Endoscopography: Deriving a 3D Textured Surface from Endoscopic Video, Then Registration with CT

Rui Wang*, Qingyu Zhao, True Price, Ruibin Ma, Adam Aji, Jan-Michael Frahm, Julian Rosenman, and Stephen Pizer

***Abstract*—Endoscopy [1] enables high-resolution visualization of tissue texture and is a critical step in many clinical workflows, including diagnosis and treatment planning for cancers in the nasopharynx. However, an endoscopic video does not provide its information in 3D space, making it difficult to use in tumor localization, and it is inefficient to review. We present a new imaging approach that we call *endoscopography*. Endoscopography reconstructs a full 3D textured surface, which we call an *endoscopogram*, from an endoscopic video. This endoscopogram opens the door for novel 3D visualizations of patient anatomy derived solely from endoscopic data. It also allows information contained in the tissue texture in the endoscopogram to be transferred to 3D image such as CT through a surface-to-surface registration. In particular, through an interactive tool, the physician can draw directly on the endoscopogram surface to specify a tumor, which then can be automatically transferred to CT slices to aid tumor localization. We describe the sequence of methods we have designed to achieve this goal. Through evaluations on synthethic, phantom, and patient data, we demonstrate that our surface reconstruction method can obtain accurate results within millimeters of ground-truth, and our registration approach can handle tissue deformations and reconstruction inconsistencies across endoscopic video frames.**

***Index Terms*—Endoscopography, Endoscopogram, 3D Reconstruction, Endoscopy, Registration, nasopharynx.**

I. Introduction

There exists a variety of endoscopic examinations, and for each of these a reconstruction from the video into a 3D textured surface can be useful. The particular application that we are working on are nasopharyngoscopy and colonoscopy. In this paper we explain our methods with respects to nasopharyngoscopy applications.

Nasopharyngoscopy is a commonly used technique for nasopharyngeal cancer diagnosis and treatment planning. For radiotherapy, the planning requires tumor localization. Although nasopharyngoscopy can provide a direct, high-contrast, high-resolution visualization of a patient's interior tissue surface, it has a weakness for tumor localization in that it does not provide precise 3D spatial information. On the other hand, CT provides many critical sources of information needed in planning radiotherapy, with low distortion. However, it provides relatively low contrast and low resolution images for localization of the tumor, especially for tumors predominantly on the tissue surface, as is common in throat cancer.

Therefore, if we can leverage the advantage of tissue information in nasopharyngoscopy together with the 3D geometry information in CT scan, the accuracy of tumor localization will be increased. Our challenge is to develop technologies enabling physicians to efficiently review nasopharyngoscopies and to visualize endoscopic data directly in the CT space.

To achieve these objectives, the 2D endoscopic video must be converted into a format that 1) summarizes the overall anatomy as a single object that is easy to manipulate and 2) contains the shape information necessary for registration to the 3D patient space.

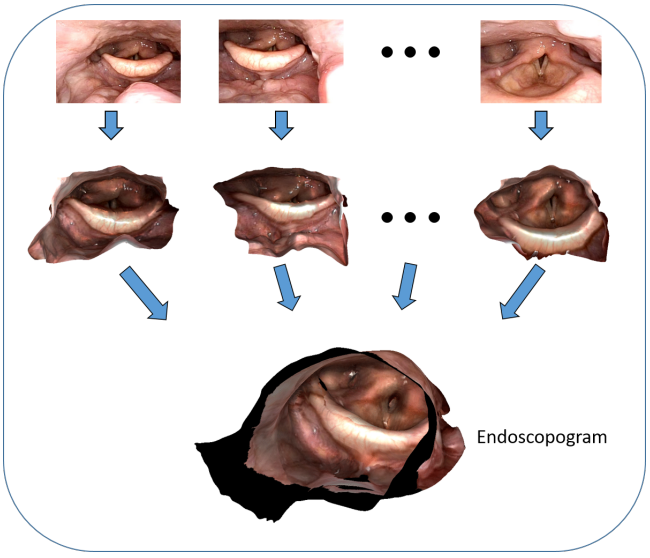

Fig. 1. Deriving endoscopogram through frame-by-frame 3D reconstruction and group-wise deformable registration.

Given an input endoscopic video sequence, our method reconstructs the throat surface as a textured 3D mesh (see Fig. 1). We term this new image type an *endoscopogram*.

The endoscopogram is generated by first reconstructing a textured 3D partial surface for each frame. Then these multiple partial surfaces are fused into an endoscopogram using a groupwise surface registration algorithm and a seamless texture fusion from the partial surfaces. Finally, the endoscopogram geometry is registered with the surface extracted from CT which enables the desired tumor transfer process.

This paper focuses on the technical details of the frame-by-frame depth reconstruction of individual endoscopic video frames, the groupwise geometry fusion of multiple partial reconstructions, their seamless texture fusion, and the registration between the endoscopogram and CT. The paper is organized as follows: Section II gives an overview of the problem and challenges for the proposed methodology. Section III presents related work and provides background on the tools we use in our methods. Section IV describes the details of our approach. Section V provides experimental validation and error analysis, and it presents qualitative analysis of

application to patients. Section VI discusses the method and addresses avenues of future work.

II. Overview

A. Challenges

Reconstructing 3D surface geometry from 2D endoscopic video and registration with the surface extracted from CT is very challenging. For example in the nasopharyngoscopy, the environment for 3D reconstruction is unknown because throat texture and shape can vary greatly from patient to patient, especially when tumors are present. Besides, due to the presence of the endoscope the throat constantly has sudden large deformations caused by the gag reflex and swallowing [2], [3]. Moreover, the specularities of the saliva-coated throat tissue and the self-occlusions of different inner structures make the reconstruction even harder.

Registration of an endoscopogram with a surface extracted from CT must deal with 1) the partial volume effect of CT, which leads to topology differences between CT and endoscopogram; 2) some anatomy not appearing in the camera view, which leads to missing data in the endoscopogram; 3) the different patient postures during CT and endoscopy procedure, which causes large deformation between the CT and endoscopogram.

Considering all these factors, a successful technique for endoscopogram construction must operate over short subsequences of the input endoscopic video without any a priori assumptions of the underlying shape, and the registration between the CT and endoscopogram must handle large deformation, missing patches and topology differences.

B. System overview

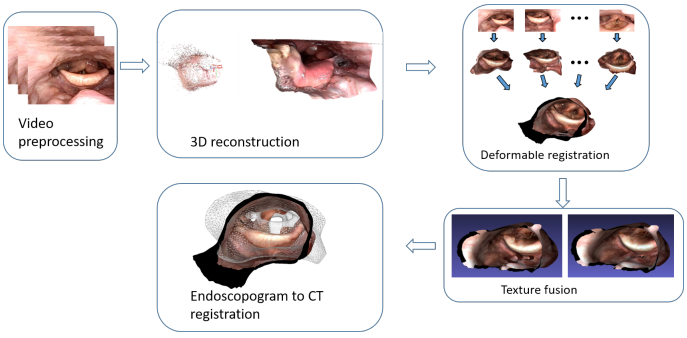

Fig. 2. System overview.

Figure 2 shows an overview of our overall system. We firstly perform a series of automatic video frames preprocessing, which includes deep learning based automatic informative frame selection and specularity removal, and a key-frame selection.

Then we perform a 3D reconstruction using the preprocessed images. Our 3D reconstruction utilizes sparse, multi-view data obtained via Structure-from-Motion (SfM) to guide Shape-from-Shading (SfS) reconstruction of the throat surface in individual frames. Novel improvements to the feature extraction and correspondence detection in SfM, and the formulation of SFS together with a new reflectance model are also introduced.

Then we apply a novel groupwise surface registration algorithm based on N-body interaction and physics-motivated modeling. This algorithm, based on a surface-to-surface registration method called Thin Shell Demons (TSD), has been described in detail in [4]. This method is improved by the estimation of nonuniform and anisotropic elasticity parameters using orthotropic physical modeling [5].

The registration between the endoscopogram and CT must handle a large deformation and the fact that both the endoscopogram and the surface extracted from CT have missing patches and topology differences. For this purpose, we use a novel method that combines TSD and estimation of the missing patches.

