

SCENE RECONSTRUCTION BEYOND STRUCTURE-FROM-MOTION AND
MULTI-VIEW STEREO

True Price

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computer Science.

Chapel Hill
2019

Approved by:

Jan-Michael Frahm

Henry Fuchs

Tamara Berg

Enrique Dunn

Ricardo Martín-Brualla

© 2019
True Price
ALL RIGHTS RESERVED

ABSTRACT

True Price: Scene Reconstruction Beyond Structure-from-Motion and Multi-View Stereo
(Under the direction of Jan-Michael Frahm)

Image-based 3D reconstruction has become a robust technology for recovering accurate and realistic models of real-world objects and scenes. A common pipeline for 3D reconstruction is to first apply Structure-from-Motion (SfM), which recovers relative poses for the input images and sparse geometry for the scene, and then apply Multi-view Stereo (MVS), which estimates a dense depthmap for each image. While this two-stage process is quite effective in many 3D modeling scenarios, there are limits to what can be reconstructed. This dissertation focuses on three particular scenarios where the SfM+MVS pipeline fails and introduces new approaches to accomplish each reconstruction task.

First, I introduce a novel method to recover dense surface reconstructions of endoscopic video. In this setting, SfM can generally provide sparse surface structure, but the lack of surface texture as well as complex, changing illumination often causes MVS to fail. To overcome these difficulties, I introduce a method that utilizes SfM both to guide surface reflectance estimation and to regularize shading-based depth reconstruction. I also introduce models of reflectance and illumination that improve the final result.

Second, I introduce an approach for augmenting 3D reconstructions from large-scale Internet photo-collections by recovering the 3D position of transient objects — specifically, people — in the input imagery. Since no two images can be assumed to capture the same person in the same location, the typical triangulation constraints enjoyed by SfM and MVS cannot be directly applied. I introduce an alternative method to approximately triangulate people who stood in similar locations, aided by a height distribution prior and visibility constraints provided by SfM. The scale of the scene, gravity direction, and per-person ground-surface normals are also recovered.

Finally, I introduce the concept of using crowd-sourced imagery to create *living 3D reconstructions* — visualizations of real places that include dynamic representations of transient objects. A key difficulty here is that SfM+MVS pipelines often poorly reconstruct ground surfaces given Internet images. To address this, I introduce a volumetric reconstruction approach that leverages scene scale and person placements. Crowd simulation is then employed to add virtual pedestrians to the space and bring the reconstruction “to life.”

To Luci

Thank you for loving me, and thank you for being there.

In memory of Atul,

an excellent friend who made the perfect cup of tea.

ACKNOWLEDGEMENTS

First, I would like to thank my adviser, Jan-Michael Frahm, who has given me so many diverse opportunities to expand my areas of expertise, work with amazing collaborators across many different domains, and challenge myself with supremely interesting research endeavors. This dissertation would not exist without Jan's patience, commitment, and openness all these years. I would also like to thank the rest of my committee members, Henry Fuchs, Tamara Berg, Enrique Dunn, and Ricardo Martín-Brualla, for their support and feedback. And, I would be remiss if I did not thank Steve Pizer and Julian Rosenman for their guidance, positivity, and unwavering support. I battled so much insecurity and doubt after my first year in graduate school — not that the feeling ever goes away completely, of course! — and I do not think I would have made it without the guidance of Steve, Julian, and Jan-Michael.

Additionally, I would like to thank my labmates, collaborators, and friends at UNC, who made my time in graduate school superb: Matt Adams, Adam Aji, Phil Ammirato, Sarah Andrabi, Akash Bapat, Andy & Kat Best, Rohan Chabra, Young-Woon Cha, Andrew Chi, Michael Deakin, Joe Di Natale, Marc Eder, Cheng-Yang Fu, Rohit Gupta, Xufeng Han, Victor Heorhiadi, Junpyo Hong, Max Hudnell, Dinghuang Ji, Hadi Kiapour, Hyo Jin Kim, Andreas Kuhn, John Lim, Wei Liu, Conny Lu, Jisan Mahmud, Sahil Narang, Vicente Ordóñez-Román, Eunbyung Park, Ric Poirson, Srihari Pratapa, Helen Qin, Nick Rewkowski, Atul Rungta, Johannes Schönberger, Misha Shvets, Sirion Vittayakorn, Thanh Vu, Ke Wang, Rui Wang, Zhen Wei, Jan Werner, Yi Xu, Licheng Yu, Dongxu Zhao, Qingyu Zhao, Enliang Zheng, and Yipin Zhou. Thanks to all my other collaborators over the years, as well, including students, postdocs, and professors at UNC and abroad. If there is anyone I should have named but inadvertently left out, please know that it was not intentional! Feel free to give me grief for it. I would especially like to thank Akash, Johannes, and Marc for their frequent collaboration and friendship.

I have to extend huge thank you to all of the staff at UNC, as well. From IT staff like Murray Anderegg, Bil Hays, and John Sopko to administrators like Missy Wood, Jody Gregoritsch, and Denise Kenney, there are so many supportive individuals who have gone out of their way to ensure that everything runs smoothly (and, indeed, to keep me in line!). And a special thanks to Jim Mahaney, without whom several research projects would have been dead in the water.

Thanks to all of my co-workers and fellow interns at Organic Motion, Bosch, and Google. I was extremely lucky to get to work with such talented individuals on such exciting projects. These experiences were invaluable to my development as a professional. A special thanks to my mentors: Shaun Kime, Zhixin Yan, and Hugues Hoppe.

I also would like to thank my parents, Clay and MaryAnn, as well as my sister and her husband, Joanna and Jeff, who have supported me wholeheartedly for many years. Thanks to other longtime friends who have supported me from afar, including Austin, Jared, and Michelle. And finally, I could not have gotten this far in life without my partner, Luci, who has provided me with unconditional love and patience throughout the entire Ph.D process. Thank you for so many wonderful years. I am so excited to begin the next chapter of our lives, with our dog, Tavern, in tow.

TABLE OF CONTENTS

LIST OF TABLES	xii
LIST OF FIGURES	xiii
LIST OF ABBREVIATIONS	xvii
CHAPTER 1: INTRODUCTION	1
1.1 Thesis Statement	4
1.2 Outline of Contributions	4
CHAPTER 2: BACKGROUND AND RELATED WORK	6
2.1 Shape-from-Shading and Shading-based Surface Reconstruction	6
2.2 3D Reconstruction of Endoscopic Imagery	9
2.2.1 Monocular SfM, MVS, and SLAM Techniques	10
2.2.2 Shading-based Approaches and Reflectance Estimation Methods	11
2.2.3 Stereo Endoscopy and Other Semi-Controlled Capture Scenarios	12
2.2.4 Combined Sparse or Dense Reconstruction with Shading Estimation	14
2.2.5 Template-based Reconstruction and Alignment of Pre-operative 3D Scans ..	14
2.2.6 Monocular Depth Estimation via Convolutional Neural Networks	16
2.3 Modeling Transient Objects in Crowd-sourced Imagery	17
2.4 Crowd Simulation in Virtual Representations of Real Environments	19
CHAPTER 3: 3D RECONSTRUCTION OF ENDOSCOPIC VIDEO	21
3.1 Background	24
3.1.1 Reflectance Models	24
3.1.2 Surface Model for Shape-from-Shading	25

3.1.3	Structure-from-Motion	27
3.2	Method	28
3.2.1	Initial PDE	30
3.2.2	Regularization	30
3.2.3	Solving the PDE	32
3.2.3.1	Discretization	32
3.2.3.2	Applying the Lax-Friedrichs Hamiltonian	33
3.2.3.3	Fast Sweeping Scheme and Boundary Conditions	34
3.2.4	Computing $\sigma_{i,j}^x$ and $\sigma_{i,j}^y$ for Arbitrary BRDFs	35
3.2.5	Image-weighted Finite Differences	37
3.2.6	Reflectance Model	38
3.2.6.1	BRDF Basis	38
3.2.6.2	Proposed Reflectance Model	39
3.2.6.3	Relation to Other Reflectance Models	40
3.2.7	Iterative Update Scheme	42
3.2.7.1	Warping	43
3.2.7.2	Reflectance Model Estimation	43
3.2.7.3	SfS with Estimated BRDF	44
3.2.7.4	Iteration	44
3.2.8	Accounting for Interreflections in Real Endoscopic Scenarios	46
3.3	Evaluation	48
3.3.1	Comparison of BRDF Fits	48
3.3.2	Ground-truth Geometric Evaluation	52
3.3.3	Results on Patient Data	54
3.4	Discussion	54
CHAPTER 4: 3D RECONSTRUCTION OF TRANSIENT OBJECTS		60

4.1	Approach	62
4.1.1	Person Detection and Gravity Estimation	62
4.1.2	Voting-based Scale Estimation	65
4.1.3	Scale Refinement, Height Estimation, and Ground Surface Estimation	68
4.1.4	Ground Surface Reconstruction	72
4.1.5	Visualization	72
4.2	Evaluation	73
4.3	Ablative Analysis	78
4.3.1	Visibility Constraint During Voting	78
4.3.2	Effect of Scale Refinement Terms	79
4.3.3	Effect of Parameters during Refinement	79
4.3.4	Comparing Scale Voting and Scale Refinement	80
4.4	Discussion	80
CHAPTER 5: LIVING 3D RECONSTRUCTIONS		82
5.1	Robust Surface Reconstruction	84
5.1.1	Truncated Signed Distance Function Aggregation	84
5.1.2	Regularizing the Distance Field	87
5.1.3	Optimizing the Distance Field	90
5.1.4	Gravity-aligned Surface Prior	91
5.1.5	Scenario-specific Considerations and Implementation	93
5.2	Triangle Color Estimation and Walkable Area Extraction	94
5.3	Crowd Simulation and Visualization	96
5.4	Results	96
5.5	Discussion	97
CHAPTER 6: CONCLUSION AND FUTURE WORK		107
6.1	Future Directions	108

6.1.1	Extensions to Shading-based Endoscopic Reconstruction	108
6.1.2	Extensions to 3D Reconstruction of Transient Objects and Living 3D Reconstructions	110
APPENDIX A:	DERIVATION OF ARTIFICIAL VISCOSITY VALUES IN SFS PDE SOLUTION	113
REFERENCES	115

LIST OF TABLES

Table 3.1 – Radiance fitting accuracy for MERL materials. For each material, the yellow cell marks the smallest K for which the proposed model achieved a smaller error than the Phong model.	51
Table 3.2 – Accuracy of the proposed SfM+SfS approach for different reflectance models on simulated and real data across 100 images. Example renderings are show in the right column.	55
Table 3.3 – Ablation analysis of the proposed method on ground-truth endoscopic data with $K = 2$	59
Table 3.4 – Accuracy of the proposed SfM+SfS approach on real endoscopic video without accounting for surface interreflections.	59
Table 4.1 – Quantitative results on the proposed method for scale and placement. “% Error” gives the amount that the method over/under-estimated the distance of one unit in the reconstruction. n_p and n_c show the number of placed detected people and photographers, respectively, recovered by the method.	76
Table 4.2 – Ablative analysis on the importance of different parts of the proposed algorithm. GT: Ground-truth scene scales (reconstruction units per meter). Initial/Final: Estimates from the voting and refinement stages. No Height/Visib.: Height/visibility terms removed from final optimization. ± 10 : With all parameters modified by ten percent. Red cells: Results where the estimated length of one unit in the reconstruction was incorrect by $>10\%$	81

LIST OF FIGURES

<p>Figure 3.1 – Surface estimates using multi-view stereo reconstruction tend to be noisy and/or incomplete for endoscopic data. Top row: Fused point cloud obtained via MVS (Schönberger et al., 2016) and an untextured surface reconstructed from this point cloud using a method based on the Delaunay tetrahedralization approach of Labatut et al. (2009). Second row: Textured and untextured views of the surface obtained using Poisson surface reconstruction (Kazhdan and Hoppe, 2013) on the MVS point cloud. Bottom two rows: Same results for a different patient.</p>	23
<p>Figure 3.2 – An endoscopogram is constructed via non-rigid registration of multiple SfMS reconstructions of individual video frames.</p>	24
<p>Figure 3.3 – Structure-from-Motion results for endoscopic video. Individual 3D surface points (colored dots) and camera poses (blue) are jointly recovered.</p>	28
<p>Figure 3.4 – Diagram of the proposed iterative approach for dense surface reconstruction of a single video frame.</p>	29
<p>Figure 3.5 – Compared to the ground-truth surface (left), the boundary conditions suggested by Kao et al. (2004) can lead to strong artifacts on the edges of the image for endoscopic applications (second from left). A minor change to these conditions can correct for this, although problematic artifacts can still occur (second from right), which can be alleviated by limiting the maximum surface slope along the image boundary (right).</p>	36
<p>Figure 3.6 – The 16 basis functions used in the proposed reflectance model with $K = 8$.</p>	41

Figure 3.7 – Example fitting results for the $K = 5$ model, using the ground-truth surface (left), warped surface (middle), and sparse SfM points (right) on a synthetic image. The top row shows the initial surfaces and points use for the fitting; these target values are scattered in the graphs in the bottom row, with each value colored by its observed intensity for visualization. The red curves in the bottom row plot the reflectance function $\eta(\theta_i; \Theta)$ whose parameters Θ have been robustly fit to the plotted points. The middle row shows a re-rendering of the ground-truth surface using these fit functions. Fitting to the SfM points alone is more reliable than fitting to the entire warped surface, which may contain errors in depth as well as in $\cos \theta_i$. In this example, the near-specular effects of the material are better captured by the fit using the SfM points.	45
Figure 3.8 – Estimated 1D radiance functions for different materials from the MERL database (Matusik et al., 2003). The top row of each image shows the color radiance, $\cos(\theta_i)\text{BRDF}_\lambda(\theta_i)$, and the second row shows the luminance equivalent. Subsequent rows show least-square fits to the luminance function for different BRDF models.	49
Figure 3.9 – Visual comparison of surfaces generated by the proposed approach for an image from a 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 the proposed BRDF ($K = 2$) without image-weighted derivatives, (3) using the 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) and the flattened curve of the epiglottis in column (1).	53
Figure 3.10 – Example results for three images from a live endoscopic video. Left: Original image. Right: Surface estimated from the image using the proposed algorithm.	56
Figure 3.11 – Example results for three images from a live endoscopic video. Left: Original image. Right: Surface estimated from the image using the proposed algorithm.	57
Figure 3.12 – Example results for three images from a live endoscopic video. Left: Original image. Right: Surface estimated from the image using the proposed algorithm.	58
Figure 4.1 – The pipeline of the proposed reconstruction system.	62

Figure 4.2 – To accurately localize 2D ground points for detected people, a planar torso model in 3D (left) is first fit to detected 2D neck, shoulder, and hip joints (middle-left). Right: Coordinate axes for the planar model.	65
Figure 4.3 – Scale scoring curve for a model of the Pantheon. The peak is chosen as the initial scale estimate.	68
Figure 4.4 – Overhead views (left) and sample renderings with ground and person avatars (middle) for the proposed method. Examples of real photos are shown on the right. The green dots in the overhead views show person placements, with cameras as red dots and detected people as green dots. Black dots show static structure. From top: Dubrovnik, Croatia; the Pantheon; San Marco Plaza, Venice; and the area around the Colosseum and Roman Forum in Rome.	74
Figure 4.5 – Overhead visualizations of person placements (left) versus aerial views from Google Earth (right). From top to bottom: Buckingham Palace, the Palace of Westminster, the Sacré Cœur in Paris, Trafalgar Square, and Trevi Fountain.	77
Figure 4.6 – Result of the scale voting scheme with (blue) and without (orange) the visibility constraint. The ground-truth scale is near 0.01 reconstruction units per meter.	78
Figure 4.7 – Ratio of the estimated neck distance $s N_i $ to the visibility threshold $v_i(s)$ for the ground-truth scale (GT), and for larger/smaller scales. Values are sorted and clipped to $[0.5, 1.5]$	79
Figure 5.1 – Left: A scene mesh generated by point cloud fusion from MVS depthmaps and Delaunay tetrahedralization with visibility optimization (Schönberger et al., 2016). Middle: Mesh for the same approach, with 3D ground points of detected people added to the point cloud. Right: Scene mesh generated by the proposed method.	83
Figure 5.2 – Initial labeled ground triangles (white) and additional triangles added after the proposed region-growing approach (red).	96
Figure 5.3 – Qualitative, ablative comparison of scene reconstruction results for the Tower of London under the proposed implementation at a voxel resolution of 1m.	98
Figure 5.4 – Reconstruction of Buckingham Palace.	99
Figure 5.5 – Reconstruction of the Castel Sant’Angelo in Rome.	100
Figure 5.6 – Reconstruction of the Notre Dame Cathedral in Paris.	101

Figure 5.7 – Reconstruction of the Old Town Square in Prague. 102

Figure 5.8 – Reconstruction of the Piazza San Marco in Venice. 103

Figure 5.9 – Reconstruction of the Sacré Cœur Basilica in Paris. 104

Figure 5.10 – Reconstruction of the Tower of London. 105

Figure 5.11 – Additional views of the reconstruction of the Tower of London. 106

LIST OF ABBREVIATIONS

BRDF	Bidirectional Reflectance Distribution Function
GT	Ground Truth
LF	Lax-Friedrichs
MVS	Multi-View Stereo
PDE	Partial Differential Equation
SfM	Structure-from-Motion
SfMS	Shape-from-Motion-and-Shading
SfS	Shape-from-Shading
TSDF	Truncated Signed Distance Function
TV	Total Variation

CHAPTER 1: INTRODUCTION

The observable world is made up of tangible materials with well-defined physical properties and spatial relationships. For example, we could identify a building as being covered in brick and having a specific length, width, and height in meters. The action of *observing* such an object, however, is an indirect process. For instance, an observer such as the human eye or a pinhole camera does not “see” a brick building in a three-dimensional sense, but rather collects information about the intensity and distribution of visible light rays irradiating and then reflecting from the building’s surface towards the observer. This is the driving problem of 3D reconstruction in computer vision: Given that an image only provides us with a 2D slice of visible light rays, how can we recover the underlying 3D surfaces that effected the image?

In a very general sense, the vast body of work in tackling the 3D reconstruction problem for visible light imagery can be divided into two categories: single-image reconstruction and multi-image reconstruction. Single-image reconstruction methods seek to recover a “depth map,” or “2.5D surface representation,” that defines the distance from the observer to the nearest physical surface for every viewing ray in the coordinate frame of the observer. This is an ill-posed problem without prior constraints, since an infinite number of underlying surfaces can generate a given image. One classical approach to this problem is known as Shape-from-Shading (SfS) (Horn, 1970; Zhang et al., 1999; Prados and Faugeras, 2005). To constrain the solution, SfS assumes the material reflectance properties of the underlying surface are known, and that the light source for the image is explicitly known. Since the strength of the reflecting light off the surface is proportional to the angles between the incoming light, the surface normal, and the direction to the observer, these conditions force the recovered surface to agree with the observed light intensities (shading) in the captured image. Apart from SfS, deep learning approaches have recently been introduced to predict surfaces from single

images; these approaches attempt to learn shape constraints automatically from 2D appearance by training on a large number of images, given ground-truth depth maps or corresponding stereo image views (Eigen et al., 2014; Garg et al., 2016; Godard et al., 2017; Zhou et al., 2017; Li and Snavely, 2018).

