple datasets, including the Middlebury benchmark and several public datasets with indoor and outdoor scenes. These public datasets have camera calibration parameters available. We also provide three real datasets, i.e., *Buddha*, *Totoro* and *bell*, taken from mobile phone or digital camera. For these three datasets, cameras are calibrated using the Bundler software [44]. More experimental results are provided in the supplementary file.

For the detail-preserving similarity measure, the standard deviation \( \phi \) for Gaussian convolution is set to 25, the windows size is fixed to be 7 × 7, the tradeoff parameter \( \kappa \) in Eq. (31) is set to 0.001 at the beginning of surface evolution and gradually increase with \( \kappa^{(i+1)} = \min(1.05 \times \kappa^{(i)}, 0.02) \). For the content-aware mesh denosing, we set the initial value of \( \lambda \) in Eq. (38) as 0.01\( l_{\text{ave}}^2 y_{\text{ave}} \), where \( l_{\text{ave}} \) is the average edge length and \( y_{\text{ave}} \) is the average dihedral angle of a current mesh surface. Then all the parameters \( p, \theta, \sigma \) and \( \lambda \) are adaptively estimated along with iterations.

A. Initialization and Implementation Details

In our experiments, we consider two initialization methods: visual hull and PMVS+PSR (Poisson Surface Reconstruction). The visual hull is the intersection of the visual cones associated with all image silhouettes, and can provide good initialization for scenes where the background can be easily segmented from foreground (e.g., most indoor scenes). The PMVS is an open source software designed by Furukawa and Ponce [8]. PMVS+PSR is used to initialize the scene where the background is not easy to be segmented from foreground. PMVS+PSR generally consists of two stages, where a set of dense patches are first generated from PMVS with its default parameters and then a triangular surface mesh is estimated by using PSR [45] with octree depth fixed to 8. Since our interested object is foreground, the post-processing is employed to refine the PMVS+PSR result. In particular, we removed the large triangles with a signal value threshold provided by PSR (the signal value threshold is set to 7 in this work), and removed the additional isolated clutters from the background by retaining the largest connected component of the mesh. For the temple dataset, because some small protruding structures tend to be over-smoothed when large concave region is recovered in the back of temple, we also use PMVS+PSR to generate an initial mesh. The statistics of all the datasets used in our experiments are listed in Table I, including the number of images, image resolution, initial points and running time (CPU i7, 2.4Ghz).

Two issues, non-convexity and topology adaptivity, are considered in our implementation. Since the \( \ell_{p} \)-sparsity (0 ≤ \( p \) ≤ 1) is used, the objective functional of DCV is non-convex, making the algorithm sensitive to local minimum. To alleviate this, we adopt a multi-resolution scheme. We first minimize the energy on low resolution mesh and downsample images accordingly, and then optimize it on high-resolution mesh and full-size images. The Gaussian pyramid is used to downsample the image. The Qslim algorithm [46] is used to simplify the mesh, and \( \sqrt{3} \)-subdivision scheme [47] is used to subdivide the mesh to a higher resolution. Another issue is the topology adaptivity of mesh-based methods, which needs an
initial surface of the model with an approximately consistent topology. We use two initialization methods, visual hull and PMVS+PSR, in the experiments. Other initialization methods can also be deployed, e.g., those methods based on fusion of depth maps and volumetric optimization.

**TABLE I: Datasets used in our experiments**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of images</th>
<th>Resolution</th>
<th>Initialization</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>dino sparse</td>
<td>16</td>
<td>640x480</td>
<td>visual hull</td>
<td>90</td>
</tr>
<tr>
<td>dino ring</td>
<td>48</td>
<td>640x480</td>
<td>visual hull</td>
<td>150</td>
</tr>
<tr>
<td>temple sparse</td>
<td>16</td>
<td>640x480</td>
<td>pmvs+psr</td>
<td>105</td>
</tr>
<tr>
<td>temple ring</td>
<td>47</td>
<td>640x480</td>
<td>pmvs+psr</td>
<td>170</td>
</tr>
<tr>
<td>Beethoven</td>
<td>33</td>
<td>1024x768</td>
<td>visual hull</td>
<td>180</td>
</tr>
<tr>
<td>Herzjue-P8</td>
<td>8</td>
<td>3072x2048</td>
<td>pmvs+psr</td>
<td>150</td>
</tr>
<tr>
<td>Totoro</td>
<td>8</td>
<td>1504x1004</td>
<td>pmvs+psr</td>
<td>45</td>
</tr>
<tr>
<td>Buddha</td>
<td>5</td>
<td>2400x1800</td>
<td>pmvs+psr</td>
<td>30</td>
</tr>
<tr>
<td>bell</td>
<td>3</td>
<td>1504x1004</td>
<td>pmvs+psr</td>
<td>20</td>
</tr>
<tr>
<td>statuegirl</td>
<td>50</td>
<td>2592x3888</td>
<td>pmvs+psr</td>
<td>750</td>
</tr>
</tbody>
</table>