III. Related Work

Our work is mostly related to the problems of 3D reconstruction and non-rigid registration in the literature.

A. 3D Reconstruction

To date, most work on combining motion-based reconstruction with shading information has utilized shading to augment an existing shape template or model priors [6]. Wu *et al.* [7] proposed to first build coarse-scale dynamic models from multi-view video and then leverage shading appearance to estimate fine-scale, temporally varying geometry. Fine-scale shading correction has also been used to refine dense surfaces obtained via depth sensor [8], [9]. In endoscopic applications, a related method by Tokgozoglu *et al.* [10] used multi-view stereo to derive a low-frequency model of the upper airway, then applied Lambertian SFS on albedo-normalized images to endow the existing surface with higher-resolution shape. For monocular reconstruction of deforming environments, several efforts have been made to extend the Shape-from-Template problem [11] to utilize shading information. In [12], [13], [14], Malti, Bartoli, and Collins proposed a two-stage approach for surgery of the uterus: Pre-surgery, an initial 3D template is recovered under rigid scene assumptions, and reflectance parameters are estimated for the surface. In surgery, the deforming surface is recovered via conformal deformations of the template surface, and subsequent shading refinement is performed using the estimated reflectance model. We address the problem of dense reconstruction in conditions where dense shape templates are unavailable or difficult to derive. Laryngoscopy is a good example of this (Figure 7) because the anatomic shapes in this region are highly patient-specific, and surfaces extracted from, for example, CT scans are typically low-resolution and have a notably different shape compared to endoscopy. Multi-view stereo also tends to fail in this scenario, as the combination of strong illumination changes and limited non-deforming image sequences is prohibitive. Motivated by our observation that SfM works over short temporal sequences for these data, we develop a method for dense single-view surface estimation that leverages sparse 3D geometry obtained from SfM.

B. Non-rigid surface Registration

Non-rigid 3D registration has been a common topic in medical image analysis. In particular, we are mostly interested in non-rigid surface registration methods.

Surface embedding is one class of surface registration methods. [15], [16] proposed a multidimensional scaling embedding method that can place the two surface vertices in a low-dimensional Euclidean space, where the nearest-neighbor matching method can be performed to yield the correspondences. Gu *et al.* [17] proposed to use conformal mapping with angle-preserving constraint to embed the surfaces into a common disc or sphere domain. However, such methods requires the surfaces to have the same intrinsic geometry that it cannot handle surface topology change or missing patches.

Matching-based methods [18], [19], [20] use hand-crafted feature descriptors to perform feature matching, which produce a set of corresponding points. However, without any regularization the outliers produced in the feature matching will lead to non-smooth or even incorrect deformations. Zeng *et al.* [21] proposed to use MRF to regularize the deformation field. LDDMM [22] have provided an elegant mathematical framework that produces diffeomorphic deformations between surfaces by comparing their normal fields.

Thirion *et al.* [23] proposed a Demons algorithm which optimize a per pixel wise displacement field. The forces that apply on each pixel were inspired from the optical flow equations. The idea of the Demons algorithm is appealing because it has no assumptions about the surface properties.

IV. Methods

A. Video frame preprocessing

The 3D reconstruction algorithm requires consecutive and clear views of the target surface. We propose an automatic video frame preprocessing pipeline that contains three steps 1) informative frame selection; 2) specularity removal; 3) key-frame selection. These are crucial steps to make our overall pipeline fully automatic and efficient.

1) Informative frame selection: Endoscopic video contains a large fraction of non-informative frames. Non-informative frames includes tissue surface being obscured by fecal matter, motion blur, the camera being too close to the tissue surface, water flushing (in colonoscopic video), etc. Explicitly extracting features and training classifiers to identify these various kinds of non-informative frames is very difficult. A deep learning method, on the other hand, can directly learn from raw images to distinguish informative frames from non-informative frames without the need of manually crafted features; thus it is very suitable for this task.

Distinguishing informative frames from non-informative frames is a binary classification problem. We have adopted the VGG16 [24] network architecture; other network architectures such as googlenet or resnet can certainly be used as well. The input to the network is a single RGB frame and the output is its probability of being an informative-frame.

Figure 3 shows an example of informative and non-informative frames in a colonoscopic video. We manually divided 50000 images from five patients into two classes as our training data. We tested the performance of the trained model on two other patients and it achieves 98.6 % accuracy.

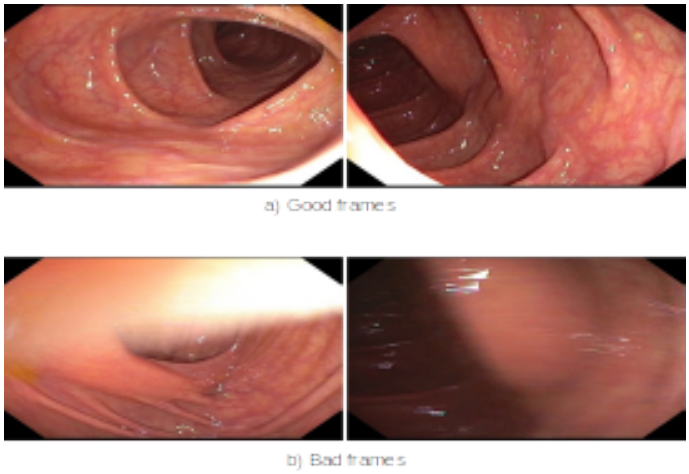

Fig. 3. Example of informative and non-informative frames in colonoscopic video

2) Specularity removal: Specular points are very common in endoscopic videos because the light source is very close to tissue surface. Moreover, because the surface is moist, the specularities are quite intense. Specularity causes a lot of problems in 3D reconstruction including wrong feature detection and matching, and saturated shading information. We proposed a deep learning-based specularity removal method that can remove specular points in real time.

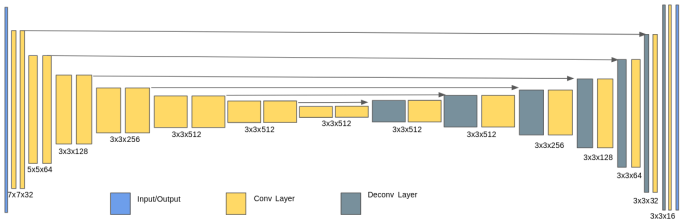

Fig. 4. DispNet architecture for our specularity removal task.

We use the DispNet [25] architecture for our specularity removal network. The DispNet has an encoder-decoder architecture as shown in Figure 4. The input and output of our network has the same size and number of channels.

The training data is generated using a software named Meitu XiuXiu such that by using some of its functions the specular points in endoscopic images can be manually removed. We manually generated 256 frames as training data. Figure 5 shows an example of our specularity removal results.

3) Key-frame selection: A critical step in our 3D reconstruction algorithm is SfM, which provides an estimation of the 3D point positions and the camera locations. However, performing SfM on the whole video is very time-consuming because of the large number of frames. Chronologically close frames contain almost the same contents, which would result in ambiguity in the step of triangulation in SfM. Moreover in some stable time domains, having many redundant frames can hide the most informative moving scenes from being reconstructed. Therefore, a key-frame selection technique is needed

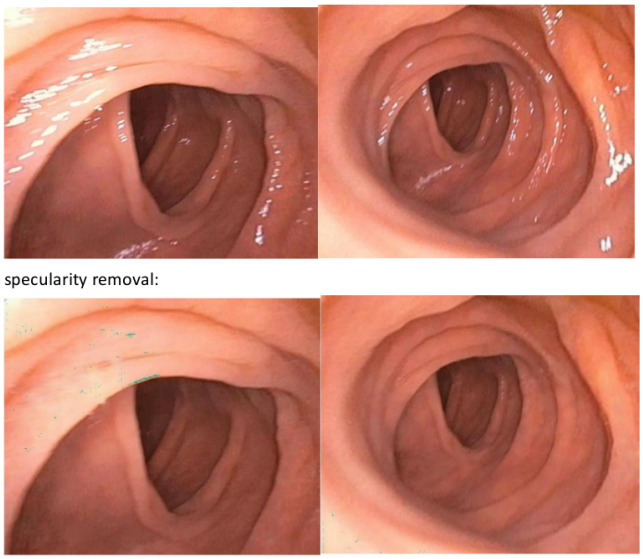

Fig. 5. Example of specularity removal results.

to exclude the redundant frames and keep the informative moving frames.