Multi-image reconstruction methods seek to recover the underlying surfaces of an environment using images taken from multiple vantage points within the space. Approaches in this vein typically consist of a “sparse, then dense” pipeline, although many variations exist for different reconstruction scenarios. The most general “sparse” reconstruction approach is known as Structure-from-Motion (SfM) (Pollefeys et al., 2004; Snavely et al., 2006, 2008; Frahm et al., 2010; Agarwal et al., 2011; Crandall et al., 2011; Wu, 2013; Wilson and Snavely, 2014; Heinly et al., 2015; Schönberger and Frahm, 2016).¹ Given a set of images, SfM aims to jointly recover camera intrinsics, relative image poses, and the 3D position for corresponding points in the individual images. The method is “sparse” because, rather than obtaining a surface with fixed fiducial sampling in the 3D world or 2D pixel space, points on the 3D structure are determined only for locations in the images with highly distinguishable 2D appearance. SfM is typically used as a preprocessing step for “dense” multi-image reconstruction techniques such as multi-view stereo (MVS) (Furukawa and Ponce, 2010; Furukawa et al., 2010, 2015; Schönberger et al., 2016). Like SfS, MVS recovers a depth map for each individual image. However, instead of using strong assumptions on illumination and surface conditions, MVS utilizes the fact that an image point will have similar appearance in nearby viewpoints if it is lifted into 3D and then reprojected into the other view. The correct underlying surface for an image, therefore, is determined by maximizing appearance similarity after reprojection. After multiple depth maps are obtained via MVS, a final model of the scene can be recovered by fusing and meshing these individual surfaces within the global space of the reconstruction (Curless and Levoy, 1996; Labatut et al., 2009; Jancosek and Pajdla, 2011; Schönberger et al., 2016).

¹Further discussion and earlier references are also provided by Hartley and Zisserman (2003, Chapter 18).

All three methods – SfS, SfM, and MVS – and their variants have limitations that impair or prohibit reconstruction in certain scenarios. For example, the equations governing SfS may be difficult to model for sufficiently complex surfaces and lighting conditions, and in any case, the underlying material reflectance properties must be well defined. SfM and MVS, on the other hand, strictly assume that the underlying 3D surfaces are stationary in all images; they are unable to handle dynamic objects. Although non-rigid SfM (NRSfM) with monocular (Bregler et al., 2000; Xiao et al., 2004; Akhter et al., 2009; Garg et al., 2013; Russell et al., 2014) and multi-view (Zheng et al., 2015; Ji et al., 2016; Innmann et al., 2019) formulations² have proven successful in certain reconstruction scenarios, rigid and non-rigid methods alike are strongly limited by the distinguishability of the underlying surface appearance and by the conditions in which the images were taken. For instance, homogeneous regions in images lack distinguishing texture and thus are difficult to reliably identify between images, which reduces reconstructability. Even for potentially well-textured surfaces, appearance can change due to the time of day or weather, and surfaces like the ground may only be captured from unfavorable angles in the majority of images; these imaging conditions frequently occur in Internet photo-collections (Kuhn et al., 2017). For dynamic object reconstruction, temporal sampling also comes into play, as existing approaches like NRSfM require the temporal order of images to be known and well-sampled. This works for reconstructing objects in video sequences, but for objects that are only imaged once, such as pedestrians in a temporally sparse photo-collection, it is necessary to develop methods that do not assume temporal contiguity. Finally, due to the properties of perspective projection, all reconstruction approaches are only accurate up to scale without additional prior knowledge on the expected size of imaged objects.

In this dissertation, I address 3D reconstruction scenarios with imaging conditions that are unfavorable due to an inability to leverage temporal consistency and/or due to insufficient surface texture for discriminative dense multi-view correspondence identification. Tackling these challenging reconstruction problems requires significant and novel adaptations to the traditional approaches listed above. Details of the individual research thrusts that support this thesis are provided below.

²I generally will only address unsynchronized multi-view scenarios in this dissertation, since approaches that assume synchronization usually employ a significantly expanded set of constraints for reconstruction.

1.1 Thesis Statement

In 3D reconstruction scenarios where the typical conditions of structure-from-motion and multi-view stereo are violated for specific objects or surfaces, more complete 3D representations can be obtained through additional processing that combines multi-view reasoning and scenario-specific constraints.

1.2 Outline of Contributions

The body of work in this dissertation covers multiple scenarios in 3D computer vision that traditional robust modeling techniques cannot handle:

3D Reconstruction of Endoscopic Video: I introduce a new approach for reconstructing dynamic, poorly textured surfaces inside the human body. To overcome the difficulties in this 3D modeling scenario, my method employs a combination of sparse 3D modeling, shading constraints, and integrated regression of surface reflectance parameters. This work, detailed in Chapter 3, has been partly described in several publications (Zhao et al., 2015, 2016; Wang et al., 2017). Chapter 3 also contains expanded research regarding the approach, detailing new aspects of the formulation and optimization that lead to improved accuracy.

3D Reconstruction of Transient Objects: I propose a novel approach for augmenting 3D reconstructions by recovering the 3D position of people in individual images in large-scale Internet photo-collections. Structure-from-Motion (SfM) and Multi-View Stereo (MVS) approaches cannot be directly used in this scenario, since no two unique images capture the same person in the exact same place. To overcome the difficulties in 3D modeling, my method reasons about possible 3D person placements according to how many people in different images would be placed “nearby” for a given scene scale. This work is described in Price et al. (2018) and detailed in Chapter 4.

Living 3D Reconstructions: Having obtained context of where people exist within a 3D environment by recovering the 3D position of people in individual images, I propose an extended reconstruction approach that aims to bring the virtual environment “to life.” The idea here is to add virtual pedestrians to the scene and animate them to walk around the environment. From a 3D reconstruction perspective, a key difficulty lies in how to define the walkable surfaces in the 3D environment, given that ground surfaces are largely unable to be captured using SfM or MVS. To recover the ground, I propose a volumetric approach that reconstructs all surfaces in the scene, with sparse person placements from the above approach (Price et al., 2018) guiding the ground surface reconstruction, and with additional modeling constraints on the scene. Color-based segmentation is used to delineate walkable ground regions in the scene, and crowd simulation is then applied to move virtual agents between different 3D locations of people detected in the individual images. Details of this approach are provided in Chapter 5.

CHAPTER 2: BACKGROUND AND RELATED WORK

The three following chapters of this dissertation generally fall into two major categories: performing 3D reconstruction on endoscopic video data using shading and structure, and modeling humans in 3D environments reconstructed from Internet photo-collections. In the following sections, I provide a broad general background of endoscopic reconstruction and outline approaches related to 3D reconstruction and human modeling in large-scale datasets.

2.1 Shape-from-Shading and Shading-based Surface Reconstruction

First introduced in the 1970 thesis of Horn (1970), Shape-from-Shading (SfS) 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 (Zhang et al., 1999; Prados and Faugeras, 2006; Durou et al., 2008). 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) (Zhang et al., 1999; Durou et al., 2008). The PDE formulation of SfS has received the most attention, starting with Prados and Faugeras (Prados and Faugeras, 2005), who introduced a novel, provably convergent approach for solving the problem as a PDE. Fast marching (Prados and Soatto, 2005; Tankus et al., 2005) and fast sweeping (Ahmed and Farag, 2006) methods have also been successfully applied to solve the SfS PDE problem. A major criticism of SfS has been it requires too-strong constraints on surface and lighting conditions (Durou et al., 2008). I present a complete formulation of SfS for endoscopy in Chapter 3 and address modeling considerations therein.

While older methods exist for SfS under general illumination and reflectance, for example the work of Zheng and Chellappa (1991) and Tsai and Shah (1994), these methods have long been known to perform poorly even given synthetic data, partially due to simplified assumptions of reflectance, lighting, and camera projection (Zhang et al., 1999). Traditional PDE formulations of SfS assume a Lambertian reflectance model for the scene (Visentini-Scarzanella et al., 2012), which may be a poor assumption for real-world data (Zhang et al., 1999; Ahmed and Farag, 2006). Some work has investigated non-Lambertian models for the SfS PDE formulation. Ahmed and Farag (2006) introduce a SfS method for the Oren-Nayar reflectance model, which describes reflectance for rough diffuse surfaces; the authors later demonstrated an approach for the Ward reflectance model (Ahmed and Farag, 2007). Vogel et al. (2009) present a method for SfS on Phong-type surfaces, which is itself an extension of the Lambertian model with added ambient and specular terms. Quéau et al. (2017) formulated a SfS PDE for scenes exhibiting known natural illumination and albedo. For endoscopic applications, I propose a reflectance model that subsumes the Lambertian and Phong models and, in general, is suitable for surfaces with arbitrary reflectance properties. To avoid a need to know the reflectance model *a priori*, I also introduce an approach for using Structure-from-Motion (SfM) to bootstrap reflectance model estimation and guide the SfS solution.

A number of methods have been proposed for jointly predicting a combination of surface reflectance, illumination, and/or shape for a single image. Barron and Malik (2014) formulated this problem as an “intrinsic image” technique, where shading is determined as a function of shape and illumination. Their method learns separate priors on reflectance, depth, and illumination and computes a maximum-likelihood solution with the constraint that an image rendered under the given parameters should appear as similar as possible to the observed image. Oxholm and Nishino (2015) proposed an approach that assumes a complete environment map is available for the space surrounding the object. By leveraging a directional-statistics BRDF model, their method is able to compute shape and reflectance for single- and multi-image capture scenarios. Johnson and Adelson (2011), in contrast, assume a known reflectance map but unknown natural illumination; Huang and

Smith (2011) utilize silhouette constraints to avoid the requirement of an explicit reflectance map. Both of these methods utilize the fact that diffuse surfaces act as low-order filters of the environment illumination, and thus the illumination can be approximated using low-order spherical harmonics.

Deep learning methods have much promise in robustly modeling the complex shading behaviors found in real-world applications of the inverse graphics problem. In particular, Li et al. (2018) recently introduced a convolutional neural network approach for jointly estimating albedo, specular roughness, surface normal, and depth for an object captured in a single flash-illuminated image. The approach also estimates environment illumination for the image and introduces an internal network architecture to recover images formed from multiple bounces of light off of the surface; an analytical rendering layer is used to produce the direct illumination image based on the regressed surface parameters. The network applies a multi-stage refinement of estimated parameters to achieve state-of-the-art recovery of shape and reflectance parameters from a single image.

Outside of endoscopic applications, many works on combining motion-based reconstruction with shading information have utilized shading to *augment* an existing shape template or model priors (Salzmann and Fua, 2010). Wu et al. (2011) proposed to first build coarse-scale dynamic models from multi-view video and then leverage shading to estimate fine-scale, temporally varying geometry. Fine-scale shading correction has also been used to refine dense surfaces obtained using a depth sensor (Han et al., 2013; Zollhöfer et al., 2015). Among multi-view methods that leverage shading directly in the shape estimation, Gallardo et al. (2016) introduced a template-based method for reconstructing low-texture deforming surfaces leveraging Lambertian shading constraints. More recently, the same authors introduced a template-free non-rigid SfM method (Gallardo et al., 2017) that considers Lambertian shading. Finally, the theoretical constraints on shape estimation with unknown reflectance under camera motion have been outlined in a number of works by Chandraker (2014a,b, 2015).

2.2 3D Reconstruction of Endoscopic Imagery

In Chapter 3 of this thesis, I motivate endoscopic surface reconstruction for 3D review during treatment planning. That is, given an endoscopic video, use an offline reconstruction process to build a textured surface model of the target area that a physician can for enhanced visualization, video review, or procedure post-analysis. In addition to treatment planning, the literature is rife with methods that target goals for augmented reality and real-time 3D applications during surgery. I address the general themes and research areas in this section.

To provide some historical context, methods for achieving 3D reconstruction of endoscopic imagery have existed for at least three decades, dating back at least to the early work of Badiqué et al. (1988) that investigated correlation-based matching and 3D visualization for stereoscopic endoscopy. The work of Oda et al. (1994, 1995a) was perhaps the first to outline a full approach for 3D reconstruction from monocular endoscopic video. Similar to the standard pipeline of today’s reconstruction methods, this method introduced a SfM-type sparse reconstruction approach with inter-frame feature tracking and proposed a method for patch-based multi-view depthmap estimation, with later extensions to estimate the scale of the reconstruction based on light intensity (Oda et al., 1995b). Perhaps the earliest applications of SfS for endoscopy were introduced by Deguchi (1996), Okatani and Deguchi (1997), and Yeung et al. (1999), who used a method for estimating equal-depth contours to recover shape from a single endoscopic image assuming a known — but material-agnostic — 1D Bidirectional Reflectance Distribution Function (BRDF). The first two works actually utilize multi-view information as part of the method, initializing the estimation using a sparse, multi-frame surface estimation algorithm (Deguchi et al., 1994). The third work is the first that I know of to empirically measure a BRDF for use in endoscopic SfS.

In the following subsections, I outline various approaches for endoscopic 3D reconstruction that have been proposed since these initial works. Controlled capture settings that allow for per-frame depth estimation have perhaps enjoyed the longest success in endoscopic surface reconstruction (Mountney et al., 2010; Lin et al., 2016). SfM and Simultaneous Localization and Mapping (SLAM) approaches, both sparse and dense, rigid and motion-compensating, have also been investigated.

Still other methods — *e.g.*, Shape-from-Template algorithms — have combined these approaches with pre-operative CT scans that serve as a shape prior for reconstruction or as a target space for reconstruction alignment. Finally, a number of alternative capture strategies such as range imaging and depth-from-focus have been proposed.

2.2.1 Monocular SfM, MVS, and SLAM Techniques

The vast majority of endoscopic procedures are carried out using monocular (single-view) endoscopes, and thus many methods have established monocular approaches to 3D reconstruction that seamlessly integrate with existing treatment planning and surgical workflows (Maier-Hein et al., 2013). Burschka et al. (2005) proposed a SLAM approach for sinus surgery that obtained a sparse surface reconstruction entirely from monocular endoscopic video. Reconstruction scale was obtained via rigid alignment to a CT scan. Other sparse SLAM approaches for monocular endoscopy include the work of Grasa et al. (2011, 2013), which leveraged extended Kalman filters to improve reconstruction accuracy for handheld endoscopic video capture; the work of Marcinczak and Grigat (2014), which adopted a photometric, volumetric approach (Newcombe et al., 2011) that accounts for surface specularities in its photometric cost; and the work of Chen et al. (2018), which applied intraoperative meshing to the reconstructed point cloud with a goal of real-time 3D visualization. Marmol et al. (2018) introduced a keypoint-based SLAM approach for anthroscopy in minimally invasive surgery scenarios; this approach was later extended to perform dense PatchMatch-based MVS (Bleyer et al., 2011) on SLAM keyframes to form a dense global reconstruction (Marmol et al., 2019). Mahmoud et al. (2019) also recently introduced a monocular SLAM system with dense multi-view depth estimation for selected keyframes.

Among non-SLAM methods, Koppel et al. (2007) introduced an approach combining SfM and MVS estimation with specific applications for colonoscopy. Hu et al. (2012) explored SfM-based video reconstruction for stereo and monocular endoscopes alike, with careful consideration for missing and outlier data. Collins et al. (2014) performed SfM reconstruction for a small number of endoscopic images of the uterus with manual partial labeling of the organ of interest. A preoperative

scan was then aligned to this sparse result to compute a 2D/3D registration. Several approaches have proposed to obtain a reconstruction solely from SfM and then perform single-frame or global surface reconstruction from the resulting point cloud Thormahlen et al. (2002); Sun et al. (2013); Lurie et al. (2017). However, these methods are strongly dependent on the completeness and accuracy of the SfM result. Considering newer technologies for external 6-DoF tracking of the endoscopic device, Garbey et al. (2018) assessed sparse surface reconstruction for laparoscopic scenarios where the absolute camera pose is known.

2.2.2 Shading-based Approaches and Reflectance Estimation Methods

As mentioned above, single-image shape-from-shading approaches have long been applied to endoscopic imagery. For example, Tankus et al. (2005) demonstrated some of the first SfS results on medical images following the introduction of the perspective PDE formulation for SfS. Visentini-Scarzanella et al. (2012) applied Lambertian SfS on endoscopic images with a non-co-located light source and proposed an approach for scale recovery by triangulating surface specularities. Wu et al. (2010) introduced a multi-view surface reconstruction approach leveraging Lambertian shading and known camera motion in the context of bone reconstruction. Their approach first performs single-view SfS on individual images, aligns these individual surfaces, and progressively introduces multi-view surface consistency constraints to refine and fix the estimated SfS depthmaps. I further discuss multi-view extensions of shading-based surface estimation in a later subsection.

Several works have investigated surface reflectance estimation and illumination modeling for enhanced surface visualization. An early method by Kitoh et al. (1997) investigated color correction in endoscopic video while explicitly accounting for interreflections of the light off of the surface, an assumption that is often ignored in shading-based approaches. Perhaps most pertinent to my work is the method of Chung et al. (2004, 2006), which uses a 2D/3D rigid registration algorithm (Deligianni et al., 2004) to align a CT scan with bronchoscopic video. Given the aligned CT surface, they estimate a cubic 1D BRDF and a cubic light attenuation function (decrease in brightness based on depth) for the video. This illumination and reflectance model then used to more realistically

render the CT surface from novel views. Nunes et al. (2017) similarly estimate the BRDF of a liver using a video-aligned CT scan after manual non-rigid 2D/3D alignment.

Finally, one quite different single-frame depth estimation method that is worth mentioning is the approach of Hong et al. (2009, 2014), which is specifically tailored for the 3D reconstruction of colonoscopic images. This approach explicitly models the “tube with folds” anatomy the colon and uses reasoning about light intensity to compute slant directions. However, the approach does not generalize to other anatomical structures.

2.2.3 Stereo Endoscopy and Other Semi-Controlled Capture Scenarios

Many methods have focused on reconstruction using binocular stereo endoscopes, which recover per-frame depth using photometric matching between a synchronized pair of cameras; this synchronized matching leads to a much more controlled reconstruction problem Moutney et al. (2010). Stereo approaches have frequently been combined with SfM- or SLAM-type approaches for complete surface reconstruction, and 3D stereoscopic endoscopy has recently shown potential for improving treatment outcomes versus traditional monocular endoscopy (Albrecht et al., 2016; Egi et al., 2016; Best, 2019; Bickerton et al., 2019). Considering multi-view reconstruction approaches, one early SfM-type approach from Kitoh et al. (1998) proposed to use a stereo endoscope for accurate scale estimation. Later efforts include the work of Lau et al. (2004) that proposed a method using stereo endoscope observations to monitor cardiac deformations caused by heartbeats and respiration. Moutney et al. (2006) introduced the first (sparse) stereoendoscopic SLAM approach for minimally invasive surgery; further work introduced coarse surface stitching from the sparse reconstruction (Moutney and Yang, 2009) and an altered SLAM approach to compensate for the periodic motion caused by respiration (Moutney and Yang, 2010). A number of other stereo SLAM approaches have been introduced since this initial work, targeting areas such as robust tracking in rigid Chang et al. (2014) and deforming (Lin et al., 2013) environments.

Surface reconstruction via stereo depthmap fusion has also been explored. For example, Reichard et al. (2015) used stereo endoscopy with organ segmentation and depthmap fusion to achieve

complete surface reconstruction, which can be used in a stereoscopic SLAM system (Reichard et al., 2016). Motivated towards real-time AR surgery applications, Chen et al. (2017) also proposed a stereo SLAM approach with depthmap fusion. Recently, Song et al. (2018) introduced a real-time stereo SLAM approach that is able to handle deforming surfaces and demonstrated aligned depthmaps for a number of endoscopic video sequences. Apart for multi-view reconstruction approaches, a number of works have explored improving, post-processing, and evaluating stereo depthmap estimation algorithms for endoscope-specific applications, particularly with a goal of overcoming the inherent difficulties of stereo reconstruction for low-texture surfaces (Röhl et al., 2011; Stoyanov et al., 2010; Chang et al., 2013; Parchami and Mariottini, 2014; Totz et al., 2014; Wang et al., 2018; Zampokas et al., 2018).