### B. Comparison with the competing methods

We first evaluate the effectiveness of DCV on the Middlebury benchmark [4] by using two performance indicators: accuracy and completeness. The accuracy is measured by the distance $d$ such that the distance between 90% of the reconstructed surface and the ground truth surface is less than $d$. Completeness is measured by the percentage $f$ such that the distance between percentage $f$ of the ground truth surface and the reconstructed surface is less than 1.25 mm.

We test DCV on the dino ring (48-views), dino sparse (16-views), temple ring (47-views) and temple sparse (16-views), respectively. The smaller the number of views is, the more difficult and challenging the reconstruction will be. The accuracy and completeness of DCV on these four datasets are shown in Fig. 8. These results are also publicly available on the Middlebury evaluation page [48] and can be compared with state-of-the-arts. It is worthy to mention that (at the time that this paper is submitted) our DCV method achieves the best result on the dino ring and dino sparse in terms of both model completeness and accuracy. Though our results on the temple ring and temple sparse datasets are not top ranked, it can be seen that the visual quality of our results is better than most of the other top ranked methods. Fig. 9 compares the reconstruction results by DCV and several top ranked methods on the temple ring datasets.

The quantitative comparison results between DCV and several state-of-the-art methods [6–8], [11], [12], [14], [17], [31] are listed in Table II. Since some reconstruction results were not reported by authors, we labeled them as ‘-’.

A visual comparison with four representative methods [8], [12], [14], [17] on the dino sparse dataset is presented in Fig. 10. The methods proposed in [14] and [17] use the ZNCC and SSD for reprojection error minimization and use isotropic mesh smoothing. The method proposed in [12] uses an anisotropic weighted minimal surface functional. The method [8] is a combination of patch-based method and isotropic surface refinement.

We further apply DCV to two other public datasets: the Beethoven dataset [11], and the statuegirl dataset [49]. Since groundtruths of these datasets are not available, qualitative evaluation of the reconstruction results is adopted.

The Beethoven dataset contain thirty three 1024 × 768 images. It was captured by a set of synchronized cameras and presents textureless/smooth surface. The reconstruction results of Beethoven by several self-implemented state-of-the-art methods [7], [14], [17] and the proposed DCV are shown in Fig. 11. Thanks to the content-aware $\ell_p$-minimization based denoising algorithm, DCV is able to effectively suppress noise and outliers while keeping the sharp features of surface.

The statuegirl is an outdoor dataset including fifty 2592 × 3888 images. The input images, initial surface generated by PMVS+PSR, the reconstruction results by DCV and commercial 3D reconstruction software Smart3Dcapture (free edition) [49] are shown in the first row of Fig. 12, respectively. The close-ups images in different surface regions are shown in the second and the third rows. The images have been down-sampled by half before performing the reconstructions. For the Smart3Dcapture software, a complete and robust reconstruction pipeline has been integrated, including camera calibration, dense reconstruction and visualization. We have used the software’s ultra high precision option to recover more details. For our DCV, bundler is used for calibration and PMVS+PSR is used for initialization. The comparison results show that DCV can generally obtain similar results to Smart3Dcapture, and in some part (e.g., toes) it can recover more fine-scale details.