Our key-frame selection method consists of three major components:

1) Sort the frames according to their sharpness: the integral of the square of the magnitude of the color gradients.
2) Define a motion score between two images using optical flow and the normalized correlation coefficient (NCC). In detail, an optical flow vector field is calculated by Flownet2.0 [26], and one of the two images is warped to the other one. Then calculate the NCC between the warped image and the target image (only taking into account the pixels that have correspondence).
3) Inspect each frame from low sharpness to high sharpness. If the motion score between its remaining chronological neighbors is less than a threshold, the frame is considered as unnecessary to build up connection and will be deleted from the time sequence. Otherwise it will be taken as a keyframe.

	Total frames	Classified as good	Keyframes	Percentage
Case 1	7248	1429	408	5.6%
Case 2	4974	2963	420	8.4%
Case 3	5122	2458	374	7.3%
Case 4	7522	3106	882	11.7%
Case 5	3682	2007	336	9.1%
Average	5709.6	2392.6	484	8.5%

Fig. 6. Key-frames in 5 cases.

Figure 6 shows the number and percentage of key-frames selected in five cases.

B. Temporally local 3D reconstruction

In this stage, our 3D reconstruction method recovers depth information separately for the selected images in the input video sequence. Our approach is a combination of SfM and SfS, therefore we name it Structure-from-Motion-and-Shading (SfMS). Figure 7 shows example results of our SfMS method.

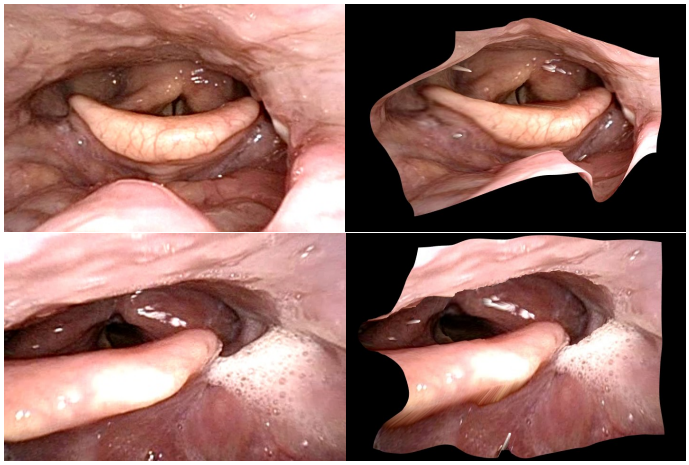

Fig. 7. Example results of SfMS on live endoscopy from two different patients. Left: Original image. Right: Surface estimated from the image using our algorithm.

Our approach achieves this using a new Shape-from-Shading formulation that utilizes the sparse, but accurate, 3D point data obtained via Structure-from-Motion. In this section, we detail the main contributions of the current work that enable this enhanced depth estimation: First, we introduce a regularized formulation of SFS that allows for a trade-off between predicted image intensity and similarity to an existing estimated surface. We also suggest a way to account for errors along occlusion boundaries in the image using intensity-weighted finite differences. Second, we propose a general reflectance model for use in our SFS framework that can more accurately capture real-world illumination conditions. Finally, we develop an iterative update scheme that (1) warps an estimated surface to the SfM point cloud, (2) estimates a reflectance model using this warped surface and the given image, and (3) produces a new estimated surface using the regularized SFS method.

SfM and SFS. Our novel SfMS method is based on two classical methods: Structure-from-Motion (SfM) and Shape-from-Shading (SfS). We first sketch these two methods, as well as the reflectance model used in SfMS.

SfM [27], [14], [13] is the simultaneous estimation of camera motion and 3D scene structure from multiple images taken at different viewpoints. Typical SfM methods produce a sparse scene representation by first detecting and matching local features in a series of input images, which are the individual frames of the endoscope video in our application. Then, starting from an initial two view reconstruction, these methods incrementally estimate both camera poses (rotation and position for each image) and scene structure. The scene structure is parameterized by a set of 3D points projecting to corresponding 2D image features.

Our motivation for using SfM is that it provides a prior on depth, albeit at sparse locations, that provides constraints for surface geometry and reflectance model estimation. Figure

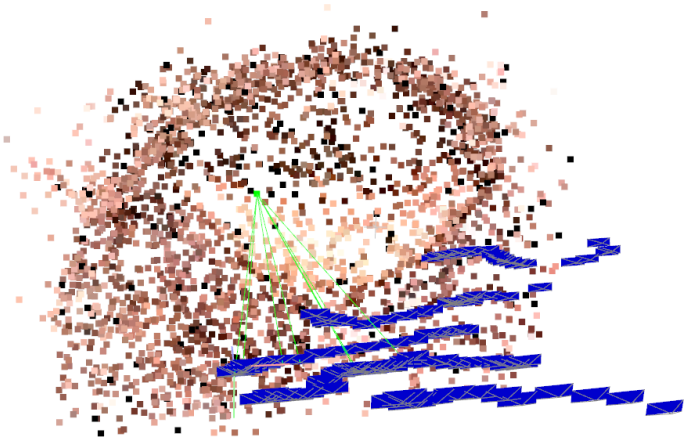

Fig. 8. Structure-from-Motion results for endoscopic video. Individual 3D surface points (colored dots) and camera poses (blue) are jointly recovered.

8 shows an example SfM reconstruction of endoscopic data using several segments from the overall video. One limitation to the generality of our method is that sparse non-rigid reconstruction in medical settings is an unsolved problem [3]. However, the approach we propose can handle any sparse data as input, so the method could easily be integrated with non-rigid SfM in future work.

SFS, first introduced in the 1970 thesis of Horn [28], is a monocular method of depth estimation that, given a single image viewing a scene, recreates the three-dimensional shape of the scene under given assumptions about the lighting conditions and surface reflectance properties [29], [30], [31]. A number of different formulations have been proposed to solve the SfS problem, including energy minimization, recovery of depth from estimated gradient, local shape estimation, and modeling as a partial differential equation (PDE) [29], [31]. Over the last decade, the PDE formulation of SFS has received the most attention, starting with Prados and Faugeras [32], who introduced a novel, provably convergent approach for solving the problem as a PDE.

Our primary motivation for using SFS is that many of its simplifying assumptions are well adjusted to general endoscopic devices. In particular, use of an endoscope allows us to assume a co-located camera and light source, which greatly simplifies the modeling of surface reflectance in the scene. We next describe what this simplification entails, which sets the stage for introducing our proposed reflectance model.

Reflectance Models. The amount of light reflecting from a surface can be modeled by a wavelength-dependent Bidirectional Reflectance Distribution Function (BRDF) that describes the ratio of the radiance of light reaching the observer $I_{\lambda r}$ to the irradiance of the light hitting the surface $E_{\lambda r}$ [33]. Generally, a BRDF is given as a function of four variables: the angles (θ_i, ϕ_i) between the incident light beam and the normal, and the reflected light angles (θ_r, ϕ_r) with the normal; that is,

$$\text{BRDF}_\lambda(\theta_i, \phi_i, \theta_r, \phi_r) = \frac{I_{\lambda r}}{E_{\lambda i}}, \tag{1}$$

where λ represents light wavelength. In the following, we implicitly assume the wavelength dependence of the BRDF.

The irradiance for an incoming beam of light is itself a function of θ_i and the distance r to the light source:

$$E_i = I_i \frac{A}{r^2} \cos\theta_i, \tag{2}$$

where I_i is the light source intensity and A relates to the projected area of the light source.

We make two simplifying assumptions about the BRDF that help the overall modeling of the problem. First, we assume surface isotropy of the BRDF, which constrains it to only depend on the relative azimuth, $\Delta\phi = |\phi_i - \phi_r|$, rather than the angles themselves [34]. Second, we assume that the light source is approximately located at the camera center relative to the scene, which is a reasonable model for many endoscopic devices. In this case, the incident and reflected light angles are the same, *i.e.*, $(\theta_i, \phi_i) = (\theta_r, \phi_r)$. Under these assumptions, the observed radiance simplifies to

$$I_r(r, \theta_i) = I_i \frac{A}{r^2} \cos(\theta_i) \text{BRDF}(\theta_i). \tag{3}$$

The reflectance model we propose is based on the set of BRDF basis functions introduced by Koenderink *et al.* [34]. These functions form a complete, orthonormal basis on the half-sphere derived via a mapping from the Zernike polynomials, which are defined on the unit disk.