Active techniques are a promising alternative for obtaining per-frame depth information in endoscopy without a need for well-textured surfaces, although such approaches require substantial changes to the endoscopic hardware Maier-Hein et al. (2014). For example, Penne et al. (2009) introduced the first time-of-flight endoscope, which is able to reconstruct the depth for a given endoscopic frame based on detected phase shifts in projected infrared light. Haase et al. (2013) also proposed a time-of-flight endoscope prototype. Parot et al. (2013) introduced a modified endoscope design that enables photometric stereo endoscopy by switching between different illumination configurations. This multi-illumination approach provides controlled constraints for surface reconstruction under assumptions of Lambertian reflectance. Edgcumbe et al. (2015) introduced a structured light approach wherein a small checkboard light projector is inserted into the body; the underlying surface can then be recovered using a stereo endoscope or by tracking the pose of the projector relative to the camera. Several other methods using projected light have been proposed for improved tracking, surface reconstruction, and registration to CT (Jin et al., 2007; Qiu and Ren, 2018). Visentini-Scarzanella et al. (2015) proposed an endoscopic system that leverages both structured light projection and photometric stereo. This approach recovers semi-dense surfaces, in part by leveraging a Blinn-Phong reflectance model for the photometric stereo estimation. See

Lin et al. (2016) and Bernhardt et al. (2017) for further discussion on active and stereo endoscopic reconstruction methods.

2.2.4 Combined Sparse or Dense Reconstruction with Shading Estimation

A number of works have explored the combination of shading models with multi-view reconstruction methods. Kaufman and Wang (2008) proposed to use SfM to obtain camera motion and Lambertian SfS to obtain per-frame depth; however, the authors reported that the success of depthmap fusion in their method was hindered by inaccuracies in the SfS estimates. Tokgozoglu et al. (2012) 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. Turan et al. (2017) achieved non-rigid SLAM by using SfS to estimate per-frame depth combined with inter-frame point tracking and depthmap-to-fused-model surface registration. These authors later introduced a camera tracking method that performs per-frame depth estimation using Lambertian SfS and then feeds the resulting RGB-D image into a recurrent neural network to regress 6-DoF camera motion (Turan et al., 2018). Several other works have explored shading-based alignment of pre-operative 3D scans to endoscopic imagery. I discuss these approaches in the following subsection.

2.2.5 Template-based Reconstruction and Alignment of Pre-operative 3D Scans

For monocular reconstruction of deforming environments, several efforts have been made to extend the Shape-from-Template problem (Bartoli et al., 2015) to utilize shading information. 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 (Malti et al., 2011, 2012; Malti and Bartoli, 2014). 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. Earlier rigid reconstruction methods in this vein include the work of Shoji et al. (2001), who aligned a pre-operative CT scan to video in a

two-stage process of texture-based alignment followed by shading-based refinement; the CT texture is obtained from the previous video frame (assuming an initial alignment), and shading is refined by rendering the CT surface and minimizing the overall squared intensity difference. Later work expanded this approach to include image-based tracking with alignment to the CT scan (Mori et al., 2002). Helferty and Higgins (2002) used a preoperative CT scan as a geometry proxy for camera tracking in bronchoscopy under rigid surface assumptions. Assuming an initial alignment of the CT to the video is available, this method estimates the relative camera motion between frames via an optical flow (*i.e.*, intensity matching) formulation that utilizes the 2D motion constraints induced by project of the CT surface. The method was later extended to use an alignment procedure assuming Lambertian shading (Helferty et al., 2007) and was used to perform image-based texturing of the CT mesh (Rai and Higgins, 2006).

Rigid registration was also used by Vagvolgyi et al. (2008) to align single-frame stereo endoscope depth estimates to a CT mesh. Mirota et al. (2009) proposed to use a trimmed iterative closest point approach to rigidly align a preoperative CT scan to a 3D point cloud created using an SfM-type type approach for endoscopic video (Wang et al., 2008). Bernhardt et al. (2015) performed rigid 3-DoF camera-to-CT-surface alignment assuming local Lambertian shading and albedo in the image; this approach is in contrast to other shading-based alignment approaches that assume reflectance properties hold globally across the image. Billings and Taylor (2015) introduced an iterative alignment procedure to rigidly align two oriented point clouds. Sinha et al. (2018) extended this work to deformably register a surface representation of the nasal cavity and sinuses to a meshed SfM point cloud recovered from endoscopic video; unlike the methods mentioned above, this approach uses a shape space learned from extracted CT surfaces and does not require a patient-specific CT scan.

Among other methods that employ non-rigid modeling, Deligianni et al. (2004, 2006) introduced a method that constructs an active shape model from multiple CT scans of a patient and then applies a deformable registration of this model to 2D endoscopy. For the CT-to-video registration a linear SfS is first applied to obtain a surface normal map for the image, to which the shape model is

subsequently adjusted to match. Others have proposed methods for deformable registration of a CT preoperative scan to sparse stereo-based reconstructions (Haouchine et al., 2014), partial surfaces Song et al. (2016), and time-of-flight endoscopic imagery (dos Santos et al., 2014).

2.2.6 Monocular Depth Estimation via Convolutional Neural Networks

The recent explosion of convolutional neural networks for image processing has encouraged interesting alternatives to classical approaches for 3D reconstruction, including for endoscopies. Reiter et al. (2016) proposed an interesting approach to train a patienti-specific neural network that regresses per-pixel depth and normal information solely as a function of position in the image and pixel color. This technique bypasses explicit surface reflectance and illumination modeling, which avoids the common pitfalls of pure shading-based modeling. However, the approach requires careful 3D alignment of the endoscopic video frames to a preoperative CT scan for each specific patient in order to train against a ground-truth surface, and it is unclear how well the method would generalize between patients or to environments where direct CT registration is impossible. Mahmood *et al.* (Mahmood and Durr, 2018; Mahmood et al., 2018) perform direct monocular depth estimation using a convolutional neural network with refinement via a conditional random field. A similar approach was taken by Visentini-Scarzanella et al. (2017), who learn to regress depth from virtual CT images and, to apply this network to real imagery, train a separate network to re-render real images to look like virtual CT images. Training this second network again requires 2D/3D registration of a CT surface to its corresponding real endoscopic video sequence.

Looking forward, network-based depth estimation approaches for endoscopy will likely benefit by incorporating temporal constraints across video sequences. The recent method of Wang et al. (2019) is one such approach that could be utilized in this vein. This approach leverages recurrent network layers to predict *both* frame-to-frame camera motion and per-pixel depth. The neural network essentially “remembers” the image properties from the previous frame and uses these to infer inter-frame parallax, which is a much stronger depth cue than single-frame image appearance, alone. Interestingly, however, the network’s recurrent design allows it to also perform single-frame

depth estimation. Having such a fallback is quite useful when processing endoscopic video. This is because an endoscopic video may only contain short snippets of “good” imagery due to, for example, constant patient motion. Since some parts of the anatomy may therefore only be glimpsed very briefly, it may be necessary to drop temporal constraints for these frames in order to reconstruct them.

2.3 Modeling Transient Objects in Crowd-sourced Imagery

There has been a strong interest in automatically obtaining 3D reconstructions from crowd-sourced images. The seminal work of Snavely *et al.* (Snavely et al., 2006, 2008) demonstrated the feasibility of reconstruction from Internet photos, and later systems robustified the reconstruction methods and tackled increasingly larger scenes and photo-collections. Today, state-of-the-art systems are able to provide highly detailed 3D models of thousands of sites around the world from one-hundred million user-uploaded images (Heinly et al., 2015; Schönberger et al., 2016). However, the resulting models are only reconstructed up to an unknown scale factor and only represent the static parts of the scenes. Transient objects such as humans are inherently missing in such reconstructions.

A number of works have leveraged human detections for single-view camera calibration, particularly for surveillance cameras, and for crowd modeling in synchronized multi-view systems. Lv *et al.* (Lv et al., 2002, 2006) and others (Krahnstoeber and Mendonca, 2005; Junejo and Foroosh, 2006; Kusakunniran et al., 2009; Micusik and Pajdla, 2010) extract head and foot positions for one or more walking humans in each frame of a video taken by a single stationary camera. Under the assumption that people stand upright and that the walking area is flat, these methods recover the vertical vanishing point and a horizon line for the scene, which can be further used to obtain camera intrinsics and the ground plane relative to the camera. If the height of one or more of the detected people is known, the absolute height of the camera above the ground can also be recovered. Notably, Liu *et al.* (Liu et al., 2011) used known human height distributions to automatically determine focal length and camera height. Other works (Hödlmoser et al., 2011; Trocoli and Oliveira, 2016)

explored increasing robustness by additionally incorporating vanishing points from the static scene. For general crowd modeling in multi-view synchronized systems (Wang, 2013), a large number of methods (*e.g.* (Ge and Collins, 2010; Fleuret et al., 2008; Otsuka and Mukawa, 2004; Focken and Stiefelhagen, 2002; Black et al., 2002)) exist to triangulate and track people in the camera space, potentially without explicit correspondences (Liu et al., 2013) or a knowledge of the system calibration (Guan et al., 2016). All of these works, however, either assume that the temporal domain is densely sampled or only perform a calibration task for a single camera. Multi-view reconstructions from internet photo-collections, in contrast, consist of potentially tens of thousands of unique, temporally disjoint images.

Among other methods for reconstructing moving humans, trajectory triangulation for dynamic objects has been well-researched for images with dense temporal sampling (Avidan and Shashua, 2000; Park et al., 2010; Zheng et al., 2015; Ji et al., 2014), but the topic has rarely been applied to unordered photo collections (Zheng et al., 2014) and has not been applied in cases where hundreds or thousands of object class instances are observed. Garg *et al.* (Garg et al., 2011) explored detecting a single, manually specified individual among sets of Internet imagery, working under the assumption that the individual is positioned in approximately the same location across many images. Martin-Brualla *et al.* (Martin-Brualla et al., 2014) pieced together separate crowd-sourced 3D reconstructions by, in part, recovering the paths of photographers moving between them; this method does not recover the behavior of non-photographers, however. Zheng *et al.* (Zheng et al., 2014) tackled the lack of temporal overlap by leveraging single-instance detections to localize object class trajectories. Their insight was that most object classes have structured motion paths in the scene, and recovering this path structure is complementary to recovering the object trajectories. The problem is formulated as a generalized minimum spanning tree (GMST), followed by a continuous optimization to refine the trajectory. However, the approach does not generalize to unstructured or weakly structured object class motions, as is often encountered in open scenes such as plazas or tourist sites. Additionally, the method carries high computational cost due to solving the NP-hard problem of computing the GMST (Myung et al., 1995).

Finally, Bulbul and Dahyot (Bulbul and Dahyot, 2016) introduced a method for obtaining representations of transient objects in map representations such as OpenStreetMap (OSM). In contrast to large-scale 3D reconstructions from unordered Internet photo-collections, OSM provides both a to-scale, geo-localized environment model and a coarse ground surface representation. The authors used social media photos with geo-localization metadata to place human avatars into the map. To obtain the camera position of a social media image, they registered the image to nearby Google StreetView (GSV) images based on its known geo-location. People in the images were placed onto the map's ground surface at a distance from the camera estimated by the size of their face in the image. For visualization, realistic poses and configurations for virtual people were introduced, and the authors also simulated crowd flow for the agents to move from different photograph locations within the OSM environment. While this approach places humans into 3D environments, the method relies on a large amount of data (scene scale, OSM models, GPS data, social media timestamps, and GSV imagery) that is typically unavailable for general large-scale 3D reconstructions.

2.4 Crowd Simulation in Virtual Representations of Real Environments

Integrating virtual agents into real imagery has a rich history in computer graphics, with computer-generated special effects (Thalmann and Thalmann, 1997) and interactive systems (Maes et al., 1995) dating back to the 1990s. A number of works have extended these ideas by combining object tracking/modeling with simulation, with a goal of creating augmented video wherein virtual agents interact with real objects in a convincing manner (Baiget et al., 2009; Fernández et al., 2011; Doğan et al., 2018). Even more works have employed computer vision to automatically learn crowd motion behavior (Lerner et al., 2007; Musse et al., 2007; Courty and Corpetti, 2007; Alahi et al., 2016; Gupta et al., 2018). To my knowledge, however, the only work that has attempted crowd simulation using large-scale photo-collections is the aforementioned approach of Bulbul and Dahyot (Bulbul and Dahyot, 2016), who render people detected in social media photos as virtual agents moving within OpenStreetMap models. Integrating crowd simulation directly into large-scale 3D reconstructions has, as yet, not been demonstrated.

In addition, crowd simulation – and overall completeness of the virtual environment – requires well-defined ground surfaces in the scene. Several works have joined reconstructions from ground-level imagery with reconstructions from aerial imagery to derive a more complete scene model (Frueh and Zakhor, 2003; Shan et al., 2013). For ground-level imagery, an effective approach has been to leverage semantic constraints in volumetric reconstruction (among several works, see for example Häne et al., 2013, 2016; Cherabier et al., 2018). Given a voxelization of the scene, these approaches aggregate distance fields from a set of depthmaps considering semantic labels for each image. A separate distance field is aggregated for each semantic class. This multi-label volume is then refined using learned priors on class transitions; for example, a ground-labeled voxel is likely to be surrounded by other ground voxels but unlikely to exist above an empty voxel. Through a variational formulation, per-voxel label probabilities are optimized to respect these shape priors while accounting for the surfaces observed by the input imagery and depthmaps. Final per-class surfaces are extracted using the most probable voxel labels. I do not rely on training data or semantic labeling for my method, although the option is an intriguing direction for future work.

CHAPTER 3: 3D RECONSTRUCTION OF ENDOSCOPIC VIDEO

Endoscopy is a common medical procedure wherein a camera with a light attached is inserted into the body, allowing physicians to gain a direct view of the internal surfaces of a patient without resorting to strongly invasive methods. Endoscopic medical vision applications constitute a steadily growing field of 3D computer vision research with much potential to improve patient outcomes without significant alterations to existing physician workflows. For example, a doctor performing laparoscopic surgery uses video to as a navigational aid during the procedure. By performing online 3D reconstruction on this video as it is captured, medical vision technologies can augment the surgeon’s spatial reasoning during the procedure.

Offline 3D reconstruction is also a potentially invaluable tool for treatment planning and review. To provide a driving scenario, consider nasopharyngoscopy, *i.e.*, endoscopic video of the upper throat. Cancerous tumors in the throat are often superficial, perhaps less than 2mm in thickness. However, treatment planning workflows typically rely on computed tomography (CT) scans that usually have a resolution of 3mm. To localize a tumor in the CT scan, a treating physician will perform a nasopharyngoscopy on their patient to obtain a visual confirmation of the tumor’s location. They will then manually label the tumor on the throat surface in the CT, often from memory and without clear geometric cues, the latter of which is due to the low CT resolution. If, instead, a textured 3D surface representation – an *endoscopogram* – was reconstructed from the nasopharyngoscopic video, the physician would be able to use the color data to more easily and, importantly, more accurately label the tumor; the labeling could then be transferred to the CT surface via deformable registration of the two surfaces. The endoscopogram also serves as a convenient mechanism for endoscopy review, as it condenses minutes of video into a single unified surface. This is especially important for procedures like colonoscopy, where the overall procedure

can exceed 40 minutes in length, which prohibits review to determine whether growths were missed or potentially unobserved.

Constructing an endoscopogram, however, is a challenging task for several reasons, especially in the case of nasopharyngoscopy. For one, the inside surfaces of the throat are stationary for only very short time windows due to the patient breathing and swallowing. This makes traditional SfM techniques difficult to leverage, since the relative camera motion is quite slight relative to surface; from experience, sparse point triangulation is generally possible but quite prone to noise. Compounding this, the throat surfaces are quite homogeneous in appearance, which limits the amount of feature points that can be reliably matched between video frames; this is unfavorable for NRSfM approaches, which typically rely on *ad hoc* point clusterings to build a dynamic motion model. The homogeneous textures and poor triangulation angles also make MVS depth estimates quite noisy, resulting in degraded surface estimates (Fig. 3.1).

Given these difficulties for multi-image reconstruction methods, single-image methods like SfS seem to be a reliable alternative, since they are agnostic to texture homogeneity, camera motion, and surface dynamics. Unfortunately, the near-surface lighting conditions in nasopharyngoscopy make it impossible to derive a global model of illumination/material reflectance that can be applied for all video frames. A successful SfS approach must be able to refine illumination properties on a frame-by-frame basis.

To bootstrap endoscopogram construction, I introduce a new *SfM-guided* SfS method to recover a dense surface representation for individual frames of an endoscopic video. The insight here is that the sparse geometry obtained from SfM, while being far from perfect, provides a sufficient geometric prior to guide a refinement of the material reflectance model and an overall SfS procedure. The proposed method, shape-from-motion-and-shading (SfMS), alternates between three stages: First, the current SfS depth map (starting from some initialization) is warped to a set of 3D SfM points for the given image. Next, this warped surface is used to update a surface illumination/reflectance model. Finally, this reflectance model is used within a *regularized SfS* framework to obtain a new depth map for the image. The regularized SfS approach is a new derivation that strikes a balance

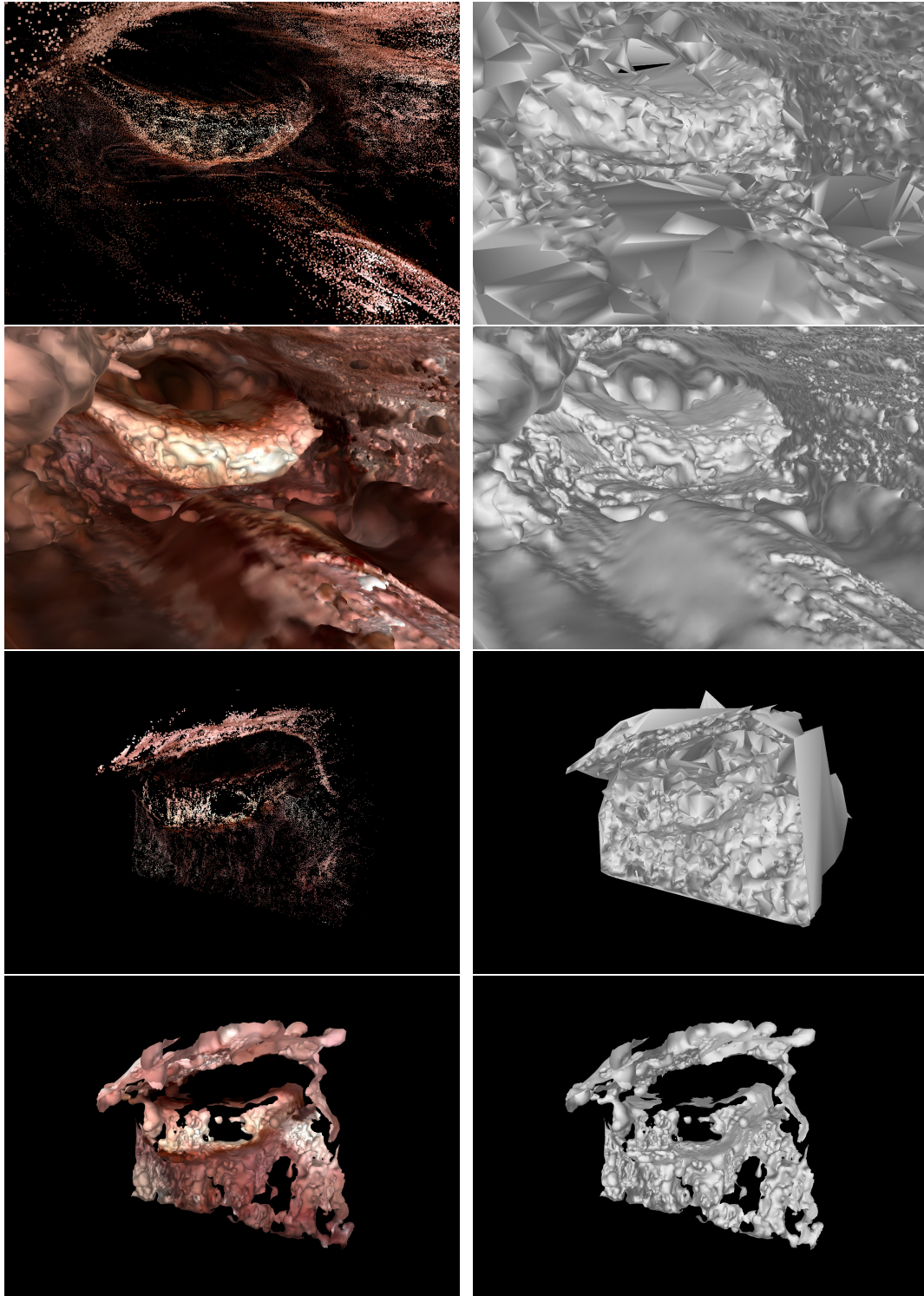


Figure 3.1: Surface estimates using multi-view stereo reconstruction tend to be noisy and/or incomplete for endoscopic data. Top row: Fused point cloud obtained via MVS (Schönberger et al., 2016) and an untextured surface reconstructed from this point cloud using a method based on the Delaunay tetrahedralization approach of Labatut et al. (2009). Second row: Textured and untextured views of the surface obtained using Poisson surface reconstruction (Kazhdan and Hoppe, 2013) on the MVS point cloud. Bottom two rows: Same results for a different patient.