### C. Evaluation on Detail-preserving Similarity Measure

Several state-of-the-art methods, e.g., Kostrikov et al. [6], Vu et al. [7], Kolev et al. [11], [12], Zhaareescu et al. [14] are based on ZNCC, and their results have been submitted to the Middlebury benchmark. In Table II, we compared DCV with these ZNCC-based methods [6], [7], [11], [12], [14]. One can see that, for dino sparse, dino ring, and temple sparse, our DCV performs better than the ZNCC-based methods [6], [7], [11], [12], [14] in terms of both accuracy and completeness. Moreover, Fig. 10 shows the results by DCV and the methods in [12], [14] on dino sparse. Fig. 11 show the results by DCV and the methods [7], [14] on Beethoven. One can see that DCV can preserve more fine-scale details than the ZNCC-based methods [7], [12], [14].

In Fig. 3 and Fig. 4, using the datasets Buddha (five 2400 × 1800 images) and temple sparse, we have compared...
TABLE II: Quantitative comparison between DCV and several state-of-the-art methods on the Middlebury data sets in terms of accuracy/completeness

<table>
<thead>
<tr>
<th>Method</th>
<th>dino sparse</th>
<th>dino ring</th>
<th>temple sparse</th>
<th>temple ring</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCV</td>
<td>0.3mm/100%</td>
<td>0.28mm/100%</td>
<td>0.66mm/97.3%</td>
<td>0.73mm/98.2%</td>
</tr>
<tr>
<td>Vu [7]</td>
<td>–</td>
<td>0.53mm/99.7%</td>
<td>–</td>
<td>0.45mm/99.5%</td>
</tr>
<tr>
<td>Kostrikov [6]</td>
<td>0.37mm/99.3%</td>
<td>0.35mm/99.6%</td>
<td>0.79mm/95.8%</td>
<td>0.57mm/99.1%</td>
</tr>
<tr>
<td>Zaharescu [14]</td>
<td>0.45mm/99.2%</td>
<td>0.42mm/98.6%</td>
<td>0.78mm/95.8%</td>
<td>0.55mm/99.2%</td>
</tr>
<tr>
<td>Kolev2 [11]</td>
<td>0.53mm/98.3%</td>
<td>0.43mm/99.4%</td>
<td>1.04mm/91.8%</td>
<td>0.72mm/97.8%</td>
</tr>
<tr>
<td>Kolev3 [12]</td>
<td>0.48mm/98.6%</td>
<td>0.42mm/99.5%</td>
<td>0.97mm/92.7%</td>
<td>0.7mm/98.3%</td>
</tr>
<tr>
<td>Gargallo [31]</td>
<td>0.76mm/90.7%</td>
<td>0.6mm/92.9%</td>
<td>1.05mm/81.9%</td>
<td>0.88mm/84.3%</td>
</tr>
<tr>
<td>Delaunoy [17]</td>
<td>0.89mm/93.9%</td>
<td>–</td>
<td>0.73mm/95.9%</td>
<td>–</td>
</tr>
<tr>
<td>Furukawa3 [8]</td>
<td>0.37mm/99.2%</td>
<td>0.28mm/99.8%</td>
<td>0.63mm/99.3%</td>
<td>0.47mm/99.6%</td>
</tr>
</tbody>
</table>

Fig. 9: Comparison of reconstruction results on the Middlebury temple ring dataset. The names of the comparison methods follow the entries in the Middlebury evaluation website. From left to right: Vu [7], Acc. 0.45, Comp. 99.8%; Campbell [9], Acc. 0.48, Comp. 99.4%; Furukawa3 [8], Acc. 0.47, Comp. 99.6%; Hernandez [13], Acc. 0.52, Comp. 99.5%; the proposed DCV method, Acc. 0.73, Comp. 98.2%; and groundtruth.

Fig. 10: Comparison of reconstruction results on the dino sparse dataset. From left to right: results by [17], [12], [14], [8], the proposed DCV, and groundtruths, respectively. It is obvious that DCV has better performance in preserving the details and sharp features while filtering the noises.

Fig. 11: Reconstruction results by several state-of-the-art methods and the proposed DCV on the Beethoven datasets. From left column to right column: input images, results by [14], [17], [7], and DCV in two views, respectively.
Fig. 12: Results on the *statuegirl* dataset. First row, from left to right: one of input images, initial surface by PMVS+PSR, results by DCV and results by Smart3dCapture Free Edition with ultra high precision setting. Second and third rows, from left to right: the close-ups of reconstruction results in the first row.