We adapt the BRDF basis of Koenderink *et al.* to produce a multi-lobe reflectance model for camera-centric SFS. First, taking the light source to be at the camera center, we have $\theta_i = \theta_r$ and $\Delta\phi_{ir} = 0$; this gives

$$\text{BRDF}(\theta_i) = \sum_{k=0}^{K-1} \left(\alpha_k + \beta_k \sin\left(\frac{\theta_i}{2}\right) \right) \cos^k \theta_i, \tag{4}$$

where α_k and β_k are coefficients that specify the BRDF.

Surface Model. Let $(x, y) \in \Omega$ represent image coordinates after normalization by the intrinsic camera parameters (centering around the principal point and dividing by the focal length). For a given camera pose, the surface function $f : \Omega \to \mathbb{R}^3$ maps points in the image plane to 3D locations on a surface viewed by the camera. Under perspective projection,

$$f(x, y) = z(x, y) \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}, \tag{5}$$

where $z(x, y) > 0$ is a mapping from the image plane to depth along the camera's viewing axis. The distance r from the surface to the camera center is

$$r(x, y) = \|f(x, y)\| = z(x, y)\sqrt{x^2 + y^2 + 1}, \tag{6}$$

and the normal to the surface is defined by the cross product between the x and y derivatives of f:

$$\mathbf{n}(x, y) = f_x \times f_y = z \begin{pmatrix} -z_x \\ -z_y \\ xz_x + yz_y + z \end{pmatrix}. \tag{7}$$

Given a co-located light source, the light direction vector for a point in the image is the unit vector $\hat{\mathbf{l}}(x,y) = \frac{1}{\sqrt{x^2+y^2+1}}(x,y,1)$. The cosine of the angle between the normal and light direction vectors is then equal to their dot product:

$$\cos\theta_i = \hat{\mathbf{n}} \cdot \hat{\mathbf{l}} = \frac{z}{\sqrt{(x^2+y^2+1)\left(z_x^2+z_y^2+(xz_x+yz_y+z)^2\right)}}, \quad (8)$$

where ("carat") represents normalization to unit length.

Prados and Faugeras [32] note that Eq. (8) can be simplified using the change of variables $v(x,y) = \ln z(x,y)$:

$$\hat{\mathbf{n}} \cdot \hat{\mathbf{l}} = \frac{1}{\sqrt{(x^2+y^2+1)\left(v_x^2+v_y^2+(xv_x+yv_y+1)^2\right)}}. \quad (9)$$

This transformation allows us to separate terms involving v from those involving its derivatives in our shading model, which is important for our subsequent formulation.

1) Adapted PDE framework: In the following, we modify the traditional SFS PDE to include regularization against a pre-existing estimated surface. Then, we address an implementation for solving this regularized SFS equation. Finally, we propose the use of weighted finite differences to mitigate the effect of smoothness assumptions in the implementation that cause inaccurate depth measurements along surface occlusion boundaries.

Original PDE. Eq. (3) models observed intensity for a generic, isotropic BRDF with the assumption that the light source is co-located with the camera. Joining this with Eqs. (6) and (9) and multiplying by r^2, we have

$$(x^2+y^2+1)I_r e^{2v} - I_i A\cos(\theta_i)\text{BRDF}(\theta_i) = 0 \quad (10)$$

(note $e^{2v} = z^2$). This is a static Hamilton-Jacobi equation of the form

$$\begin{cases} Le^{2v} - H(v_x, v_y) = 0, & (x,y) \in \Omega \\ v(x,y) = \psi(x,y), & (x,y) \in \partial\Omega, \end{cases} \quad (11)$$

where the dependence of H and L on x and y is implied. $\psi(x,y)$ defines boundary conditions for the PDE.

Regularized Equation. The PDE introduced above is dependent on the accuracy of the BRDF modeling the scene. To prevent surface mis-estimations arising from an inaccurate BRDF, we use the 3D points obtained from SfM as an additional set of constraints for our estimated log-depths, v.

We add a simple regularization to the SFS PDE (Eq. (11)) that constrains the solution to be similar to a warped surface generated from the 3D SfM points. Instead of a proper PDE, we consider the following energy function:

$$E(v) = \frac{1}{2}\left(e^{2v} - \frac{1}{L}H(v_x, v_y)\right)^2 + \frac{\lambda}{2}\left(e^{2v} - z_{\text{warp}}^2\right)^2, \quad (12)$$

where $z_{\text{warp}}(x,y)$ is the depth of the warped surface at a given image coordinate, and the parameter $\lambda(x,y) \geq 0$ controls the influence of the right term, which regularizes on depths. We define λ when we introduce our iterative algorithm, below. Minimizing $E(v)$ w.r.t v, we obtain

$$\frac{\partial E}{\partial v} = \left[e^{2v} - \frac{1}{1+\lambda}\left(\frac{1}{L}H(v_x,v_y) + \lambda z_{\text{est}}^2\right)\right] 2e^{2v} = 0. \quad (13)$$

Incorporating boundary conditions, we have the following optimization problem:

$$\begin{cases} e^{2v} - \frac{1}{1+\lambda}\left(\frac{1}{L}H(v_x,v_y) + \lambda z_{\text{est}}^2\right) = 0 & (x,y) \in \Omega \\ v(x,y) = \psi(x,y). & (x,y) \in \partial\Omega. \end{cases} \quad (14)$$

Solving the Regularized SFS Equation. We employ the fast-sweeping method proposed for SFS by Ahmed and Farag [35], itself based on a method by Kao *et al.* [36], to solve our regularized SFS equation. This approach uses the Lax-Friedrichs (LF) Hamiltonian, which provides an artificial viscosity approximation for solving static Hamiltonian-Jacobi equations. At a high level, the algorithm presented in [35] initializes the log-depth values $v(x,y)$ to a large positive constant and proceeds to iteratively update these values to progressively closer depths. We refer the reader to [35] for the full algorithm of the fast-sweeping scheme, as the order of sweeping directions, treatment of boundary conditions, and convergence criterion presented in [35] are the same as for our method.

2) Iterative Update Scheme: We now describe our iterative updating scheme. Our method has an EM flavor in the sense that it iterates a step optimizing a set of parameters (the reflectance model) based on the existing surface followed by a step computing expected depths using these parameters.

Algorithm 1 Shape-from-Motion-and-Shading

Input: An endoscopic image F_i and the associated 3D SfM points C_i
1. Warping $S^n_{warp}(x,y) = \rho(x,y)S^n_{warp}(x,y)$
2. Reflectance model estimation $E(\Theta) = \sum_\Omega \left(I_r(x,y) - I_{\text{est}}(x,y;\Theta)\right)^2$
3. Solve the SfS PDE using the estimated reflectance model parameters Θ and the warped surface S^n_{warp} to generate a newly estimated surface f_{n+1}
4. Re-warp f_{n+1} and repeat step 1-3

The proposed algorithm takes as input an observed image and the 3D SfM points associated with that image. It outputs a dense surface using depth-correcting warpings, the proposed reflectance model, and the proposed PDE framework.