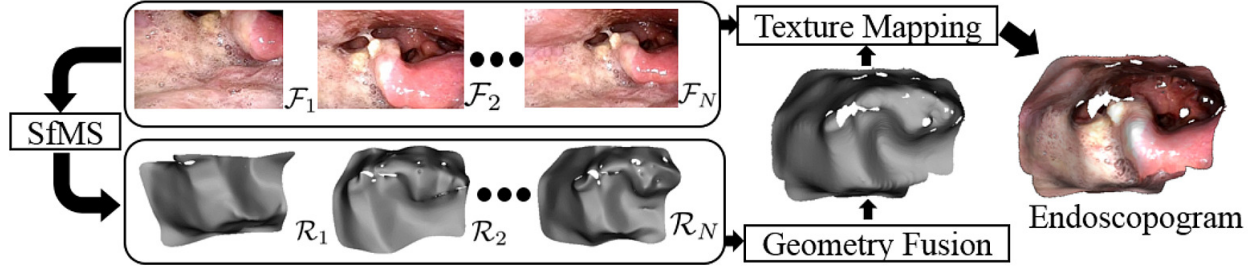


Figure 3.2: An endoscopogram is constructed via non-rigid registration of multiple SfMS reconstructions of individual video frames.

between the previous SfS surface and the shading constraints that arise under the newly estimated illumination/reflectance model. I also propose a new, generalized reflectance model that better accounts for the illumination conditions present in endoscopy, compared to the Lambertian model typically adopted for SfS. Moreover, I propose a simple approach for modeling light interreflections within the endoscopic space that are not accounted for in traditional SfS approaches, and I show that this can substantially improve the accuracy of the depth estimated by SfMS.

Fig. 3.2 outlines the overall process of constructing the final endoscopogram. Once depthmaps have been computed using SfMS for a set of endoscopic video frames, a final endoscopogram can be formed via non-rigid registration of the individual surfaces. This more complete surface can then itself be non-rigidly aligned to the CT surface for visualization within the original treatment planning space. Details of these fusion and registration procedures are described in Zhao et al. (2015), Zhao et al. (2016), and Zhao (2017).

3.1 Background

3.1.1 Reflectance Models

The amount of light reflecting off 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_{λ_r} to the irradiance of the light hitting the surface E_{λ_r} (Cook and Torrance, 1982). The behavior of the BRDF is specific to the material of the surface. 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,

$$\text{BRDF}_\lambda(\theta_i, \phi_i, \theta_r, \phi_r) = \frac{I_{\lambda r}}{E_{\lambda i}}, \quad (3.1)$$

where λ represents light wavelength. In the following, the wavelength dependence of the BRDF is implicitly assumed.

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

$$E_i = I_i \frac{A}{r^2} \cos \theta_i, \quad (3.2)$$

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

For the case of endoscopy, two simplifying assumptions about the BRDF can be made that help the overall modeling of the problem. The first assumption is that the BRDF exhibits surface isotropy, which constrains it to only depend on the relative azimuth, $\Delta\phi = |\phi_i - \phi_r|$, rather than the angles, themselves (Koenderink et al., 1996). While this sacrifices some generality, it provides a good approximation for surfaces with low anisotropy. Second, it is assumed 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

$$I_r(r, \theta_i) = I_i \frac{A}{r^2} \cos(\theta_i) \text{BRDF}(\theta_i). \quad (3.3)$$

3.1.2 Surface Model for Shape-from-Shading

Let $(x, y) \in \Omega$ represent image coordinates after normalization by the intrinsic camera parameters (accounting for lens distortion, centering around the principal point, and dividing by the focal length). For a given camera pose, the surface function $f : \Omega \rightarrow \mathbb{R}^3$ maps points in the image plane

to 3D locations on a surface viewed by the camera. Under perspective projection,

$$f(x, y) = z(x, y) \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}, \quad (3.4)$$

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

$$r(x, y) = \|f(x, y)\| = z(x, y)\sqrt{x^2 + y^2 + 1}, \quad (3.5)$$

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

$$\mathbf{n}(x, y) = f_x \times f_y = z \begin{pmatrix} -z_x \\ -z_y \\ xz_x + yz_y + z \end{pmatrix}. \quad (3.6)$$

The lighting conditions of the endoscope allow us to assume that the scene is illuminated by a single light source located at the optical center of the camera. In this case, 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 $\theta_i(x, y)$ between the normal and light direction vectors is then equal to their dot product:

$$\cos \theta_i = \hat{\mathbf{n}} \cdot \hat{\mathbf{l}} = \frac{z}{\sqrt{(x^2 + y^2 + 1)(z_x^2 + z_y^2 + (xz_x + yz_y + z)^2)}}, \quad (3.7)$$

where “carat” represents normalization to unit length and the dependence of all variables on (x, y) is implied.

Prados and Faugeras (2005) note that Eq. (3.7) can be simplified using the change of variables $v(x, y) = \ln z(x, y)$:

$$\hat{\mathbf{n}} \cdot \hat{\mathbf{l}} = \frac{1}{\sqrt{(x^2 + y^2 + 1) (v_x^2 + v_y^2 + (xv_x + yv_y + 1)^2)}}. \quad (3.8)$$

This transformation allows us to separate terms involving v from those involving its derivatives in our shading model, which is important for PDE formulations of the SfS model.

3.1.3 Structure-from-Motion

As mentioned previously, Structure-from-Motion (SfM) (Hartley and Zisserman, 2003; Pollefeys et al., 2004; Schönberger and Frahm, 2016) 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 and scene structure. The scene structure is parameterized by a set of 3D points projecting to corresponding 2D image features.

In the case of endoscopy, the motivation for using SfM is that it provides a (sparse) prior on depth, which supplies adequate constraints for surface geometry and reflectance estimation. Because SfM uses rich feature descriptors to identify image correspondences, compared to the weaker photo-consistency metrics of multi-view approaches, experience shows that SfM produces substantially more reliable, albeit sparse and typically noisy, geometry for endoscopic datasets. Fig. 3.3 shows an example SfM reconstruction of endoscopic data using several segments from the overall video.

One limitation to the generality of the method is that sparse non-rigid reconstruction in medical settings is an unsolved problem (Stoyanov, 2012; Münzer et al., 2018). However, the proposed approach can handle any sparse data as input, and thus the method could easily be integrated with

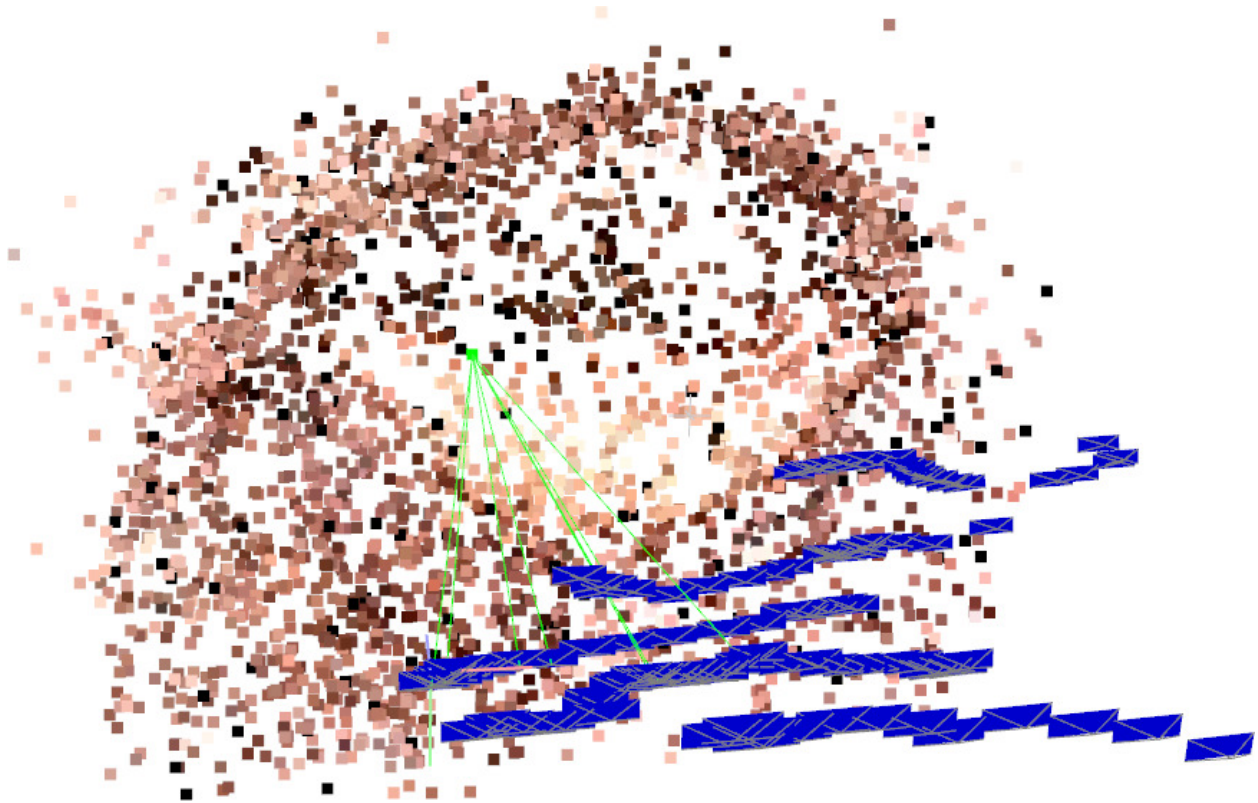


Figure 3.3: Structure-from-Motion results for endoscopic video. Individual 3D surface points (colored dots) and camera poses (blue) are jointly recovered.

non-rigid SfM in future work. In the experiments on live endoscopy, rigid SfM is employed on small intervals of temporally neighboring frames with minimal surface deformation. When slight scene motion does occur in these images, SfM has proven to be fairly robust against distortion of the resulting sparse geometry. While this justifies the use of the approach for scenes with small deformation, the method could benefit from the development of robust sparse methods non-rigid modeling that work in difficult endoscopic scenarios, if they were able to provide more accurate sparse point triangulations.

3.2 Method

The ultimate goal of the proposed method is to produce a dense, geometrically accurate surface for a given image in a video sequence. My approach achieves this using a new Shape-from-Shading formulation that utilizes the sparse 3D point data obtained via Structure-from-Motion. In this section,

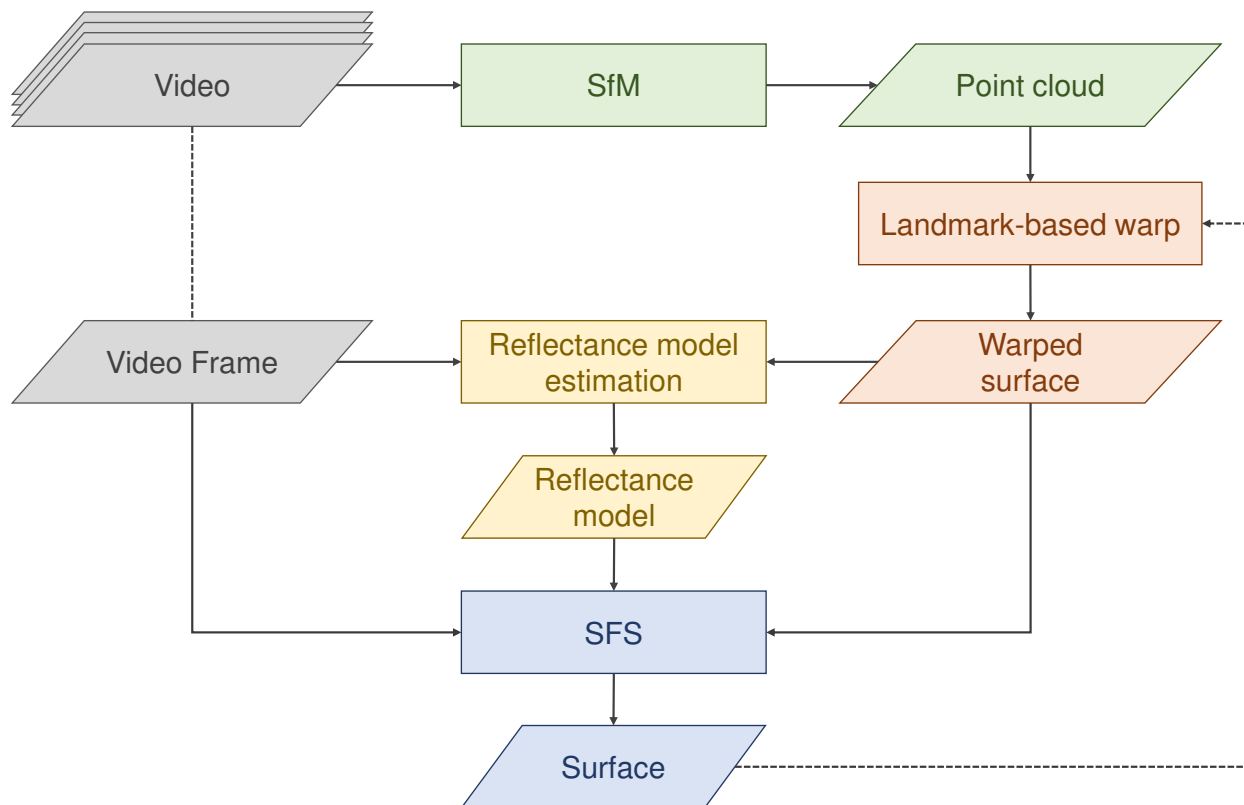


Figure 3.4: Diagram of the proposed iterative approach for dense surface reconstruction of a single video frame.

I detail the main contributions of the current work that enable this enhanced depth estimation: First, I introduce a regularized formulation of SfS that allows for trade-off between predicted image intensity and similarity to an existing estimated surface. This formulation is integrated into a Lax-Friedrichs (LF) (Kao et al., 2004; Ahmed and Farag, 2006) partial differential equation (PDE) solver. To improve the accuracy of the solution, I suggest a way to account for errors along occlusion boundaries in the image using intensity-weighted finite differences, and I also outline how parameters for the LF solver can be computed for general 1D reflectance models. Second, I propose a novel reflectance model for use in SfS that can more finely capture real-world illumination conditions. Finally, I develop an iterative update scheme (see Fig. 3.4) 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.

3.2.1 Initial PDE

Eq. (3.3) above models observed intensity $I_r(r(x, y), \theta_i(x, y))$ for a generic, isotropic BRDF with the assumption that the light source is colocated with the camera. In practice, the values of I_r are obtained directly from the input (grayscale) image G , *i.e.*, $I_r(r(x, y), \theta_i(x, y)) = G(x, y)$. Joining Eq. (3.3) with Eqs. (3.5) and (3.8) and multiplying by r^2 , we have

$$(x^2 + y^2 + 1)Ge^{2v} - I_i A \cos(\theta_i) \text{BRDF}(\theta_i) = 0 \quad (3.9)$$

(note $e^{2v} = z^2$). The dependence of G , v , and θ_i on (x, y) is implied. The ultimate goal of the following formulation is to solve for log-depth v (and from this, to derive the depth z) at each point in the image.

To simplify the notation, denote $L(x, y) = (x^2 + y^2 + 1)G(x, y)$. Also, at each point (x, y) , let $\eta(v_x, v_y) = I_i A \cos(\theta_i) \text{BRDF}(\theta_i)$. (Recall that θ_i can itself be expressed as a function of v_x and v_y , according to Eq. (3.8).) Using these substitutions and adopting appropriate boundary conditions to handle the image domain, we can write Eq. (3.9) as a static PDE of v and its derivatives:

$$\begin{cases} Le^{2v} - \eta(v_x, v_y) = 0, & (x, y) \in \Omega \\ v(x, y) = \psi(x, y), & (x, y) \in \partial\Omega, \end{cases} \quad (3.10)$$

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

3.2.2 Regularization

The PDE introduced above is dependent on the accuracy of the BRDF modeling the scene. To prevent inaccuracies arising from errors in the BRDF fit, I propose to use the 3D points obtained from SfM as an additional set of constraints for the estimated log-depths, v . Naïvely, the attempt could be made to directly add these known depths as point constraints – *i.e.* for a given 2D

feature point (x_k, y_k) with estimated depth z_k , we would require $v(x_k, y_k) = \ln z_k$. However, such constraints are ineffective in the PDE formulation, as they have no effect on the solution outside of that 2D location (Horowitz and Kiryati, 2004). The 3D point cloud acquired via SfM can also potentially yield noisy or outlier depth measurements, especially for scenes where the camera motion is small, which results in larger depth uncertainty for 3D triangulation. Moreover, even minor surface deformations can further degrade triangulation accuracy in live endoscopy. Thus, it is inadvisable to fix the depths estimated by SfM to exact values.

Instead, assume there exists a current estimate $f_{\text{est}}(x, y)$ of the surface viewed by the camera. In the iterative scheme introduced below, $f_{\text{est}}(x, y)$ is a warped surface that passes near the 3D SfM points. A simple regularization is added to the SfS PDE (Eq. (3.10)) that constrains the solution to be similar to the estimated surface in high-confidence regions (*i.e.* regions where the warped surface agrees with the SfM feature points). This is captured in the following energy function:

$$E(v) = E_0(v) + \int_{\Omega} \frac{\lambda}{2} (e^v - z_{\text{est}})^2 d\mathbf{x}. \quad (3.11)$$

The term $E_0(v)$ denotes an energy functional effecting the original SfS PDE, *i.e.*, $\frac{\partial E_0}{\partial v} = Le^{2v} - \eta(v_x, v_y)$. The function $z_{\text{est}}(x, y)$ is the (fixed) depth of the existing surface at a given image coordinate, and the parameter $\lambda(x, y) \geq 0$ controls the influence of the regularization term. An approach for calculating $\lambda(x, y)$ is defined below, when the final iterative algorithm is introduced. The squared loss term is a design choice, of course; in principle, robust choices such as the absolute difference could be adopted, to help alleviate gross errors in the current estimated surface.