Fig. 14: Reconstruction results on the *bell* dataset. First row, from left to right: one of the input images, the initial reconstruction using PMVS+PSR, result by ZNCC + isotropic smoothing, result by detail-preserving similarity measure + isotropic smoothing, and result by DCV. Second row, from left to right: the point clouds generated by PMVS, and close-up images corresponding to the red rectangle regions in the first row, respectively. One can see that DCV preserves well the fine details and smooth surface.

ZNCC and the proposed similarity measures based on our DCV framework, i.e. we implement two variants of our methods: (i) ZNCC + isotropic regularizer [8], and (ii) Detail-preserving similarity + isotropic regularizer [8]. As shown in Fig. 3 and Fig. 4, with the same regularization method, detail-preserving similarity is more effective in reconstructing fine-scale details and sharp features. The *barrel* datasets consists of fifteen 864×1536 images taken by a cellphone. We fixed surface regularization to the isotropic regularizer adopted in [8] and changed the methods of similarity measure, i.e. ZNCC and the proposed measure. The results are shown in Fig. 13. It can be seen that the proposed similarity measure recovers more fine-scale details.

We further compare three variants: (i) isotropic ZNCC similarity + isotropic regularization, (ii) detail-preserving similarity + isotropic regularization [8], and (iii) DCV. The chosen isotropic regularization combines the first order and second order Laplacian [8]. The results on the *bell* dataset are shown in Fig. 14. The *bell* dataset consists of only three 1504×1004 images, and thus only a partial surface of bell is reconstructed. The less number of observed images makes the regularization scheme more important. It is clear that DCV presents the best results for both fine details and surface smoothness among all competing methods.

D. Evaluation on Content-aware Mesh Denoising

Our DCV method consists of two components, i.e., detail-preserving similarity measure and content-aware $\ell_p$ mesh denoising. The effectiveness of the former component has been validated in Section V-C. To evaluate the effectiveness of the latter component, we implement four variants of DCV by substituting the content-aware $\ell_p$ mesh denoising with four
Fig. 15: Reconstruction results by different mesh denoising methods on the *Herzjesu-P8* dataset. First row, from left to right: two of input images, results by combining the first order and second order Laplacian [8] and results by Sun et al.’s method [25]. Second row, from left to right: results by Zhang et al.’s bilateral normal filtering [26], results by He et al.’s $\ell_0$ denoising [27] and results by our content-aware $\ell_p$ denoising method. All the models are flat-shaded to show faceting effect.

Fig. 16: Reconstruction results by using different mesh denoising methods on the *Totoro* dataset. First row, from left to right: two of input images, results by combining the first order and second order Laplacian [8], results by Sun et al.’s method [25]. Second row, from left to right: results by Zhang et al.’s bilateral normal filtering [26], results by He et al.’s $\ell_0$ denoising [27], results by our content-aware $\ell_p$ denoising method. All the models are flat-shaded to show faceting effect.

Fig. 13: Reconstruction results on the *barrel* dataset. First row, from left to right: two of input images, initial surface by PMVS+PSR, the results by ZNCC-based similarity measure and by our proposed detail-preserving similarity measure. Second row shows close-ups of the results in the first row. It can be clearly seen that the proposed similarity measure recovers more fine-scale details.

VI. Conclusion

In this paper, we proposed a detail-preserving and content-aware variational (DCV) method for multi-view stereo (MVS) reconstruction. First, by connecting guided image filtering with image registration, a novel similarity measure was proposed.
to preserve the fine-scale details in reconstruction. Second, by the hyper-Laplacian modelling of surface gradients, a content-aware mesh denoising method based on $\ell_p$ minimization was presented to suppress the noises and outliers while preserving sharp features. Compared with state-of-the-art MVS methods, the proposed DCV method is capable of reconstructing a smooth and clean surface with finely preserved details and sharp features.