Warping. We denote the warped surface at iteration n of our scheme as S^n_{warp}. For initialization, we define an estimated surface S^0_{warp} having $r(x,y) = 1$, where r is defined in Eq. (6). First, we perform an image-space warp of S^n_{warp} using the 3D SfM points with known distance $\hat{r}_i(x_i,y_i)$ as control points. For each SfM point, we estimate the ratio $\rho_i = \hat{r}_i/r_i$, where r_i is the point's (bilinearly interpolated) distance on S_n. To minimize the effect of outlier points from SfM, we adopt a nearest-neighbor approach to define our warping function: For each pixel (x,y) in the image, we compute the

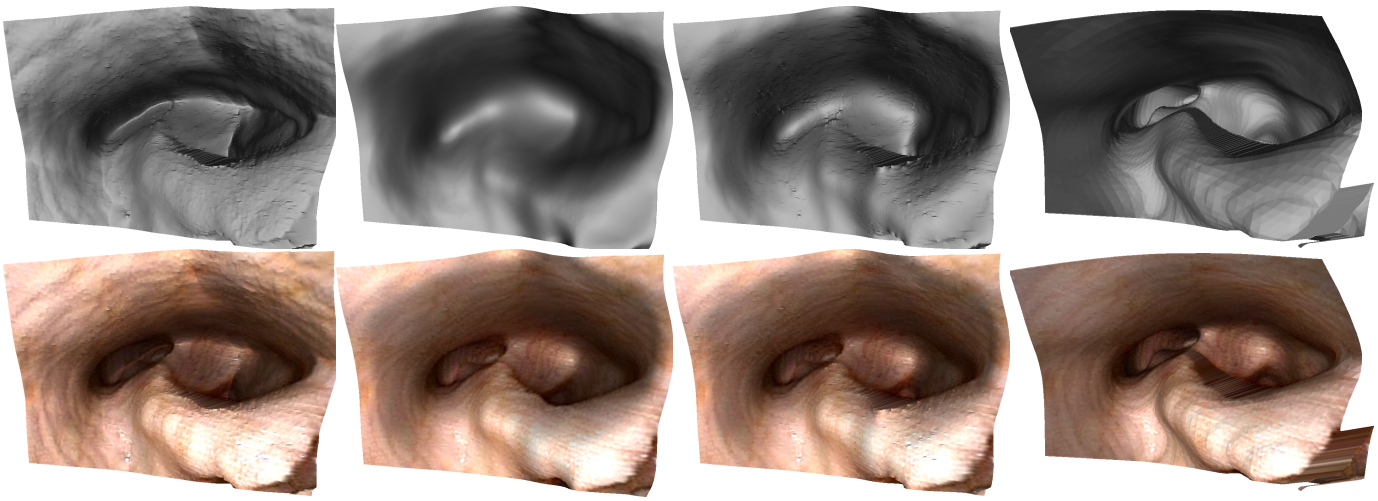

Fig. 9. Visual comparison of surfaces generated by our approach for an image from our ground truth dataset. Top/bottom rows: Visualization of the surface without/with texture from the original image. Columns from left to right: (1) using a Lambertian BRDF, (2) using our proposed BRDF ($K = 2$) without image-weighted derivatives, (3) using our proposed BRDF ($K = 2$) with image-weighted derivatives, and (4) the ground-truth surface. Note the oversmoothing along occlusion boundaries in column (2) versus column (3).

N closest SfM points in the image plane. In our experiments, we use $N = 10$. Then, we define the warp function at that pixel as $\rho(x, y) = \sum w_i \rho_i / \sum w_i$, where the sums are over the neighboring SfM points. We set $w_i = \exp(-d_i)$, where d_i is the distance in the image plane between (x, y) and the SfM point (x_i, y_i). The new surface is calculated as $S^n_{warp}(x, y) = \rho(x, y) S^n_{warp}(x, y)$.

Reflectance Model Estimation. From this warped surface, we optimize reflectance model parameters Θ for the specified BRDF (where the parameters depend on what BRDF we choose). This optimization is done by minimizing the least-squares error

$$E(\Theta) = \sum_{\Omega} \left(I_r(x, y) - I_{\text{est}}(x, y; \Theta) \right)^2 , \tag{15}$$

where $I_{\text{est}}(x, y; \Theta)$ is the estimated image intensity (see Eq. (3)) as determined by S^n_{warp} and the estimated BRDF.

SfS. Following reflectance model estimation, we apply the PDE framework introduced above (Eq. (14)) using the warped surface S^n_{warp} for values of z_{est} and using the current estimated reflectance model.

Concerning values of $\lambda(x, y)$ in our PDE, $\lambda > 1$ will give greater weight to S^n_{warp}, while $\lambda < 1$ will favor a purely SFS solution. We decide the weighting based on agreement between the SfM points and S^n_{warp}. Let Δr_i be the distance between a 3D SfM point with distance $\hat{r}_i$ and its corresponding point on S^n_{warp}. We define the agreement between the warped surface and the SfM point as $\lambda_i = \log_{10} \hat{r}_i / 2\Delta r_i$. This equally weights SfM and SFS (*i.e.,* $\lambda_i = 1$) when Δr_i is 5% of $\hat{r}_i$. The log term serves to increase λ_i by 1 for every order-of-magnitude decrease in $\Delta r_i / \hat{r}_i$. Just as for $\rho(x, y)$ above, we use the same nearest-neighbor weighting scheme to define $\lambda(x, y)$ based on the λ_i values at the SfM control points.

Iteration. Once SFS has been performed, we have a newly estimated surface S^{n+1}_{est}. Then, we simply re-warp the surface, re-estimate the reflectance model, and re-run regularized SFS. This iterative process is repeated for a maximum number of iterations or until convergence.

3) KLT and Optical Flow-based Correspondence Detection and Tracking: We introduced our SfMS 3D reconstruction algorithm in the previous subsection. As we mentioned, SfM is used to provide prior knowledge on depth that constrains surface geometry and reflectance model estimation. Therefore, a better SfM result can lead to more accurate dense surface reconstruction.

General purpose SfM methods are designed for 3D reconstruction of unordered images. Thus, feature-based (SIFT or ORB features) localization methods are usually used. However, these methods are difficult to generalize to endoscopic videos because endoscopic images are extremely low-textured. Therefore, in order to produce more robust correspondence matching results, we leverage the temporal coherent constraints by using a KLT tracker. However, there are still cases that a simple KLT tracker cannot handle: temporal gaps. The aforementioned non-informative frame removal step in video preprocessing will sometimes result in temporal gaps. This can be understood as a short-term loop closure problem. This section presents a method that solves this problem and augments the tracking-based correspondence matching.

A common tracking algorithm is shown in Algorithm 2.

The $track$ function is a KLT tracker. Each keypoint is tracked from F_i to F_{i+1} using Lucas-Kanade optical flow. The resulting position is then tracked back from F_{i+1} to F_i. If the point comes back to the original position in F_i, the tracking will be considered successful and the position in F_{i+1} will be added into $P^{tracked}_{i+1}$.

In order to solve short-term loop closure problem, we improved upon Algorithm 2 by using a frame-skipping strategy. The algorithm detail is shown in Algorithm 3. The main idea

Algorithm 2 strictly sequential tracking

$i = 1$
while $i \leqslant N_F$ **do**
  if $N_i^{tracked} < N_P$ **then**
    detect $N_P - N_i^{tracked}$ keypoints outside the neighborhoods of $P_i^{tracked}$.
    add P_i^{new} to P_i.
  end if
  $track(F_i, F_{i+1}, P_i) \rightarrow P_{i+1}^{tracked}$
  $i = i + 1$
end while

is to track not only the immediate next frame but also track the frames after it. Each frame maintains a set of unique keys that appears in the frame. In the meanwhile, a global hash table is also maintained to record for each unique point the frames it has appeared in. The purpose of using unique keys is to save the computation if a keypoint's successor is already tracked from a even earlier frame. Therefore, for a unique point u_k of associated with m_k keypoints, only $m_k - 1$ trackings will be performed.