The minimum of $E(v)$ is a new PDE:

$$\frac{\partial E}{\partial v} = Le^{2v} - \eta(v_x, v_y) + \lambda(e^v - z_{\text{est}}) e^v \stackrel{!}{=} 0. \quad (3.12)$$

The associated PDE with boundary conditions can be written as

$$\begin{cases} (L + \lambda)e^{2v} - \lambda z_{\text{est}}e^v - \eta(v_x, v_y) = 0 & (x, y) \in \Omega \\ v(x, y) = \psi(x, y). & (x, y) \in \partial\Omega. \end{cases} \quad (3.13)$$

3.2.3 Solving the PDE

Ahmed and Farag (2006) introduced a fast-sweeping method for SfS with the Oren-Nayar reflectance model (Oren and Nayar, 1994), itself based on a method by Kao et al. (2004), that can be used to solve PDEs like the regularized equation introduced above. I adopt this solving scheme here and outline how it can be extended to any general 1D reflectance model. Their approach uses the Lax-Friedrichs (LF) Hamiltonian, which provides an artificial viscosity approximation for solving static Hamiltonian-Jacobi equations, *i.e.*, functions of the form $H(\mathbf{x}, \nabla v(\mathbf{x})) = R(\mathbf{x})$. The LF Hamiltonian is advantageous in that it is able to handle non-convex, complex Hamiltonian equations. While time-independent PDEs like Eq. (3.13) are not Hamiltonian equations due to the reliance of the variable v , Ahmed and Farag (2006) demonstrated that the LF Hamiltonian can be effectively applied to these types of equations.

3.2.3.1 Discretization

Before explaining the LF solving scheme, it is necessary to first introduce some numerics that underlie the approximation of the PDE. Let the image space be uniformly discretized into columns x_i and rows y_j with grid spacing Δx and Δy . Let $v_{i,j}$ be the log-depth at position (x_i, y_j) . Denoting $p = \frac{\partial v}{\partial x}$ and $q = \frac{\partial v}{\partial y}$, the forward- and backward-difference approximations of p can be represented as

$$p_{i,j}^+ = \frac{1}{\Delta x}(v_{i+1,j} - v_{i,j}) \quad \text{and} \quad p_{i,j}^- = \frac{1}{\Delta x}(v_{i,j} - v_{i-1,j}), \quad (3.14)$$

respectively, and similarly for q . Let

$$\bar{p}_{i,j} = \frac{p_{i,j}^+ + p_{i,j}^-}{2} \quad \text{and} \quad \bar{q} = \frac{q_{i,j}^+ + q_{i,j}^-}{2} \quad (3.15)$$

be the average of the finite differences, and let

$$\bar{v}_{i,j}^x = \frac{v_{i+1,j} + v_{i-1,j}}{2} \quad \text{and} \quad \bar{v}_{i,j}^y = \frac{v_{i,j+1} + v_{i,j-1}}{2} \quad (3.16)$$

be the average value of the grid elements adjacent to $v_{i,j}$.

3.2.3.2 Applying the Lax-Friedrichs Hamiltonian

Consider a general static Hamiltonian equation $H(x, y, p = v_x, q = v_y) = 0$. To obtain a solution of v that approximately satisfies this equation, the 2D Lax-Friedrichs Hamiltonian introduces artificial viscosity terms

$$\sigma_{i,j}^x \geq \max_{p \in [A,B], q} \left| \frac{\partial H}{\partial p}(x_i, y_j, p, q) \right| \quad \text{and} \quad \sigma_{i,j}^y \geq \max_{q \in [C,D], p} \left| \frac{\partial H}{\partial q}(x_i, y_j, p, q) \right| \quad (3.17)$$

that ensure stability of the solution scheme (Kao et al., 2004; Shu, 2007). In a global LF scheme, $[A, B]$ and $[C, D]$ cover the entire valid range of p and q , respectively, whereas in a local LF scheme, $[A, B] = [\min(p_{i,j}^+, p_{i,j}^-), \max(p_{i,j}^+, p_{i,j}^-)]$ and $[C, D] = [\min(q_{i,j}^+, q_{i,j}^-), \max(q_{i,j}^+, q_{i,j}^-)]$. See below for further discussion on these parameters. Implicitly assuming the dependence on (x_i, y_j) , the function H is approximated by the LF Hamiltonian:

$$\tilde{H}_{LF}(v_{i,j}, v_{i+1,j}, v_{i-1,j}, v_{i,j+1}, v_{i,j-1}) = H(\bar{p}_{i,j}, \bar{q}_{i,j}) + \frac{\sigma_{i,j}^x}{\Delta x} (v_{i,j} - \bar{v}_{i,j}^x) + \frac{\sigma_{i,j}^y}{\Delta y} (v_{i,j} - \bar{v}_{i,j}^y), \quad (3.18)$$

where the ‘‘bar’’ terms are from Eqs. (3.15) and (3.16), above.¹

¹This is a slightly unconventional way of expressing the LF Hamiltonian, which is typically written with artificial viscosity terms $-\frac{\sigma^x}{2}(p^+ - p^-)$ and $-\frac{\sigma^y}{2}(q^+ - q^-)$. I use this form to simplify the use of the image-weighted finite differences I introduce below.

As mentioned before, Ahmed and Farag (2006) demonstrated that this augmentation can be applied to more general equations like the original PDE in Eq. (3.13). The PDE becomes

$$(L + \lambda)e^{2v_{i,j}} - \lambda z_{\text{est}}e^{v_{i,j}} - \eta (\bar{p}_{i,j}, \bar{q}_{i,j}) + \left(\frac{\sigma_{i,j}^x}{\Delta x} + \frac{\sigma_{i,j}^y}{\Delta y} \right) v_{i,j} - \frac{\sigma_{i,j}^x}{\Delta x} \bar{v}_{i,j}^x - \frac{\sigma_{i,j}^y}{\Delta y} \bar{v}_{i,j}^y = 0, \quad (3.19)$$

plus appropriate boundary conditions that are detailed below. In a similar vein to Ahmed and Farag (2006), we can solve for the new value of $v_{i,j}$ using Newton's root-finding method, *i.e.*, expressing the left side of the above equation in a generic form of

$$g(v) = ae^{2v} - be^v + cv - d, \quad g'(v) = 2ae^{2v} - be^v + c, \quad (3.20)$$

the value of v is updated using the following equation until the solution $g(v) = 0$ is satisfied:

$$v := v - \frac{g(v)}{g'(v)}. \quad (3.21)$$

3.2.3.3 Fast Sweeping Scheme and Boundary Conditions

Kao et al. (2004) and Ahmed and Farag (2006) both outline the general algorithm for fast sweeping using the LF Hamiltonian, so I detail it on a high level, here. For initialization, the log-depth values $v_{i,j}$ are set to a large positive constant. The algorithm then proceeds to iteratively update these values to progressively closer depths, applying Eq. (3.21) to determine the new value for one $v_{i,j}$ at a time. Stable updates are maintained using diagonal "sweeps" that alternative between bottom left to top right, bottom right to top left, top left to bottom right, and top right to bottom left. For example, in the top-left-to-bottom-right sweep, the value of a general $v_{i,j}$ will be updated using values of $v_{i-1,j}$ and $v_{i,j-1}$ that have already been updated in the current sweep and values of $v_{i+1,j}$ and $v_{i,j+1}$ that have yet to be updated. Updates are applied until the total change in v over the entire image is smaller than some small positive constant.

To avoid computational catastrophe on the borders of the image, where the 4-neighborhood structure needed for \tilde{H}_{LF} is unavailable, Kao et al. (2004) propose boundary conditions to be applied on the edge of the image after every sweep. On the left border (and similarly for other three borders), their approach computes a new value of $v_{0,j}$ under two possible conditions: $p_{1,j}^+ = p_{1,j}^-$ and $p_{1,j}^+ = -p_{1,j}^-$. If the maximum of these new values is smaller than the current value of $v_{0,j}$, $v_{0,j}$ is updated to this smaller value.

In practice, I have found that taking the maximum of the two values can often lead the solution to exhibit strongly incorrect geometry near the boundary, at least for endoscopic applications (Fig. 3.5). This is due to the ground-truth surface (which is essentially a tube) near the image boundary often being very oblique w.r.t the camera’s viewing direction. Instead, taking the *minimum* of the two values seems to offer generally better results — with an additional constraint that the surface slope at the boundary is not too large. The boundary condition is thus applied on the left border using

$$v_{0,j} := \min(\max(\min(2v_{1,j} - v_{2,j}, v_{2,j}), v_{1,j} - \Delta_v^{\max})v_{0,j}), \quad (3.22)$$

where Δ_v^{\max} is the change in v corresponding to a maximum allowed incident angle (e.g., $\theta_i = 89.5^\circ$) at the image border. Similar boundary conditions are used for the other three image borders. Without the threshold on the maximum slope, sporadic artifacts can sometimes arise near the image boundaries due to specularities or dark regions (Fig. 3.5, third image). This value is somewhat sensitive – if the threshold is even 85° , I have found that the overall accuracy of the method can suffer.

3.2.4 Computing $\sigma_{i,j}^x$ and $\sigma_{i,j}^y$ for Arbitrary BRDFs

As mentioned previously, $\sigma_{i,j}^x$ and $\sigma_{i,j}^y$ (Eq. (3.17)) can be chosen using either as *global* parameters or *local* parameters. The local LF scheme is generally preferred, since it exhibits smaller dissipation in the final LF solution due to the adaptive range in which the maximum is taken (Osher and Shu, 1991; Shu, 2007). In practice, however, values for $\sigma_{i,j}^x$ and $\sigma_{i,j}^y$ may be difficult to compute

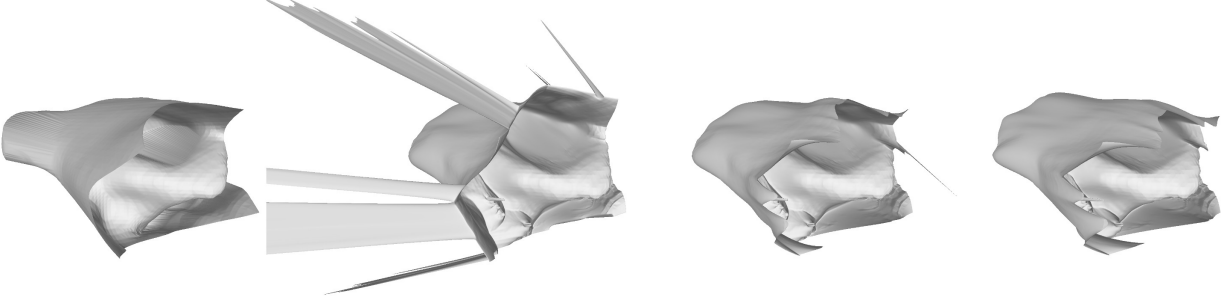


Figure 3.5: Compared to the ground-truth surface (left), the boundary conditions suggested by Kao et al. (2004) can lead to strong artifacts on the edges of the image for endoscopic applications (second from left). A minor change to these conditions can correct for this, although problematic artifacts can still occur (second from right), which can be alleviated by limiting the maximum surface slope along the image boundary (right).

at equality for both types of schemes, since the magnitude of the derivative must be evaluated over a range of p and q , and at each pixel location. An alternative to finding this exact threshold is to instead find a reasonable upper bound that is relatively simple to compute. One way to approach this for SfS is to separate the 1D BRDF and surface representation — that is, treat the PDE as a function of $\cos(\theta_i)$ (cf. Eqs. (3.9) and (3.10)), and treat $\cos(\theta_i)$ as a function of x, y, p , and q (cf. Eq. (3.8)).

To make this more clear, I next outline the computation for $\sigma_{i,j}^x$. The value $\sigma_{i,j}^y$ has a similar formulation. In the following, I use $\theta = \theta_i$ for the incident light angle to avoid confusion with (i, j) subscripts.

First, note that for the regularized SfS PDE H (Eq. 3.13), $\frac{\partial H}{\partial p} = \frac{\partial \eta}{\partial p}$, where again $p = v_x$. In Section 3.2.1, $\eta(p, q)$ was formulated as a 2D function (ignoring the dependence on x and y) to clarify the PDE formulation; however, we can equivalently express it as a 1D function, $\tilde{\eta}(\cos(\theta)) = \eta(p, q)$, with θ itself being a 2D function of p and q . Thus, considering Eq. (3.17), we

have at pixel location (x_i, y_j) that

$$\begin{aligned}
\max_{p \in [A, B], q} \left| \frac{\partial H}{\partial p} (x_i, y_j, v_{i,j}, p, q) \right| &= \max_{p \in [A, B], q} \left| \frac{\partial \eta}{\partial \cos(\theta)} \frac{\partial \cos(\theta)}{\partial p} (x_i, y_j, p, q) \right| \\
&\leq \max_{p \in [A, B], q} \left| \frac{\partial \eta}{\partial \cos(\theta)} (x_i, y_j, p, q) \right| \max_{p \in [A, B], q} \left| \frac{\partial \cos(\theta)}{\partial p} (x_i, y_j, p, q) \right| \\
&\leq \max_{\cos(\theta) \in (0, T_{i,j}^x]} \left| \frac{\partial \tilde{\eta}}{\partial \cos(\theta)} (\cos(\theta)) \right| \max_{p, q} \left| \frac{\partial \cos(\theta)}{\partial p} (x_i, y_j, p, q) \right|,
\end{aligned} \tag{3.23}$$

where $T_{i,j}^x$ is the largest possible value of $\cos(\theta)$ given $p_{i,j}^+$ and $p_{i,j}^-$, for any value of q . The lower bound of zero arises because an arbitrarily large value of q can be chosen; the upper bound of $T_{i,j}^x$ arises because the maximum value of $\cos(\theta)$ decreases monotonically with p . For the right term, I have found that a global range for $\left| \frac{\partial \cos(\theta)}{\partial p} \right|$ is more tractable to work with, so I have adopted it, here.

It turns out that both $T_{i,j}^x$ and the second maximum are computable in closed form (see Appendix A). The first maximum may or may not be easy to compute – for example, if the underlying BRDF model is assumed to be Lambertian, it is a constant, whereas other models may require a search of the entire range. For purposes of a general and efficient implementation, I use numeric differentiation to approximate $\left| \frac{\partial \eta}{\partial \cos(\theta)} \right|$ for values of $\cos(\theta)$ from 0 to 1, and I maintain a lookup table of the cumulative maximum for any value of $T_{i,j}^x$.

3.2.5 Image-weighted Finite Differences

The artificial viscosity introduced by the Lax-Friedrichs Hamiltonian can be quite dissipative (Osher and Fedkiw, 2003), meaning solution schemes involving the Hamiltonian will poorly approximate functions along discontinuities. For a surface function $f(x, y)$ (Eq. (3.4)), such discontinuities occur along self-occlusion boundaries of the surface in the image. To address this issue, I propose a simple image-intensity-based weighting scheme for $\bar{p}_{i,j}$, $\bar{q}_{i,j}$, $\bar{v}_{i,j}^x$, and $\bar{v}_{i,j}^y$ that gives higher emphasis on neighboring pixels with similar observed intensities. This approach is inspired by similar approaches in stereo-based methods (Yoon and Kweon, 2006; Gu et al., 2008).

More explicitly, consider the observed radiance $I_{i,j}$ for a pixel (x_i, y_j) normalized to the range $[0, 1]$. Define $w_{i,j}^{x\pm} = \exp\left(-\left(\frac{I_{i,j}-I_{i\pm 1,j}}{\sigma_I}\right)^2\right)$, and similarly $w_{i,j}^{y\pm}$. The parameter σ_I defines the spread of the weighting, with smaller values placing a higher penalty on intensity differences ($\sigma_I = 0.1$ for the experiments later in this chapter). The values $\bar{p}_{i,j}$ and $\bar{q}_{i,j}$ introduced above are now redefined as $\bar{p}_{i,j} = \frac{1}{w_{i,j}^{x+}+w_{i,j}^{x-}} (w_{i,j}^{x+} p_{i,j}^+ + w_{i,j}^{x-} p_{i,j}^-)$ and $\bar{q}_{i,j} = \frac{1}{w_{i,j}^{y+}+w_{i,j}^{y-}} (w_{i,j}^{y+} q_{i,j}^+ + w_{i,j}^{y-} q_{i,j}^-)$. A similar weighting is applied for $\bar{v}_{i,j}^x$ and $\bar{v}_{i,j}^y$.

3.2.6 Reflectance Model

The choice of reflectance model is key to achieving realistic SfS reconstructions. In the case of nasopharyngoscopy, the underlying surface consists of throat tissue covered by a thin layer of saliva. While throat tissue is generally Lambertian, meaning that reflected light intensity is a direct function of $\cos \theta_i$ (and thus the BRDF is a constant related to the surface albedo), the extra salivary coating induces superficial reflections that significantly alter the overall reflectivity. Since these non-diffuse effects are significantly different from those modeled by existing non-Lambertian SfS approaches (Ahmed and Farag, 2006; Vogel et al., 2009), I propose to instead model the saliva/tissue reflectance using a general basis for 1D BRDFs.

3.2.6.1 BRDF Basis

The proposed reflectance model is based on the set of BRDF basis functions introduced by Koenderink et al. (1996). 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. Assuming Helmholtz's reciprocity² and surface isotropy, the basis consists of a set of functions $S_{nm}^l(\theta_i, \theta_r, \Delta\phi_{ir})$, where $\Delta\phi_{ir} = |\phi_i - \phi_r|$, $0 \leq l \leq m \leq n \leq N$, and the quantities $(n - l)$ and

²In this context, Helmholtz's reciprocity is the principle that a 4D BRDF remains constant if the light and camera are interchanged. Of course, this is trivially true for the 1D case.

$(m - l)$ are even. N represents the order of the BRDF. The basis functions have the form

$$S_{nm}^l(\theta_i, \theta_r, \Delta\phi_{ir}) = (\Theta_n^l(\theta_i)\Theta_m^l(\theta_r) + \Theta_m^l(\theta_i)\Theta_n^l(\theta_r)) \cos l\Delta\phi_{ir}. \quad (3.24)$$

Here, $\Theta_a^b(\theta)$ is proportional to the radial function $R_a^b(\sqrt{2}\sin(\frac{\theta}{2}))$, which itself takes the form of a terminating hypergeometric series. $\Theta_a^b(\theta)$ can thus be expressed as a polynomial of $\sin(\frac{\theta}{2})$ with powers ranging from b to a :

$$\Theta_a^b(\theta) = \sum_{k=b}^a c_k \sin^k\left(\frac{\theta}{2}\right), \quad (3.25)$$

where the coefficients c_k are proportional to coefficients of the terminating hypergeometric series. (The exact value of c_k is not important for this exposition, as it will later be combined with parameters for the BRDF.) The final BRDF is a sum of the individual basis functions:

$$\text{BRDF}(\theta_i, \theta_r, \Delta\phi_{ir}) = \sum_{nml} c_{nml} S_{nm}^l(\theta_i, \theta_r, \Delta\phi_{ir}), \quad (3.26)$$

where the coefficients c_{nml} are parameters that dictate the specific BRDF.