This work provides a new insight for MVS by employing detail-persevering similarity measure and content-aware mesh denoising, which might shed light on developing new MVS or other 3D reconstruction methods. However, our DCV may not perform well for the scenario with complex reconstruction noise, and the computational cost is relatively high. In our future work, to deal with complex noise, we will consider using more flexible form of noise (e.g., mixture of Gaussian [50]) or extend our model to the blind denoising framework [51] for joint noise estimation and denoising. To improve efficiency of our DCV, GPU-based parallel implementation on the main parts of gradient computation will be adopted by using the Nvidia CUDA framework.

Acknowledgement

The authors would like to thank the associate editor and the anonymous reviewers for their constructive suggestions, thank Dr. Daniel Scharstein for evaluating our results on the Middlebury datasets, and thank Dr. Hoang Hiep Vu for providing the statuegirl dataset. This research is partly supported by Hong Kong RGC GRF grant (PolyU 5313/13E), the NSFC grants (61271093, 61173086, 61571165 and 61373114), and the program of ministry of education for new century excellent talents (NCET-12-0150).

References


Zhaoxin Li received the B.Sc. degree and M.Sc. degree in Computer Science from Tianjin Polytechnic University, Tianjin, China, in 2007 and 2010, respectively. He is currently pursuing the Ph.D. degree from the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. His research interests include 3D computer vision and 3D data processing.

Wangmeng Zuo (M’09, SM’15) received the Ph.D. degree in computer application technology from the Harbin Institute of Technology, Harbin, China, in 2007. From July 2004 to December 2004, from November 2005 to August 2006, and from July 2007 to February 2008, he was a Research Assistant at the Department of Computing, Hong Kong Polytechnic University, Hong Kong. From August 2009 to February 2010, he was a Visiting Professor in Microsoft Research Asia. He is currently an Associate Professor in the School of Computer Science and Technology, Harbin Institute of Technology. Dr. Zuo has published more than 50 papers in top tier academic journals and conferences. His current research interests include image modeling and blind restoration, discriminative learning, biometrics, and 3D vision. Dr. Zuo is an Associate Editor of the IET Biometrics. He is a senior member of the IEEE.

Deyu Meng received B.Sc., M.Sc., and Ph.D. degrees in 2001, 2004, and 2008, respectively, from Xi’an Jiaotong University, Xi’an, China. He is currently an associate professor with the Institute for Information and System Sciences, School of Mathematics and Statistics, Xi’an Jiaotong University, Xi’an. His current research interests include machine learning, computer vision, data mining and multimedia analysis.

Kuanquan Wang (M’01, SM’07) is a full professor and PhD supervisor with School of Computer Science and Technology at Harbin Institute of Technology. He is a senior member of IEEE, a senior member of China Computer Federation and a senior member of Chinese Society of Biomedical Engineering. His main research areas include image processing and pattern recognition, biometrics, biocomputing, virtual reality and visualization. So far, he has published over 200 papers and 6 books, got 10 patents, and won 1 second prize of National Teaching Achievement.

Lei Zhang (M’04, SM’14) received the B.Sc. degree in 1995 from Shenyang Institute of Aeronautical Engineering, Shenyang, P.R. China, the M.Sc. and Ph.D degrees in Control Theory and Engineering from Northwestern Polytechnical University, Xi’an, P.R. China, respectively in 1998 and 2001. From 2001 to 2002, he was a research associate in the Dept. of Computing, The Hong Kong Polytechnic University. From Jan. 2003 to Jan. 2006 he worked as a Postdoctoral Fellow in the Dept. of Electrical and Computer Engineering, McMaster University, Canada. In 2006, he joined the Dept. of Computing, The Hong Kong Polytechnical University, as an Assistant Professor. Since July 2015, he has been a Full Professor in the same department. His research interests include Computer Vision, Pattern Recognition, Image and Video Processing, and Biometrics, etc. Dr. Zhang has published more than 200 papers in those areas. By 2015, his publications have been cited more than 15,000 times in literature. Dr. Zhang is currently an Associate Editor of IEEE Trans. on Image Processing, IEEE Trans. on CSVT and Image and Vision Computing. He was selected as the Highly Cited Researcher by Thomson Reuters, 2015. More information can be found in his homepage http://www4.comp.polyu.edu.hk/~cszlzhang/.