Algorithm 3 frame-skipping tracking

$i = 1, n = (default)10, BackTrack = (default)False$
for $i = 1$ to N_F **do**
  for $j = i + 1$ to $min(i + n, N_F)$ **do**
    if $BackTrack$ and $P_j^{tracked}$ is not $\emptyset$ **then**
      find all points in $P_j^{tracked}$ whose unique keys are not in $P_i^{tracked} \rightarrow P_j^c$
      $track(F_j, F_i, P_j^c) \rightarrow$ add to $P_i^{tracked}$. Tracked keypoints inherent the unique keys of their origins.
    end if
  end for
  if $N_i^{tracked} < N_P$ **then**
    detect $N_P - N_i^{tracked}$ keypoints outside the neighborhoods of $P_i^{tracked} \rightarrow P_i^{new}$.
    add P_i^{new} to P_i.
    create a unique key for each keypoint in P_i^{new}.
  end if
  for $j = i + 1$ to $min(i + n, N_F)$ **do**
    find all points in P_i whose unique key is not in $P_j^{tracked} \rightarrow P_i^c$
    $track(F_i, F_j, P_i^c) \rightarrow P_j^{tracked}$. Tracked keypoints inherent the unique keys of their origins.
  end for
end for

C. Recurrent Neural Network (RNN)-based 3D reconstruction

We introduced a temporally local frame-by-frame 3D reconstruction method named SfMS in previous section that can estimate camera poses and dense depth maps for all keyframes. SfMS involves solving a large non-linear optimization and complex partial differential equations, so it can only be performed in an offline manner. However, in some applications, such as colonoscopy, a real-time 3D reconstruction is required because all the analysis need to be done during the procedure. In addition, human tissue has rich texture and complex reflectance properties, which cannot be adequately modeled using the BRDF introduced in SfMS. Therefore, we developed an RNN-based depth estimation method named DenseSLAMNet [Rui ECCV submission] that implicitly models the complex tissue reflectance property and performs depth estimation in real-time.

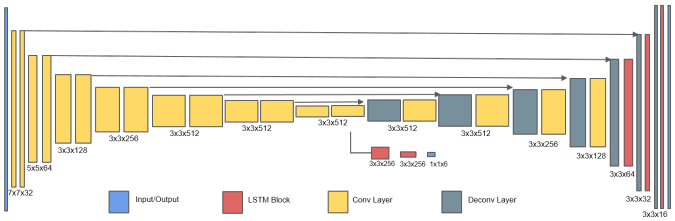

Fig. 10. (Best viewed in color) Our network architecture at a single time step. We use the DispNet architecture. The width and height of each rectangular block indicates the size and the number of the feature map at that layer. Each increase and decrease of size represents a change factor of 2. The first convolutional layer has 32 feature maps. The kernel size for all convolution layers is 3, except for the first two convolution layers, which are 7 and 5, respectively.

Figure 10 shows the network architecture of DenseSLAMNet. In our DenseSLAMNet, multiple views are incorporated into the single frame depth estimation through RNN. We use a temporal window of size $t = 10$ during training: every ten consecutive frames are grouped into one training sample and fed to the DenseSLAMNet sequentially.

Once the network is trained, video frames can be fed to it sequentially and the DenseSLAMNet will output the dense depth map and relative camera pose for each input frame.

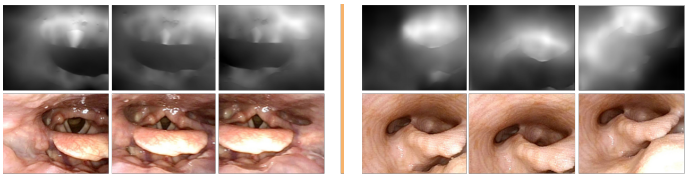

Fig. 11. Example of estimated dense depth maps of nasopharynoscopic images using the DenseSLAMNet

Figure 11 shows an example of estimated dense depth maps of nasopharynoscopic images using the DenseSLAMNet.

D. Deformable Surface registration

To fuse multiple frame-by-frame 3D reconstructions from SfMS into an endoscopogram, we use a novel groupwise surface registration algorithm involving N-body interaction. This algorithm is described in [4] and is based on Zhao *et al.* [37]'s pairwise surface registration algorithm, Thin Shell Demons. Here we only give an overview.

1) Thin Shell Demons: Thin Shell Demons is a physics-motivated method that uses geometric virtual forces and a thin shell model to estimate surface deformation. The geometric virtual forces $\{f\}$ are defined as vectors connecting vertex pairs $\{u^k, v^k\}$ between two surfaces $\{S_1, S_2\}$ (we use

k here to index correspondences). The correspondences are automatically computed using geometric and texture features. The thin shell model is a physical model which regularizes the non-parametric deformation vector field $\phi : S_1 \rightarrow S_2$. Combining these two, the algorithm is defined as an iterative energy minimization function

$$E(\phi) = \sum_{k=1}^{M} c(v^k)(\phi(v^k) - f(v^k))^2 + E_{shell}(\phi), \qquad (16)$$

where $c(v^k)$ is the confidence score based on the feature distance and E_{shell} is the thin shell deformation energy.

2) N-body Surface Registration: The endoscopogram requires registration of multiple partial surfaces. As an extension to the pairwise Thin Shell Demons, Zhao *et al.* [4] proposed a groupwise deformation scenario in which: N surfaces are deformed under the influence of their mutual forces. Mutual forces are defined as virtual forces that attract one surface by all the other surfaces. In other words, the deformation of a single surface is determined by the overall forces exerted on it. Such groupwise attractions bypass the need of a target mean.

3) Orthotropic Thin Shell Elasticity Estimation: The thin shell model that originally introduced by Zhao *et al.* assumes uniform isotropic elasticity, which contradicts human tissue elasticity being not only inhomogeneous but also anisotropic. Therefore, in order to better simulate the tissue deformation and produce more accurate registration results, Zhao recently [5] presented a statistical method that jointly estimates both the non-uniform anisotropic elasticity parameters and the material deformations from (within endoscopy deformations). As shown in figure 12, at each vertex on the surface model a canonical orthotropic model is formed by estimating the direction of its natural axes and the elasticity parameters along each axis. The estimated inhomogeneous and anisotropic elasticity parameters is shown to improve the surface registration accuracy and can help in studying within-patient deformations.

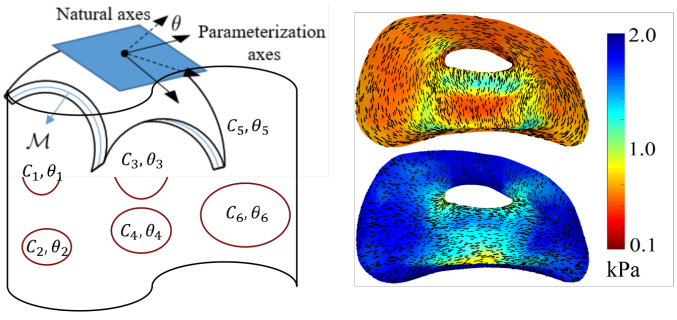

Fig. 12. Example of orthotropic elasticity estimation at each vertex on surface

E. Fusion-Guided SfMS

In the SfMS reconstruction method introduced in section IV-A there are no temporal constraints between successive frame-by-frame reconstructions. This fact and the method's reliance on reflectance model initialization lead to inconsistent reconstructions and even failure to reconstruct some frames, as shown in figure(example). As a result, manual intervention is needed for selecting partial surface reconstructions for fusion.

Wang *et al.* [38] introduced a method named fusion-guided SfMS that solves the inconsistency problem in the SfMS method so that longer sequences can be fused together without any manual intervention. The main idea of the method is to produce a single "reference model" which can be consistently used as a guidance across all frame-by-frame reconstructions so that temporal constraints are imposed among them. Such a reference model, S_{fused}, is used in Fusion-guided SfMS.

The multiple frame-based surfaces warped to fit the SfM points, $\{S^{i,j}_{warp}|j = 1, ..., n\}$ (see section IV-D), are fused to form S_{fused}. This is done using groupwise TSD. Then for each frame, a depth map that corresponds to its camera position is extracted from S_{fused} for reflectance model estimation. In such a way, all the single frame reconstructions are using a same reference surface as their a prior for reflectance model estimation so more coherent results are generated. Figure 13 shows an example of the inconsistency problem being solved by the fusion-guided SfMS method.

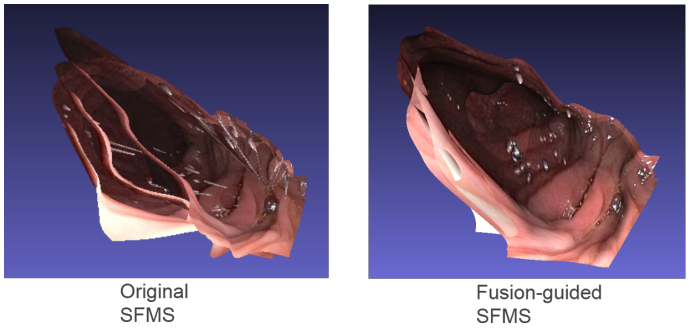

Fig. 13. Example of fusion-guided SfMS

This fits naturally to the iterative process of SfMS algorithm that descried in 1. At each iteration i a new fused reference surface S^i_{fused} is generated by fusing $\{S^{i,j}_{warp}|j = 1, ..., n\}$ together.