3.2.6.2 Proposed Reflectance Model

The BRDF basis of Koenderink et al. (1996) can be adapted 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 $\Delta\phi_{ir} = 0$. Combining with Eqs. (3.24) and (3.25), this gives

$$\begin{aligned} S_{nm}^l(\theta_i) &= 2\Theta_n^l(\theta_i)\Theta_m^l(\theta_i) \\ &= 2 \left(\sum_{k=l}^n c_k \sin^k\left(\frac{\theta}{2}\right) \right) \left(\sum_{k=l}^m c_k \sin^k\left(\frac{\theta}{2}\right) \right) \\ &= 2 \left(c_l^2 \sin^{2l}\left(\frac{\theta}{2}\right) + 2c_l c_{l+1} \sin^{2l+1}\left(\frac{\theta}{2}\right) + \dots \right) \\ &= \sum_{k=2l}^{nm} c'_k \sin^k\left(\frac{\theta_i}{2}\right), \end{aligned} \quad (3.27)$$

where c'_k is a specific coefficient for each value of k . Since $\sin^2(\theta) = \frac{1}{2}(1 - \cos(2\theta))$, Eq. (3.27) can be rewritten as

$$S_{nm}^l(\theta_i) = \sum_{\substack{k=2l \\ k \text{ even}}}^{nm} c'_k \left(\frac{1 - \cos \theta_i}{2} \right)^{k/2} + \sin \left(\frac{\theta_i}{2} \right) \sum_{\substack{k=2l+1 \\ k \text{ odd}}}^{nm} c'_k \left(\frac{1 - \cos \theta_i}{2} \right)^{(k-1)/2}. \quad (3.28)$$

Note that each element in both sums can be expanded in to a polynomial of $\cos \theta_i$. Abstracting the coefficients and combining all summed values, each basis function can be expressed simply by

$$S_{nm}^l(\theta_i) = \sum_{k=0}^{\lfloor nm/2 \rfloor} \left(a_k + b_k \sin \left(\frac{\theta_i}{2} \right) \right) \cos^k \theta_i, \quad (3.29)$$

where a_k and b_k are, again, specific coefficients for each value of k .

Using the above equation in Eq. (3.26), the camera-centric BRDF can thus be expressed as

$$\text{BRDF}(\theta_i) = \sum_{k=0}^{K-1} \left(\alpha_k + \beta_k \sin \left(\frac{\theta_i}{2} \right) \right) \cos^k \theta_i, \quad (3.30)$$

where coefficients α_k and β_k are parameters that specify the BRDF, and K is a chosen order for the BRDF.

Fig. 3.6 shows example basis functions for this “powers-of-cosine” reflectance model. The results presented in the evaluations section below demonstrate that using only a small number of low-order terms can substantially increase the performance of SfS on real data. Moreover, this BRDF is relatively cheap to use in SfS applications, as powers of $\cos \theta_i$ can easily be computed from Eq. (3.8), and the $\sin(\theta_i/2)$ term only needs to be calculated once.

3.2.6.3 Relation to Other Reflectance Models

The reflectance model introduced above has some similarities with one-dimensional versions of previously proposed reflectance models, although it cannot directly model some physical phenomena.

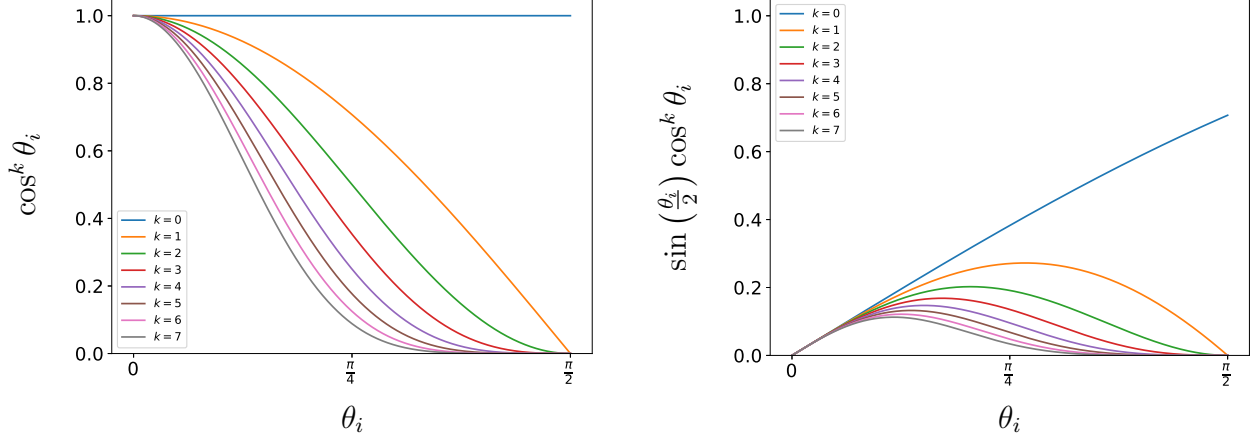


Figure 3.6: The 16 basis functions used in the proposed reflectance model with $K = 8$.

Lambertian and Phong: The proposed reflectance model can trivially capture the Lambertian and Phong reflectance models. The 1D Phong reflectance model (Phong, 1975) is defined as

$$\text{BRDF}_P(\theta_i) = a_0 + a_n \cos^n(\theta_i), \quad (3.31)$$

for some coefficient n . (This assumes the ambient lighting term is zero in Phong's model, and that n is an integer.) The Lambertian model simply has a single nonzero term, a_0 , that relates to the surface albedo.

Oren-Nayar: The 1D Oren-Nayar BRDF (Oren and Nayar, 1994) is defined as

$$\text{BRDF}_{ON}(\theta_i; \sigma) = A(\sigma) + B(\sigma) \sin(\theta_i) \tan(\theta_i), \quad (3.32)$$

where A and B are specific coefficients given model parameter σ . This BRDF is not directly compatible with the proposed reflectance model, although it is compatible if a $\cos^{-1}(\theta_i)$ term is added:

$$\begin{aligned} \text{BRDF}_{ON}(\theta_i; \sigma) &= A(\sigma) + B(\sigma) \sin(\theta_i) \tan(\theta_i) \\ &= A(\sigma) + B(\sigma) \sin^2(\theta_i) \cos^{-1}(\theta_i) \\ &= A(\sigma) + B(\sigma) (1 - \cos^2(\theta_i)) \cos^{-1}(\theta_i) \\ &= a_{-1} \cos^{-1}(\theta_i) + a_0 + a_1 \cos(\theta_i), \end{aligned} \quad (3.33)$$

with $a_{-1} = -a_1$. The $\cos^{-1}(\theta_i)$ term provides a “rough surface” approximation that negates the incident angle falloff for irradiance (Eq. 3.2) in smooth surfaces. In other words, the 1D Oren-Nayar model assumes a certain percentage of light is always reflected back towards the viewer, regardless of the incident angle.

Cook-Torrance: The 1D Cook-Torrance model (Cook and Torrance, 1982) consists of Lambertian and specular terms:

$$\text{BRDF}_{CT}(\theta_i; a_0, F, m) = a_0 + \frac{F \min(2 \cos^2(\theta_i), 1)}{m^2 \cos^6(\theta_i)} \exp\left(\frac{1 - \cos^{-2}(\theta_i)}{m^2}\right), \quad (3.34)$$

where F is the Fresnel term that relates the strength of light reflectance off the surface, and m is a scale parameter for the Beckmann distribution function. In the case of a 1D BRDF, Cook and Torrance note that F can be approximated as a constant for surfaces that are not extremely specular. The specular part of the Cook-Torrance model is not directly compatible with the reflectance model I have proposed above.

3.2.7 Iterative Update Scheme

Next, I introduce an iterative updating scheme for enhancing SfS with sparse 3D scene geometry. In principle, this method is not tied to the sparse reconstruction method used (*e.g.* rigid SfM or non-rigid SfM) – it only requires 3D points associated with 2D observations in the given image. For the experiments presented below, I use a rigid SfM implementation (Schönberger and Frahm, 2016) and, for the experiments on live endoscopies, operate on small groups of temporally neighboring frames without large surface deformation.

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. The method has a “flavor” of expectation-maximization algorithms in the sense that it iterates between optimizing a set of parameters (the reflectance model) based on the existing surface and computing expected depths using these parameters.

3.2.7.1 Warping

Denote the estimated surface at iteration n of the iterative scheme as f_n . For initialization, an estimated surface f_0 is defined having $r(x, y) = 1$, where r is defined in Eq. (3.5). First, an image-space warp of f_n is performed using the 3D SfM points with known distance $\hat{r}_k(x_k, y_k)$ as control points. For each SfM point, the ratio $\rho_k = \hat{r}_k/r_k$ is estimated, where r_k is the point's (bilinearly interpolated) distance on f_n . To minimize the effect of outlier points from SfM, I adopt a nearest-neighbor approach to define the warping function: For each pixel (x, y) in the image, the N closest SfM points in the image plane are taken. In my experiments, I use $N = 10$. Then, the warp function at that pixel is defined as $\rho(x, y) = \sum w_k \rho_k / \sum w_k$, where the sum is over the set of neighboring SfM points. The per-point weight is set as $w_k = \exp(-d_k)$, where d_k is the distance in the image plane between (x, y) and the SfM point (x_k, y_k) . The new surface is calculated as $f_n^{\text{warp}}(x, y) = \rho(x, y)f_n(x, y)$.

3.2.7.2 Reflectance Model Estimation

From this warped surface, optimization is performed to update the reflectance model parameters Θ for the specified BRDF (where the parameters depend on what BRDF that is chosen, such as $\{\alpha_k, \beta_k\}$ for the model proposed above or a constant albedo for the Lambertian model). This optimization is done by minimizing the error over all SfM points (cf. Eq. (3.10)):

$$E(\Theta) = \sum_{I_k, \hat{r}_k, \theta_k} \Phi(\eta(\theta_k; \Theta) - I_k \hat{r}_k^2) + \Psi(\Theta), \quad (3.35)$$

where I_k , \hat{r}_k , and θ_k are the observed luminance, original distance, and current estimated incident angle for the k^{th} input SfM point. An example result is shown in Fig. (3.7). Because the warped surface may not exactly pass through the SfM points, each θ_k is obtained from the surface point in f_n^{warp} that is closest to the original SfM point, rather than directly from $f_n^{\text{warp}}(x_k, y_k)$. The term Φ is a robust function to help avoid outliers in the fit, and Ψ is a regularization function for the estimated

model parameters. For Φ , I use a Huber function (Huber, 1981) that is applied per point with a threshold of $\tau \hat{r}_k^2$, with $\tau = 0.1$ in my experiments.

The Ψ term is necessary to improve the conditioning of the proposed BRDF model — without this, the surfaces estimated by the method can vary widely with just a small change in the input. I use a Tikhonov regularization: $\Psi(\Theta) = \frac{\alpha^2 \mu(\hat{r}_k^2)}{2K} \|\Theta\|_2^2$, with $\alpha = 0.01$ and where $\mu(\cdot)$ represents the mean. The algorithm is somewhat sensitive to the order of magnitude of α : it cannot be too large or too small. I do not use any regularization when fitting a Lambertian model.

Instead of fitting to the SfM points only, another option is to perform the fit over the entire warped surface. However, this is often highly sensitive to geometric inaccuracies in the warped surface, and I have found that the approach gives generally inferior results (Fig. 3.7).

3.2.7.3 SfS with Estimated BRDF

Following reflectance model estimation, PDE framework introduced above (Eq. (3.13)) is then applied using the warped surface f_n^{warp} for values of z_{est} and using the current estimated reflectance model.

Concerning values of $\lambda(x, y)$ in the regularized PDE (Eq. (3.13)), $\lambda > L$ will give greater weight to f_n^{warp} , while $\lambda < L$ will favor a purely SfS solution. The weighting is decided based on agreement between the SfM points and f_n^{warp} . Let Δr_k be the distance between a 3D SfM point with distance \hat{r}_k and its corresponding point on f_n^{warp} . The agreement between the warped surface and the SfM point is defined as $\lambda_k = \max\left(\log_{10} \frac{\hat{r}_k}{2\Delta r_k}, 0\right)$. This equally weights SfM and SfS (*i.e.* $\lambda_k = 1$) when Δr_k is 5% of \hat{r}_k and $L = 1$. The log term serves to increase λ_k by 1 for every order-of-magnitude decrease in $\Delta r_k / \hat{r}_k$. Just as for $\rho(x, y)$ above, the same nearest-neighbor weighting scheme is used to define $\lambda(x, y)$ based on the λ_k values at the SfM control points.

3.2.7.4 Iteration

Once SfS has been performed, a newly estimated surface f_{n+1} is obtained. The algorithm then re-warps the surface, re-estimates the reflectance model, and re-runs regularized SfS. This iterative

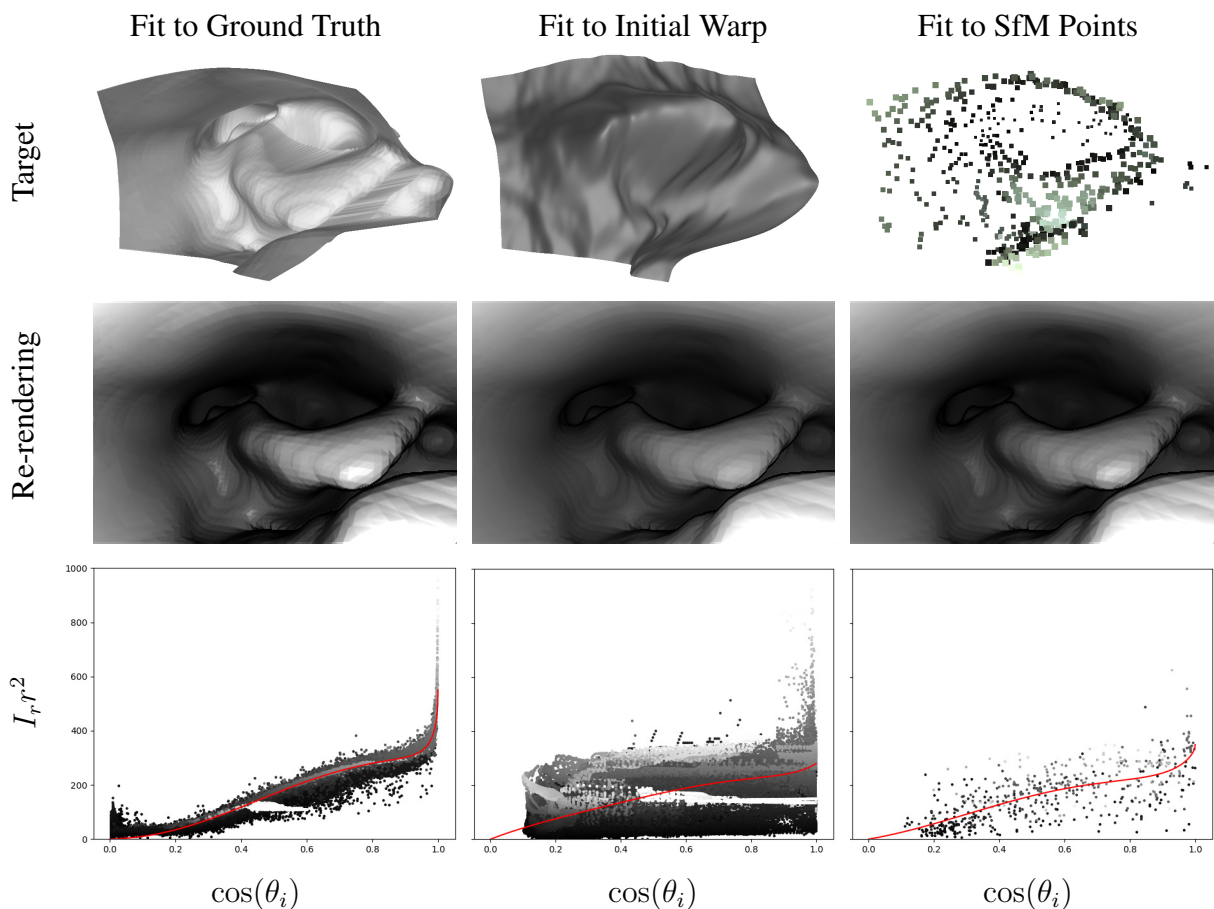


Figure 3.7: Example fitting results for the $K = 5$ model, using the ground-truth surface (left), warped surface (middle), and sparse SfM points (right) on a synthetic image. The top row shows the initial surfaces and points use for the fitting; these target values are scattered in the graphs in the bottom row, with each value colored by its observed intensity for visualization. The red curves in the bottom row plot the reflectance function $\eta(\theta_i; \Theta)$ whose parameters Θ have been robustly fit to the plotted points. The middle row shows a re-rendering of the ground-truth surface using these fit functions. Fitting to the SfM points alone is more reliable than fitting to the entire warped surface, which may contain errors in depth as well as in $\cos \theta_i$. In this example, the near-specular effects of the material are better captured by the fit using the SfM points.

process is repeated for a maximum number of iterations or until convergence. In my implementation, convergence is reached when the average change in $z(x, y)$ over the entire image is less than 1%. For the majority of images, this is usually reached within 3-6 iterations. Within each iteration, I use a tighter threshold for convergence for SfS, stopping when the maximum change in $z(x, y)$ is less than 0.1%.

3.2.8 Accounting for Interreflections in Real Endoscopic Scenarios

Up to now, we have considered a single-interaction lighting model, where each light ray is assumed to intersect with the viewed surface exactly once. In real-world applications, however, photons colliding with a surface will be scattered in all directions (which is modeled by the BRDF), and thus many photons emitted from the light source will reflect off the surface at multiple points before they collide with the camera sensor. In other words, the actual observed radiance is the sum of the radiance from the illuminant alone *plus* the strength of interreflections (Forsyth and Zisserman, 1991):

$$I_r = E_i \text{BRDF}(\theta_i) + \int_0^{\pi/2} \int_0^{2\pi} \tilde{E}_i(\theta, \phi) \text{BRDF}(\theta, \phi, \theta_r, \phi_r) d\theta d\phi \quad (3.36)$$

(cf. Eq. (3.3)). Here, $\tilde{E}_i(\theta, \phi)$ represents the amount of interreflected light irradiating a point from a given incident angle w.r.t the surface normal, and the integral is taken over the entire unit hemisphere. From an extremely pessimistic perspective, these interreflections would completely invalidate the simple model used for SfS: the light that enters the camera sensor is, in reality, more than just the light that bounced off of the surface directly towards the camera. On the other hand, since some amount of light is absorbed with each interreflection (Nayar et al., 1991), the contribution to overall radiance is mainly derived from the first few collisions. So, there is an upper bound to the error incurred by ignoring interreflections.

Intuitively, the result from the proposed SfMS approach should improve given some additional model of the interreflection function. Here, I propose a relatively simple approach for this approxi-

mation that assumes spatial contiguity of the interreflection strength. The idea is that the amount of additional interreflected light entering the camera is, in general, likely to be similar for two nearby points on the imaged surface. Interreflection can therefore be modeled to some approximation in the image domain, taking into account that nearby pixels represent nearby points on the surface except at occlusion boundaries. If we ignore occlusion boundaries and assume that the amount of interreflection varies only slowly over the image, one option for modeling interreflection is to use a low-order polynomial in x and y . The interreflection integral in Eq. (3.36) is changed to an approximating form of

$$I_r = E_i \text{BRDF}(\theta_i) + \sum_{i=0}^N \sum_{j=0}^N c_{ij} x^i y^j, \quad (3.37)$$

where each c_{ij} is a coefficient of the order- N polynomial. These extra coefficients are included as parameters during reflectance model estimation. Integrating the coefficients into the SfS formulation is also straightforward. Denoting the sum in Eq. (3.37) as I_f , the initial PDE (Eq. (3.9)) becomes

$$(x^2 + y^2 + 1)(I_r - I_f)e^{2v} - I_i A \cos(\theta_i) \text{BRDF}(\theta_i) = 0. \quad (3.38)$$

The rest of the approach needs no adaption.

In practice, I have found that $N = 2$ (9 total coefficients) gives adequate gains for the overall estimation, with diminishing returns for larger values of N . One caveat to this approach is that the low-order polynomial is dependent on the 2D placement of the SfM feature points — if the feature points are not distributed across the image, unrealistic values of I_f can frequently arise. I account for this simply by clipping values of I_f to the range $[0, 0.1]$. The reasoning here is that interreflection is always additive and should only contribute a small amount to the overall luminance (in the range $[0, 1]$). The chosen value of 0.1 is a heuristic.