F. Seamless Texture Fusion

The endoscopogram is generated by fusing both the geometry and texture from the multiple partial reconstructions. Here we present the method for fusion of the texture maps acquired from different views. Dramatically changing illumination (light binding with camera), reflection and surface deformation in endoscopic video make this problem non-trivial. The illumination changes in endoscopic images are huge even for subtle camera motions.

Therefore, we need to derive a texture map from the various frames but avoid the dramatic color differences caused by the challenges we just mentioned.

Our approach has two stages. In the first stage an initial texture is created: for each voxel on the endoscopogram surface we select the image whose reconstruction has the closest distance to that voxel to color it. A Markov Random Field (MRF) based regularization is used to make the pixel selection more spatially consistent, resulting in a texture map

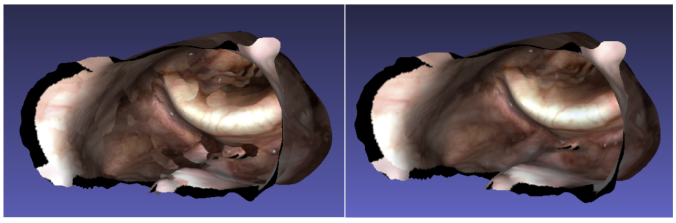

Fig. 14. Example of our seamless texture fusion. Left: Initial pixel selection result. Right: Seamless texture fusion result.

that has multiple patches with clear seams, as shown in Figure 14.

Then in the second stage, to generate a seamless texture, we minimize within-patch intensity gradient magnitude differences and inter-patch-boundary color differences.

1) Initial pixel selection and seam placement: In the fusion process used to form the endoscopogram each frame has been registered onto it. At each endoscopogram vertex $S(i)$ one of these registered frame-based surfaces $S_k^{'}$ is closest. To begin the initialization, the color from this frame is transferred to form the initial texture map for the endoscopogram. However, the irregularity of such selection results in extreme patchiness. Thus, we add a regularity energy term that depends on the labels in the local neighborhood. Then for each pixel on the endoscopogram the scheme selects the frame index k providing the color as follows:

$$\begin{aligned} D_k(i) &= \min_{j \in S_k^{'}} d(S(i), S_k^{'}(j)) \\ M(k) &= \arg\min_{k \in L} \sum_{i \in S} (D_k(i) + \lambda N_{k,i}) \end{aligned}$$

where $D_k(i)$ is the minimum distance from the surface $S_k^{'}$ to the i^{th} point on the surface S, where $N_{k,i}$ is the number of voxels in the neighboring voxel $S(i)$ that have different labels from the label k, where $k \in 1...N$ indicates the frame indices, and where M is the initial fused texture map. Such a setup is often called a Markov Random Field.

2) Texture fusion by minimizing within-patch and inter-patch differences: In this subsection we explain how the texture map M resulting from step 1 is modified through an iterative optimization to produce a seamless texture.

Let F be the set of images used to create the fused texture map. Let I_k be a single image in F. Let ω_k be all the pixels in image k that are selected to color M. We create a list ϕ that is composed of pairs of adjacent pixels in M that come from a different lighting condition, i.e., are members of different sets ω_k.

The fused texture should have low within-patch intensity gradient magnitude difference. The intuition is that the fused image should have the same details as the original images. The fused texture should also have low inter-patch-boundary color differences. Thus we wish to minimize

$$L_A = f + \lambda g + \mu ||g||^2 \tag{17}$$

where f sums the within-ω_k intensity gradient magnitudes squared and g sums the color difference magnitudes squared of pixel pairs in ϕ. That is,

$$f = \sum_{k \in F} \sum_{i \in \omega_k} || \bigtriangledown M(C(I_k(i))) - \bigtriangledown I_k(i)||_2^2 \tag{18}$$

where $I_k(i)$ is the i^{th} pixel in frame k that used to form texture map M. $C(I_k(i))$ is the coordinate in M corresponding to pixel $I_k(i)$; and

$$g = \sum_{(i,j) \in \phi} ||M(i) - M(j)||_2^2 \tag{19}$$

We use an augmented Lagrangian method to solve the optimization problem in equation 17 iteratively.

G. The Endoscopogram-to-CT Registration

After a complete endoscopogram is generated using our 3D reconstruction and groupwise geometry fusion algorithms, we can now register it to CT to achieve the fusion between endoscopic video and CT. To allow a good initialization of the registration, we first extract the tissue-gas surface from the CT and then do a surface-to-surface registration between the endoscopogram and the surface derived from the CT.

As discussed in section II, the registration between the endoscopogram and the CT extracted surface have the following challenges (1) the surface extracted from endoscopy suffers from serious missing patches due to some anatomy being not able to show up in the camera view; (2) the partial volume effect in CT leads to large topology differences between CT and endoscopogram; (3) a large anatomic deformation between CT and endoscopy results from patient posture differences and the introduction of the endoscope.

Our solutions to the above challenges is (1) using the thin shell demons registration algorithm, which is presented in detail in section IV-B, that is robust to missing surface and large topology changes; (2) applying the anisotropic elasticity parameters estimated in the groupwise registration to the endoscopogram to CT registration, which is presented in section IV-C; (3) using an expectation-maximization algorithm to estimate incompatible regions. Here we explain the incompatible regions estimation.

Because there are missing patches and topology differences between CT and endoscopogram surfaces that cause some points on either surfaces not to correspond to any point on the other surface, we must explicitly determine these patches, lest they be wrongly matched to the regions with highest matching scores. Such wrong matching will cause wrong attraction forces being generated during the registration.

Disparity Estimation. In order to solve this problem, we use a binary indicator function that indicates whether a point has a corresponding point or not. We jointly estimate the indicator function and the deformation variable iteratively using an EM algorithm. Let Ξ_1 and $\imath_2$ be the indicator functions for surfaces S_1 and S_2 respectively. The function value (0 or 1) indicates whether a vertex has a correspondence in the other surface; that is, $\Xi_1(x) = 0$ means $S_1(x)$ does not have a correspondence in S_2. The E-step in disparity estimation

computes the conditional distribution on Ξ_1, Ξ_2 given the current estimate of the deformations ϕ_1, ϕ_2:

$$p(\Xi_i|S_i, \phi_i^j) = p(S_i|\phi_i^j, \Xi_i)p(\Xi_i) \tag{20}$$

The likelihood term $p(S_i|\phi_i^j, \Xi_i)$ models how good the deformations align the compatible regions between the two surfaces. Mathematically, given the two deformed surfaces $S_1^{'} = S_1 \circ \phi_1$, $S_2^{'} = S_2 \circ \phi_2$ and their closest points on the other surfaces $C_1(x)$, $C_2(x)$,

$$p(S_i|\phi_i^j, \Xi_i) = \frac{1}{Z_0} exp(-\gamma L(S_i, \phi_i, \Xi_i)) \tag{21}$$

$$\begin{aligned} L(S_i, \phi_i, \Xi_i)) &= \sum_{x \in S_1} (\Xi_1(x) \cdot ||S_1^{'}(x) - C_1(x)||^2) \\ &= + \sum_{x \in S_2} (\Xi_2(x) \cdot ||S_2^{'}(x) - C_2(x)||^2) \end{aligned} \tag{22}$$

where the squared distance $||S_1^{'}(x) - C_1(x)||^2$ measures how well the alignment is. The M-step in this indicator function and deformation variable estimation algorithm is simply a TSD registration with attraction forces applied on compatible regions specified by the indicator functions. The algorithm initializes the two indicator functions with all ones and then iterates between the M-step and E-step until convergence.

Tumor Transfer.