This low-order approximation is admittedly quite simple, and I expect that much more elegant approaches are likely possible. For example, 2D splines could be used to better account for sharp changes in I_f , although the reliability of this more specialized fitting depends even more strongly on the 2D placement of the SfM feature points, compared with a low-order approximation. Simplicity

of the proposed method notwithstanding, I demonstrate that even using this crude approximation can significantly increase the overall accuracy of the approach.

3.3 Evaluation

In this section, I provide quantitative and qualitative evaluations of the proposed BRDF model and the overall SfMS framework.

3.3.1 Comparison of BRDF Fits

Fig. 3.8 provides a qualitative comparison of how well different 1D BRDF models are able to approximate real-world material reflectances for eight different materials taken from the MERL database (Matusik et al., 2003). In this experiment, I render each material according to its theoretical observed radiance at unit depth (*i.e.*, $I_r(\theta_i) = \rho \cos(\theta_i) \text{BRDF}(\theta_i)$, cf. Eq. (3.3)) with a normalization ρ set such that the highest intensity for any value of $I_r(\theta_i)$ is rendered with an intensity of 1 (pure white). The top row of each image shows the observed color of each material as a function of θ_i , and the second row shows the same function as a grayscale (luminance) image. Each subsequent row in each image shows a least-squares fit of a different reflectance model to the luminance function. I have taken the example to an extreme and shown fits for the proposed reflectance model up to $K = 10000$. While this is in no way practical for implementation, it helps demonstrate the full behavior of the reflectance basis versus the other models.

Table 3.1 shows an exhaustive list of radiance-fitting errors over the 100 materials in the MERL database, for each of the analyzed BRDFs. As each ground-truth reflectance function is scaled such that its largest value equals 1, the error values are not comparable between different materials; however, the results for each material are directly comparable among the different models. It is apparent that the Lambertian and Oren-Nayar BRDFs are often poor approximations for all but the most diffuse/rough surfaces. The Phong and Cook-Torrance models often perform much better, with the Phong model typically achieving slightly better approximations. Naturally, the proposed model always improves for larger values of K , and for many diffuse materials, only a small number of

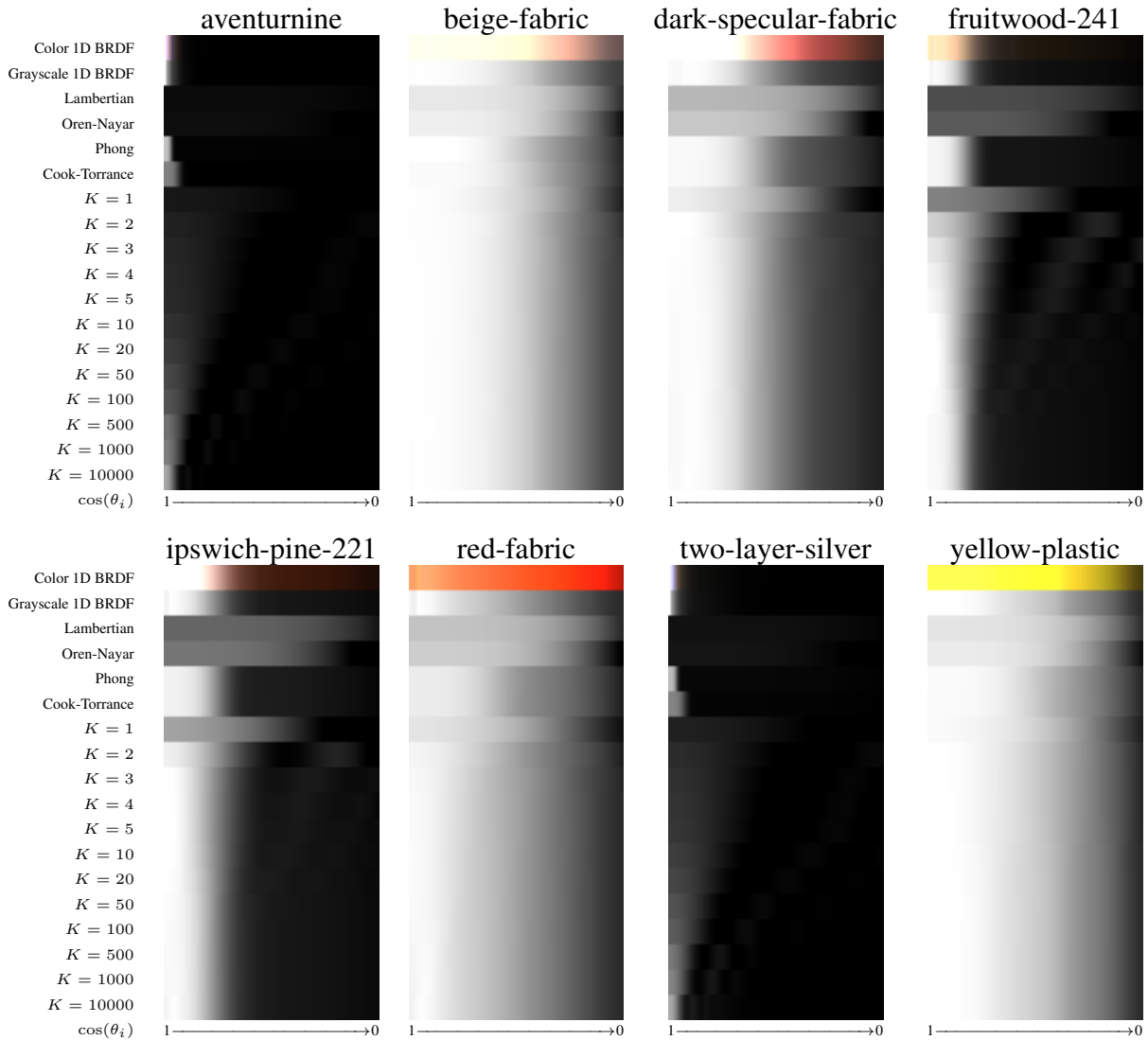


Figure 3.8: Estimated 1D radiance functions for different materials from the MERL database (Matusik et al., 2003). The top row of each image shows the color radiance, $\cos(\theta_i)BRDF_\lambda(\theta_i)$, and the second row shows the luminance equivalent. Subsequent rows show least-square fits to the luminance function for different BRDF models.

coefficients ($K \leq 5$) are necessary to outperform the other reflectance models. Specular materials typically require many more coefficients, with the proposed model only outperforming the fitting of Phong’s model when $K = 10000$, if at all. Since Phong’s model can be thought of as a sparse version of the proposed model, it follows that employing sparsity constraints in the proposed model would help the overall performance across all materials without needing a huge number of coefficients. In other words, one can use a small number of coefficients in the proposed model (say, $K = 5$ or $K = 10$) and additionally fit a specular term, $a_n \cos^n(\theta_i)$, where both a_n and n are free parameters.

However, while the addition of high-order specular terms is reasonable for fitting to known BRDFs and general graphics applications, there are some difficulties in applying these terms in computer vision applications like the SfMS approach I propose. The main issue is that specular regions are prone to *oversaturation*. Digital camera sensors operate by, to first approximation, counting the number of photons that collide with a pixel over a given period of exposure. Especially for cameras that do not have high-dynamic-range sensors (which is typically the case for endoscopic devices), the problem is that there exists an upper limit to the amount of photon charge that can be accumulated for a given pixel — past a certain point, the total number of photons has no effect on the resulting pixel value. For specular surfaces, such oversaturation frequently occurs at specularities, *i.e.*, points where $\cos(\theta_i)$ is close to 1. The effect can also occur when the camera/light is very close to the surface. This is problematic for fitting a BRDF model to observed luminance values, since the nuances of *actual* luminance are completely destroyed for such image regions. Effectively, specularities in the image can be interpreted as missing data, thus there is not a strong advantage to using high-order specular terms when fitting a BRDF model in the proposed SfMS approach. The best approach would be to detect specularities in the image, excise them, and fill in the region using an image imputation method. I do not adopt this approach in my experiments here, however.

	alum-bronze	alumina oxide	aluminum	aventu-rmine	beige fabric	black fabric	black obsid	blk-ox steel	black phenol	black plastic	blue acrylic	blue fabric	bl-metal paint	bl-metal paint2	blue rubber	brass	cherry 235	chrome	chrome steel	colonial maple
Lambertian	4.3356	2.2722	1.7267	1.5143	0.4961	0.2801	1.7332	6.0288	3.6218	5.6718	1.8640	0.1122	9.8568	1.7460	0.8565	1.9310	7.4195	1.5882	1.5518	4.7559
Oren-Nayar	3.9684	2.2165	1.6847	1.4848	0.2925	0.1171	1.6904	3.9293	3.4579	3.8878	1.8233	0.1054	7.8182	1.7074	0.5391	1.8871	6.1279	1.5564	1.5236	3.9509
Phong	0.3294	0.1057	0.1585	0.1677	0.0473	0.2801	0.2139	0.0657	0.1639	0.0354	0.1422	0.0686	0.0370	0.1420	0.0123	0.1397	0.0613	0.1764	0.1793	0.0266
Cook-Torr.	0.3295	0.2196	0.2378	0.3855	0.0231	0.2787	0.2796	0.0772	0.1644	0.0418	0.2983	0.1021	0.0395	0.3179	0.0143	0.2220	0.0631	0.2297	0.2238	0.0278
K=1	2.8097	2.0082	1.5400	1.3709	0.0545	0.2527	1.5252	0.8045	2.8495	0.8549	1.6663	0.0780	3.0317	1.5691	0.1223	1.7322	2.7016	1.4560	1.4508	1.8830
K=2	1.2928	1.8321	1.4752	1.2485	0.0072	0.0140	1.3472	0.0197	1.8163	0.0338	1.4960	0.0345	0.1505	1.4984	0.0058	1.6587	0.1722	1.4195	1.4243	0.0812
K=3	1.0321	1.7370	1.4353	1.1900	0.0029	0.0073	1.2635	0.0119	1.5986	0.0058	1.4158	0.0034	0.0788	1.4603	0.0029	1.6155	0.0383	1.3916	1.4044	0.0198
K=4	0.9254	1.6770	1.4059	1.1598	0.0010	0.0039	1.2203	0.0071	1.4684	0.0018	1.3743	0.0017	0.0675	1.4308	0.0029	1.5830	0.0329	1.3705	1.3885	0.0170
K=5	0.8302	1.6343	1.3810	1.1342	0.0006	0.0023	1.1836	0.0026	1.3692	0.0004	1.3389	0.0017	0.0640	1.4049	0.0022	1.5551	0.0298	1.3534	1.3754	0.0126
K=10	0.6079	1.4834	1.2956	1.0496	0.0004	0.0016	1.0640	0.0003	1.0513	0.0001	1.2229	0.0017	0.0403	1.3181	0.0016	1.4590	0.0236	1.2944	1.3307	0.0047
K=20	0.4336	1.3184	1.2060	0.9566	0.0003	0.0008	0.9343	0.0002	0.7596	0.0001	1.0953	0.0015	0.0124	1.2259	0.0009	1.3567	0.0161	1.2293	1.2805	0.0006
K=50	0.2493	1.0771	1.0822	0.8213	0.0001	0.0005	0.7514	0.0002	0.4407	0.0001	0.9107	0.0011	0.0010	1.0954	0.0003	1.2116	0.0141	1.1357	1.2053	0.0002
K=100	0.1488	0.8849	0.9874	0.7120	0.0001	0.0005	0.6107	0.0002	0.2666	0.0000	0.7628	0.0010	0.0008	0.9941	0.0002	1.0982	0.0119	1.0586	1.1400	0.0002
K=500	0.0480	0.4692	0.7680	0.4638	0.0001	0.0004	0.3351	0.0001	0.0752	0.0000	0.4376	0.0009	0.0007	0.7603	0.0001	0.8322	0.0010	0.8532	0.9503	0.0002
K=1000	0.0319	0.3317	0.6735	0.3724	0.0000	0.0004	0.2566	0.0001	0.0430	0.0000	0.3255	0.0009	0.0004	0.6613	0.0001	0.7171	0.0010	0.7524	0.8499	0.0001
K=10000	0.0200	0.0889	0.3773	0.1675	0.0000	0.0003	0.1287	0.0000	0.0083	0.0000	0.1035	0.0009	0.0002	0.3500	0.0001	0.3576	0.0007	0.4156	0.4767	0.0001
	color-chg paint1	color-chg paint2	color-chg paint3	dark-bl paint	dark-rd fabric	dk-spec fabric	delrin	gl-wood-241	gl-metal paint	gl-metal paint2	gl-metal paint3	gold paint	gray plastic	greased steel	green acrylic	green fabric	green latex	gr-met paint	gr-met paint2	green plastic
Lambertian	6.3303	6.9925	5.6976	4.1957	6.6435	3.6804	2.3324	7.1388	9.6879	1.3774	2.0022	8.5734	3.8949	2.6471	3.2351	3.4566	0.7041	9.7216	2.0321	1.2934
Oren-Nayar	5.8161	6.2955	5.2583	2.7733	0.2850	2.4175	1.5573	6.3747	7.4268	1.3400	1.9470	6.6182	3.6717	2.5542	3.1433	0.3006	0.4861	7.5226	1.9789	1.2714
Phong	0.0384	0.1050	0.1463	0.0313	0.0069	0.0291	0.1361	0.0315	0.0405	0.2116	0.1646	0.0491	0.1656	0.1579	0.0599	0.0648	0.0779	0.0369	0.1344	0.2175
Cook-Torr.	0.0385	0.1054	0.1464	0.0374	0.0104	0.0357	0.1368	0.0317	0.0442	0.3579	0.2022	0.0528	1.6812	0.1580	0.2917	0.1587	0.0831	0.0398	0.1924	0.3987
K=1	4.0595	4.0338	3.7506	0.6251	0.0218	0.5679	5.5539	3.9872	2.4935	1.2086	1.7528	2.3079	2.8930	2.2077	2.7918	0.1268	0.1218	2.6822	1.7787	1.1865
K=2	1.4482	1.0926	1.4786	0.0184	0.0047	0.0214	0.0653	1.0242	0.0674	1.1108	1.6570	0.0820	1.5680	1.9628	2.2820	0.0172	0.0319	1.6318	1.6318	1.1105
K=3	0.9621	0.6627	1.0425	0.0013	0.0027	0.0070	0.0216	0.4708	0.0444	1.0635	1.6046	0.0344	1.2510	1.8545	2.1399	0.0045	0.0024	0.0784	1.5640	1.0693
K=4	0.7608	0.5080	0.8596	0.0009	0.0011	0.0062	0.0144	0.2815	0.0418	1.0326	1.5646	0.0312	1.0840	1.7587	2.0386	0.0044	0.0023	0.0714	1.5063	1.0449
K=5	0.6028	0.3812	0.7091	0.0007	0.0008	0.0031	0.0139	0.1810	0.0378	1.0119	1.5300	0.0281	0.9681	1.6877	1.9588	0.0030	0.0021	0.0652	1.4647	1.0271
K=10	0.2794	0.1481	0.3834	0.0005	0.0004	0.0015	0.0136	0.0664	0.0158	0.9442	1.4143	0.0142	0.6668	1.4710	1.7146	0.0017	0.0016	0.0341	1.3335	0.9637
K=20	0.1031	0.0469	0.1774	0.0002	0.0003	0.0015	0.0122	0.0413	0.0027	0.8758	1.2940	0.0038	0.4236	1.2376	1.4525	0.0011	0.0006	0.0071	1.1893	0.8941
K=50	0.0309	0.0110	0.0555	0.0001	0.0002	0.0011	0.0113	0.0284	0.0007	0.7813	1.1291	0.0004	0.2087	0.9271	1.0866	0.0010	0.0003	0.0010	0.8746	0.7912
K=100	0.0205	0.0063	0.0202	0.0001	0.0002	0.0004	0.0103	0.0119	0.0005	0.7049	1.0251	0.0002	0.1139	0.9011	0.8111	0.0010	0.0002	0.0005	0.8346	0.7074
K=500	0.0078	0.0041	0.0030	0.0001	0.0001	0.0003	0.0047	0.0016	0.0003	0.5124	0.7332	0.0001	0.0266	0.3355	0.2842	0.0009	0.0002	0.0001	0.5165	0.5124
K=1000	0.0068	0.0036	0.0020	0.0001	0.0001	0.0002	0.0044	0.0015	0.0002	0.4318	0.6237	0.0001	0.0159	0.2403	0.1464	0.0008	0.0001	0.0001	0.4068	0.4367
K=10000	0.0038	0.0027	0.0013	0.0001	0.0001	0.0001	0.0024	0.0013	0.0001	0.2258	0.3084	0.0000	0.0024	0.0970	0.0108	0.0008	0.0001	0.0001	0.1634	0.2532
	hematite	ipswich pine-221	lt-brown fabric	lt-red paint	maroon plastic	natural rubber	neoprene	nickel	nylon	orange paint	pearl paint	pickled oak-260	pink fabric	pink fabric2	pink felt	pink jasper	pink plastic	polyeth-ylene	polyrub-ber	pure rubber
Lambertian	1.5571	7.3643	0.4773	1.0076	1.7878	6.6377	2.5573	6.6636	3.3822	0.5160	7.4396	5.2238	0.3882	0.7531	0.8006	2.4468	0.3172	0.5246	0.7008	0.5492
Oren-Nayar	1.5264	6.2714	0.0606	0.4178	1.7503	5.6969	1.8623	5.9643	2.9231	0.1789	5.2903	4.6127	0.1550	0.3809	0.3407	2.3669	0.1745	0.3008	0.0218	0.2080
Phong	0.1376	0.0477	0.4773	0.0483	0.1234	0.0327	0.0827	0.1094	0.0842	0.0145	0.0176	0.0170	0.0162	0.0162	0.1451	0.1779	0.0191	0.0144	0.7008	0.0167
Cook-Torr.	0.3194	0.0489	0.4773	0.0617	0.2415	0.0337	0.0841	0.1098	0.0845	0.0257	0.0224	0.0170	0.2871	0.0178	0.0093	0.2686	0.0897	0.0162	0.7008	0.0271
K=1	1.4102	3.0949	0.2931	0.0264	1.6059	2.9346	0.7319	3.7246	1.7951	0.0048	1.3666	2.6651	0.0373	0.0769	0.0503	2.0704	0.0267	0.0508	1.5488	0.0156
K=2	1.3287	0.3097	0.0188	0.0035	1.4574	0.2674	0.0730	1.2309	0.4083	0.0036	0.0497	0.3354	0.0041	0.0088	0.0056	1.6408	0.0055	0.0108	0.0304	0.0111
K=3	1.2927	0.0399	0.0024	0.0017	1.3830	0.0339	0.0134	0.8806	0.1850	0.0006	0.0412	0.0915	0.0024	0.0066	0.0038	1.5330	0.0007	0.0001	0.0055	0.0076
K=4	1.2599	0.0284	0.0021	0.0016	1.3446	0.0257	0.0088	0.6529	0.1085	0.0005	0.0232	0.0662	0.0023	0.0032	0.0037	1.4535	0.0005	0.0071	0.0026	0.0061
K=5	1.2347	0.0220	0.0014	0.0012	1.3128	0.0210	0.0078	0.5199	0.0686	0.0004	0.0145	0.0590	0.0022	0.0020	0.0029	1.3938	0.0005	0.0070	0.0025	0.0051
K=10	1.1567	0.0158	0.0007	0.0008	1.2055	0.0138	0.0063	0.2223	0.0254	0.0002	0.0025	0.0418	0.0012	0.0006	0.0009	1.2120	0.0004	0.0058	0.0023	0.0045
K=20	1.0690	0.0078	0.0003	0.0002	1.0876	0.0048	0.0039	0.0776	0.0188	0.0001	0.0009	0.0126	0.0005	0.0006	0.0007	1.0255	0.0003	0.0029	0.0021	0.0017
K=50	0.9430	0.0029	0.0002	0.0002	0.9158	0.0005	0.0005	0.0199	0.0165	0.0001	0.0002	0.0020	0.0001	0.0005	0.0001	0.7801	0.0003	0.0007	0.0006	0.0007
K=100	0.8435	0.0024	0.0002	0.0001	0.7769	0.0002	0.0003	0.0118	0.0147	0.0001	0.0002	0.0012	0.0001	0.0004	0.0001	0.6063	0.0002	0.0005	0.0003	0.0004
K=500	0.6195	0.0019	0.0002	0.0001	0.4636	0.0001	0.0001	0.0114	0.0137	0.0001	0.0001	0.0004	0.0000	0.0004	0.0001	0.2926	0.0002	0.0001	0.0002	

3.3.2 Ground-truth Geometric Evaluation

I validate the proposed SfMS approach against on a rigid ground truth (GT) dataset. The GT model consists of a 3D printing of the pharynx of the throat from a patient CT scan (cavity width $\sim 2\text{cm}$). This model was then re-scanned with CT to produce a highly accurate GT mesh. An endoscopic video of this model was captured at 60 Hz with a resolution of 720×240 pixels, the frames of which were corrected for radial distortion prior to Structure-from-Motion. Following SfM, the iterative closest surface algorithm (Rusinkiewicz and Levoy, 2001) was used to align the SfM point cloud to the GT model.