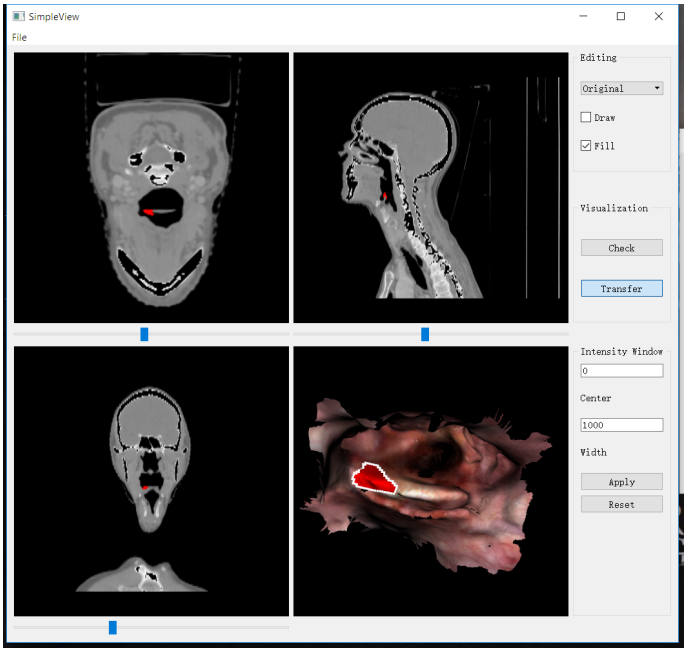

Fig. 15. Example of an ROI being drawn on the endoscopogram surface and transferred to CT image. The user drawn ROI is shown as a red region surrounded by a white contour in the lower right window.

Having the endoscopogram surface being registered to the CT extracted surface, we have created a tool for the physicians to directly draw on the endoscopogram surface. The highlighted region can then be displayed on the CT image as well as each individual endoscopic frames. Figure 15 shows an example of an ROI being drawn on the endoscopogram surface and transferred to CT image.

Acknowledgments: Bhisham Chera, Tong Zhu, Joel Tepper This work was under the partial support of NIH grant R01 CA158925. NVIDIA.

REFERENCES

[1] R. Wang *et al.*, "Recurrent Neural Network for Learning DenseDepth and Ego-Motion from Video," *ArXiv e-prints*, May 2018. 1
[2] R. J. Schwab *et al.*, "Dynamic imaging of the upper airway during respiration in normal subjects," *Journal of Applied Physiology*, vol. 74, no. 4, pp. 1504–1514, 1993. 2
[3] S. M. Kim *et al.*, "Pharyngeal pressure analysis by the finite element method during liquid bolus swallow," *Annals of Otology, Rhinology & Laryngology*, vol. 109, no. 6, pp. 585–589, 2000. 2, 5
[4] Q. Zhao *et al.*, "The endoscopogram: A 3d model reconstructed from endoscopic video frames," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 439–447. 2, 8, 9
[5] ——, "Orthotropic thin shell elasticity estimation for surface registration," in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 493–504. 2, 9
[6] M. Salzmann and P. Fua, "Deformable surface 3d reconstruction from monocular images," *Synthesis Lectures on Computer Vision*, vol. 2, no. 1, pp. 1–113, 2010. 2
[7] C. Wu *et al.*, "Shading-based dynamic shape refinement from multi-view video under general illumination," in *International Conference on Computer Vision (ICCV)*, 2011. 2
[8] Y. Han *et al.*, "High quality shape from a single rgb-d image under uncalibrated natural illumination," in *International Conference on Computer Vision (ICCV)*, 2013. 2
[9] M. Zollhöfer *et al.*, "Shading-based refinement on volumetric signed distance functions," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, 2015. 2
[10] H. N. Tokgozoglu *et al.*, "Color-based hybrid reconstruction for endoscopy," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012. 2
[11] A. Bartoli *et al.*, "Shape-from-template," *Pattern Analysis and Machine Intelligence (PAMI)*, vol. 37, no. 10, pp. 2099–2118, 2015. 2
[12] A. Malti *et al.*, "Template-based conformal shape-from-motion from registered laparoscopic images," in *Conference on Medical Image Understanding and Analysis (MIUA)*, 2011. 2
[13] ——, "Template-based conformal shape-from-motion-and-shading for laparoscopy," in *Information Processing in Computer-Assisted Interventions (IPCAI)*, 2012. 2, 4
[14] A. Malti and A. Bartoli, "Combining conformal deformation and cook–torrance shading for 3-d reconstruction in laparoscopy," *Biomedical Engineering, IEEE Transactions on*, vol. 61, no. 6, pp. 1684–1692, 2014. 2, 4
[15] A. Elad and R. Kimmel, "On bending invariant signatures for surfaces," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 25, no. 10, pp. 1285–1295, 2003. 3
[16] A. M. Bronstein *et al.*, "Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching," *Proceedings of the National Academy of Sciences*, vol. 103, no. 5, pp. 1168–1172, 2006. 3
[17] X. Gu *et al.*, "Genus zero surface conformal mapping and its application to brain surface mapping," *IEEE Transactions on Medical Imaging*, vol. 23, no. 8, pp. 949–958, 2004. 3
[18] J. Sun *et al.*, "A concise and provably informative multi-scale signature based on heat diffusion," in *Computer graphics forum*, vol. 28, no. 5. Wiley Online Library, 2009, pp. 1383–1392. 3
[19] T. Gatzke *et al.*, "Curvature maps for local shape comparison," in *Shape Modeling and Applications, 2005 International Conference*. IEEE, 2005, pp. 244–253. 3
[20] A. Zaharescu *et al.*, "Surface feature detection and description with applications to mesh matching," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 373–380. 3
[21] Y. Zeng *et al.*, "A generic deformation model for dense non-rigid surface registration: A higher-order mrf-based approach," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3360–3367. 3
[22] M. Bauer and M. Bruveris, "A new riemannian setting for surface registration," *arXiv preprint arXiv:1106.0620*, 2011. 3
[23] J.-P. Thirion, "Image matching as a diffusion process: an analogy with maxwell's demons," *Medical image analysis*, vol. 2, no. 3, pp. 243–260, 1998. 3

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 3

[25] N. Mayer *et al.*, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4040–4048. 3

[26] E. Ilg *et al.*, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2017. 4

[27] L. Maier-Hein *et al.*, "Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery," *Medical image analysis*, vol. 17, no. 8, pp. 974–996, 2013. 4

[28] B. K. Horn, "Shape from shading: A method for obtaining the shape of a smooth opaque object from one view," Dissertation, Massachusetts Institute of Technology, 1970. 5

[29] R. Zhang *et al.*, "Shape-from-shading: a survey," *Pattern Analysis and Machine Intelligence*, vol. 21, no. 8, pp. 690–706, 1999. 5

[30] E. Prados and O. Faugeras, "Shape from shading," in *Handbook of mathematical models in computer vision*, N. Paragios *et al.*, Eds. Springer, 2006, pp. 375–388. 5

[31] J.-D. Durou *et al.*, "Numerical methods for shape-from-shading: A new survey with benchmarks," *Computer Vision and Image Understanding*, vol. 109, no. 1, pp. 22–43, 2008. 5

[32] E. Prados and O. Faugeras, "Shape from shading: a well-posed problem?" in *Computer Vision and Pattern Recognition (CVPR)*, 2005. 5, 6

[33] R. L. Cook and K. E. Torrance, "A reflectance model for computer graphics," *ACM Transactions on Graphics*, vol. 1, no. 1, pp. 7–24, 1982. 5

[34] J. J. Koenderink *et al.*, "Bidirectional reflection distribution function expressed in terms of surface scattering modes," in *European Conference on Computer Vision (ECCV)*, 1996. 5

[35] A. H. Ahmed and A. A. Farag, "A new formulation for shape from shading for non-lambertian surfaces," in *Computer Vision and Pattern Recognition (CVPR)*, 2006. 6

[36] C. Y. Kao *et al.*, "Lax–friedrichs sweeping scheme for static hamilton–jacobi equations," *Journal of Computational Physics*, vol. 196, no. 1, pp. 367–391, 2004. 6

[37] Q. Zhao *et al.*, "Surface registration in the presence of missing patches and topology change." in *MIUA*, 2015, pp. 8–13. 8

[38] R. Wang *et al.*, "Improving 3d surface reconstruction from endoscopic video via fusion and refined reflectance modeling," in *Medical Imaging 2017: Image Processing*, vol. 10133. International Society for Optics and Photonics, 2017, p. 101330B. 9