For Shape-from-Shading, the observed image radiance was modeled using the L^* channel of the $L^*a^*b^*$ color space normalized to $[0, 1]$. For all analysis, the SfM point cloud was first filtered to only contain points seen by 5 or more cameras and having at least one triangulation angle (the angle between the rays of a 3D point to each of two observing cameras) of 10 degrees or greater. For each image, the algorithm only considers 3D points with corresponding 2D features in the image.

The proposed method was applied to 100 endoscopic video frames from the GT video on both simulated and actual images. For simulation, renderings of the GT surface were generated from the camera poses and parameters obtained from SfM on the real video. In each simulated image, the virtual surface was illuminated by a point light source co-located with the camera, and different surface material properties were recreated using 8 GT BRDFs from the MERL database (Matusik et al., 2003). Algorithm parameters and SfM points were the same for all trials – only the rendered images differed. Since the synthetic images were rendered without modeling surface interreflections, interreflection coefficients were only estimated for the real endoscopy sequence.

The proposed iterative framework is evaluated for the Lambertian BRDF and for the proposed reflectance model (Eq. 3.30) with K from 1 to 5. Table 3.2 provides statistics (mean and std. dev. across the 100 images) on the percentage of pixels in each image within a given threshold (0.5, 1, 1.5, and 2mm) of the GT surface. For the simulated data, the algorithm estimates depth to within 2mm of the GT for the majority of the pixels; there is a fall-off in accuracy under tighter thresholds that qualitatively correlates to the specularities of the rendered material. The proposed reflectance

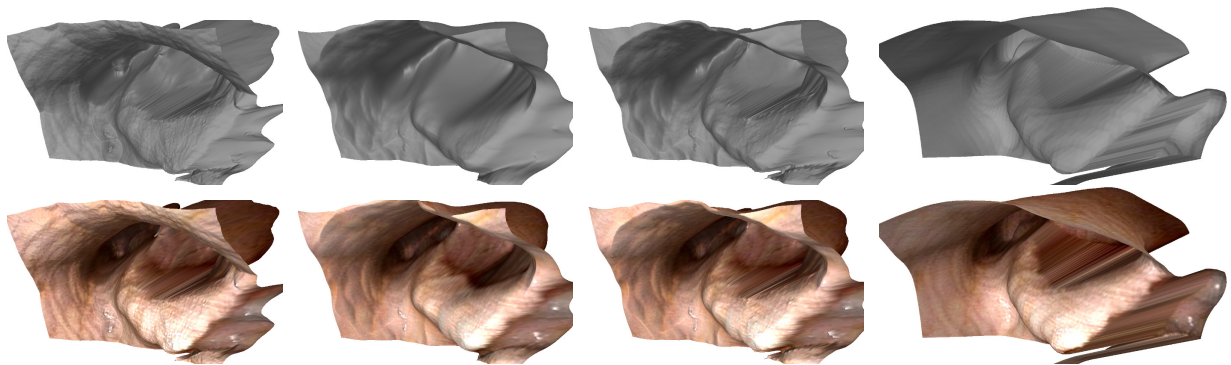


Figure 3.9: Visual comparison of surfaces generated by the proposed approach for an image from a ground-truth dataset. Top/bottom rows: Visualization of the surface without/with texture from the original image. Columns from left to right: (1) using a Lambertian BRDF, (2) using the proposed BRDF ($K = 2$) without image-weighted derivatives, (3) using the 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) and the flattened curve of the epiglottis in column (1).

model achieves comparable performance to the Lambertian model for more diffuse BRDFs (*e.g.* “beige-fabric” and “red-fabric”) and better performance for surfaces with more non-Lambertian properties (*e.g.* “fruitwood-241” and “two-layer-silver”). On the real endoscopy of the GT model (see the last entry in Table 3.2), the proposed reflectance model recovers approximately 6-10% more pixels, on average, within the given thresholds compared to the Lambertian BRDF. Fig. 3.9 provides a visual comparison between using the Lambertian and proposed BRDFs on the GT sequences.

An evaluation of the proposed use of image-weighted derivatives is also performed, as well as an assessment regarding how the number of available SfM points affects reconstruction accuracy in the method. Regarding the latter, frames in the GT video sequence observe between 300 and 600 SfM points. I randomly select a subset of SfM points in each frame (25, 50, 100, and 150 points) and evaluate how the algorithm performs with a smaller number of 3D points. Table 3.3 summarizes the resulting performance, which increases incrementally with the number of SfM points. While the use of image-weighted derivatives only increases accuracy by up to 1%, the approach greatly reduces smoothing of the solution along occlusion boundaries (see Fig. 3.9).

Table 3.4 shows results of the method on the real endoscopic video sequence without the proposed approach to modeling surface interreflections. Compared to the results in Table 3.2, ignoring interreflections drastically reduces the accuracy of the method on real data.

3.3.3 Results on Patient Data

I have also applied the method to live endoscopic datasets using manually selected intervals (typically 4-6s) with minimal surface deformation. The parameters used for SfMS in these experiments are the same as those used for the experiments on the phantom dataset. Example output for different patients is shown in Figs. 3.10, 3.11, and 3.12, using the proposed reflectance model with $K = 2$.

3.4 Discussion

In this chapter, I introduced a method combining Structure-from-Motion and Shape-from-Shading to reconstruct a surface for a single endoscopy image. SfM was used as a sparse prior for the underlying surface of the scene, under the principle that certain points in the scene can be triangulated with an approximate certainty, but that poor texturing, difficult illumination and reflectance behaviors, limited camera motion, and surface deformation (admittedly assumed to be minimal in my implementation) prevents a traditional SfM+MVS approach from achieving an accurate surface reconstruction. The SfM point cloud was used to guide and regularize the SfS solution, an approach that to my knowledge has not been explored previously. The SfM result is also used to bootstrap and refine the estimation of BRDF and coarse illumination parameters for the image, which are crucial for SfS, and I introduced a new 1D BRDF basis to improve on the Lambertian shading models that have been traditionally employed.

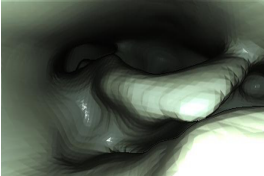
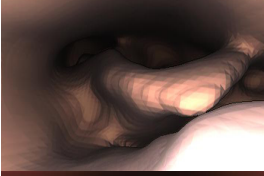
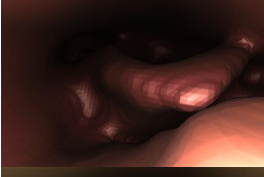
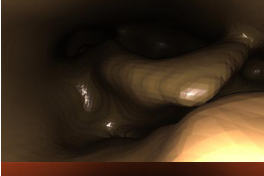
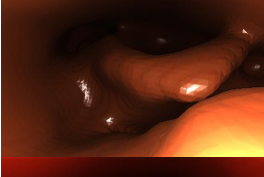
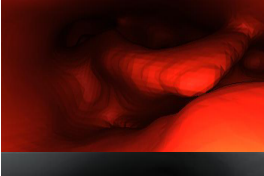

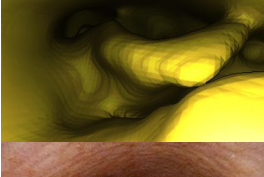
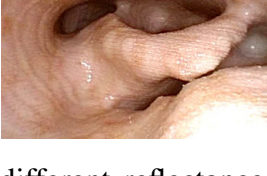
Ref. Model	Mean (Std. Dev.) Proportion of Pixels within X mm of GT				Sample Rendered Image
	0.5mm	1mm	1.5mm	2mm	
aventurine					
Lamb.	0.266 (0.072)	0.438 (0.086)	0.550 (0.096)	0.632 (0.100)	
$K = 1$	0.294 (0.064)	0.463 (0.091)	0.577 (0.104)	0.655 (0.109)	
$K = 2$	0.293 (0.073)	0.479 (0.096)	0.599 (0.108)	0.671 (0.114)	
$K = 3$	0.289 (0.074)	0.474 (0.094)	0.596 (0.108)	0.671 (0.113)	
$K = 4$	0.292 (0.075)	0.480 (0.094)	0.602 (0.109)	0.674 (0.113)	
$K = 5$	0.293 (0.075)	0.483 (0.095)	0.602 (0.111)	0.673 (0.115)	
beige-fabric					
Lamb.	0.325 (0.095)	0.482 (0.126)	0.588 (0.140)	0.668 (0.137)	
$K = 1$	0.325 (0.106)	0.492 (0.136)	0.604 (0.146)	0.682 (0.140)	
$K = 2$	0.266 (0.071)	0.451 (0.095)	0.573 (0.105)	0.655 (0.105)	
$K = 3$	0.273 (0.076)	0.449 (0.100)	0.563 (0.111)	0.644 (0.112)	
$K = 4$	0.275 (0.076)	0.445 (0.102)	0.555 (0.114)	0.636 (0.117)	
$K = 5$	0.268 (0.075)	0.439 (0.099)	0.550 (0.112)	0.634 (0.117)	
dark-specular-fabric					
Lamb.	0.296 (0.089)	0.490 (0.114)	0.625 (0.112)	0.715 (0.099)	
$K = 1$	0.297 (0.091)	0.495 (0.108)	0.631 (0.101)	0.719 (0.088)	
$K = 2$	0.249 (0.086)	0.430 (0.106)	0.564 (0.108)	0.662 (0.102)	
$K = 3$	0.238 (0.089)	0.409 (0.114)	0.542 (0.123)	0.640 (0.120)	
$K = 4$	0.231 (0.092)	0.397 (0.121)	0.526 (0.132)	0.623 (0.133)	
$K = 5$	0.228 (0.095)	0.392 (0.124)	0.519 (0.136)	0.614 (0.136)	
fruitwood-241					
Lamb.	0.381 (0.116)	0.561 (0.129)	0.673 (0.116)	0.745 (0.101)	
$K = 1$	0.407 (0.114)	0.588 (0.122)	0.690 (0.110)	0.757 (0.095)	
$K = 2$	0.386 (0.103)	0.579 (0.113)	0.694 (0.099)	0.768 (0.077)	
$K = 3$	0.380 (0.104)	0.576 (0.111)	0.692 (0.097)	0.768 (0.076)	
$K = 4$	0.381 (0.105)	0.576 (0.112)	0.693 (0.097)	0.768 (0.077)	
$K = 5$	0.380 (0.103)	0.577 (0.112)	0.693 (0.098)	0.767 (0.081)	
ipswich-pine-221					
Lamb.	0.367 (0.099)	0.535 (0.120)	0.654 (0.121)	0.732 (0.111)	
$K = 1$	0.424 (0.115)	0.598 (0.127)	0.692 (0.118)	0.751 (0.104)	
$K = 2$	0.419 (0.100)	0.601 (0.103)	0.700 (0.096)	0.764 (0.084)	
$K = 3$	0.418 (0.101)	0.604 (0.105)	0.703 (0.099)	0.766 (0.084)	
$K = 4$	0.419 (0.102)	0.605 (0.108)	0.704 (0.100)	0.766 (0.085)	
$K = 5$	0.418 (0.102)	0.604 (0.110)	0.703 (0.102)	0.764 (0.087)	
red-fabric					
Lamb.	0.345 (0.109)	0.532 (0.131)	0.649 (0.129)	0.719 (0.121)	
$K = 1$	0.345 (0.107)	0.520 (0.126)	0.636 (0.123)	0.708 (0.115)	
$K = 2$	0.319 (0.093)	0.496 (0.110)	0.610 (0.110)	0.686 (0.104)	
$K = 3$	0.326 (0.097)	0.496 (0.114)	0.608 (0.114)	0.687 (0.108)	
$K = 4$	0.328 (0.096)	0.497 (0.114)	0.609 (0.117)	0.687 (0.112)	
$K = 5$	0.329 (0.094)	0.499 (0.113)	0.612 (0.118)	0.688 (0.115)	
two-layer-silver					
Lamb.	0.283 (0.061)	0.454 (0.077)	0.565 (0.079)	0.644 (0.078)	
$K = 1$	0.287 (0.065)	0.455 (0.079)	0.565 (0.079)	0.644 (0.077)	
$K = 2$	0.279 (0.084)	0.461 (0.102)	0.578 (0.098)	0.659 (0.087)	
$K = 3$	0.283 (0.087)	0.461 (0.107)	0.574 (0.102)	0.656 (0.089)	
$K = 4$	0.286 (0.087)	0.464 (0.106)	0.575 (0.102)	0.656 (0.091)	
$K = 5$	0.297 (0.085)	0.468 (0.104)	0.577 (0.101)	0.657 (0.091)	
yellow-plastic					
Lamb.	0.337 (0.096)	0.494 (0.128)	0.596 (0.138)	0.669 (0.132)	
$K = 1$	0.341 (0.103)	0.510 (0.137)	0.615 (0.143)	0.684 (0.134)	
$K = 2$	0.275 (0.073)	0.465 (0.101)	0.586 (0.113)	0.664 (0.114)	
$K = 3$	0.279 (0.075)	0.461 (0.102)	0.580 (0.114)	0.659 (0.116)	
$K = 4$	0.277 (0.076)	0.456 (0.105)	0.574 (0.120)	0.653 (0.122)	
$K = 5$	0.271 (0.073)	0.451 (0.101)	0.568 (0.116)	0.649 (0.120)	
real endoscopic video					
Lamb.	0.106 (0.029)	0.222 (0.044)	0.334 (0.049)	0.433 (0.051)	
$K = 1$	0.167 (0.046)	0.321 (0.072)	0.445 (0.084)	0.544 (0.085)	
$K = 2$	0.172 (0.053)	0.328 (0.082)	0.451 (0.092)	0.545 (0.090)	
$K = 3$	0.162 (0.050)	0.312 (0.082)	0.430 (0.092)	0.525 (0.091)	
$K = 4$	0.161 (0.047)	0.311 (0.075)	0.426 (0.084)	0.520 (0.084)	
$K = 5$	0.164 (0.052)	0.312 (0.081)	0.426 (0.090)	0.520 (0.090)	

Table 3.2: Accuracy of the proposed SfM+SfS approach for different reflectance models on simulated and real data across 100 images. Example renderings are show in the right column.

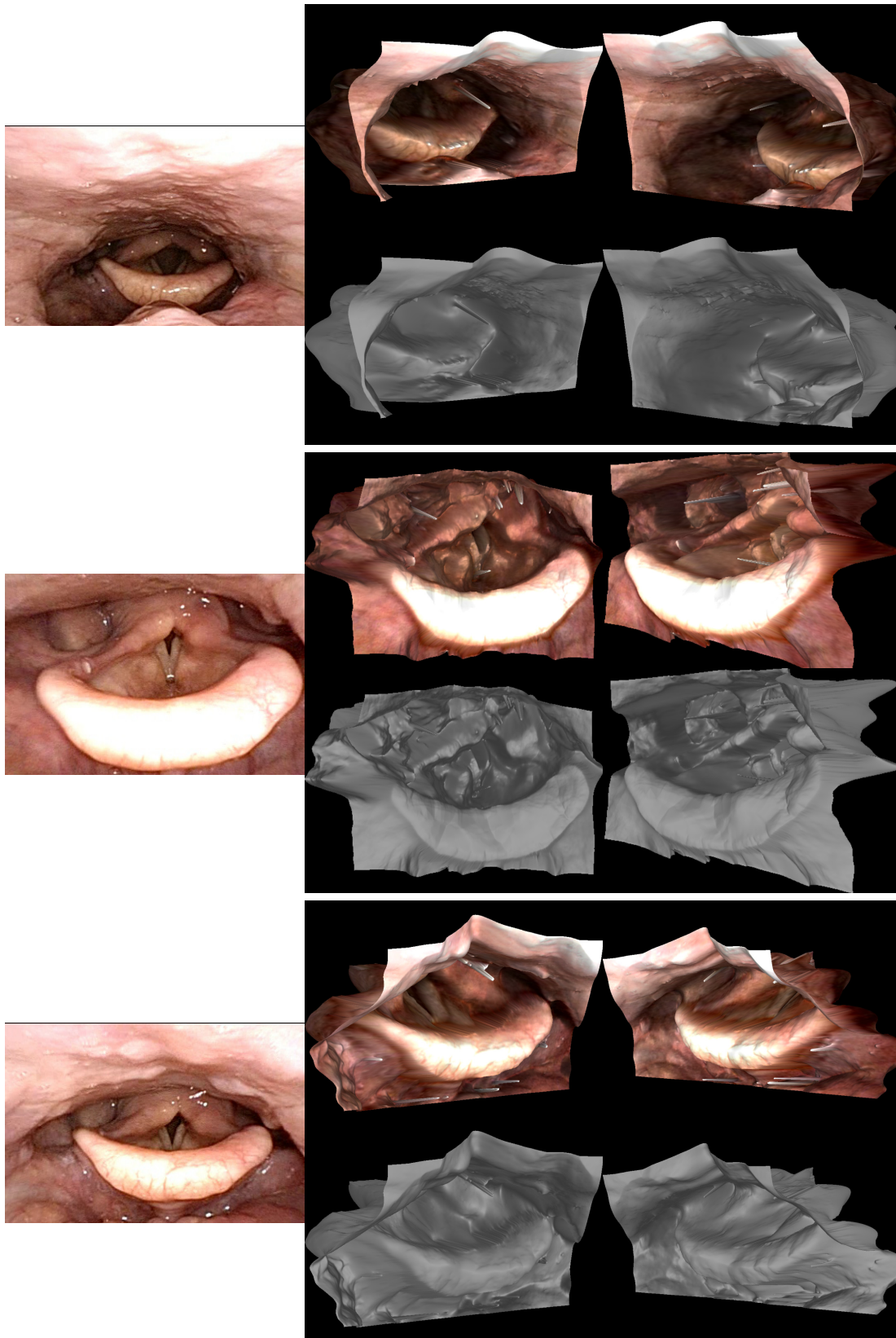


Figure 3.10: Example results for three images from a live endoscopic video. Left: Original image. Right: Surface estimated from the image using the proposed algorithm.

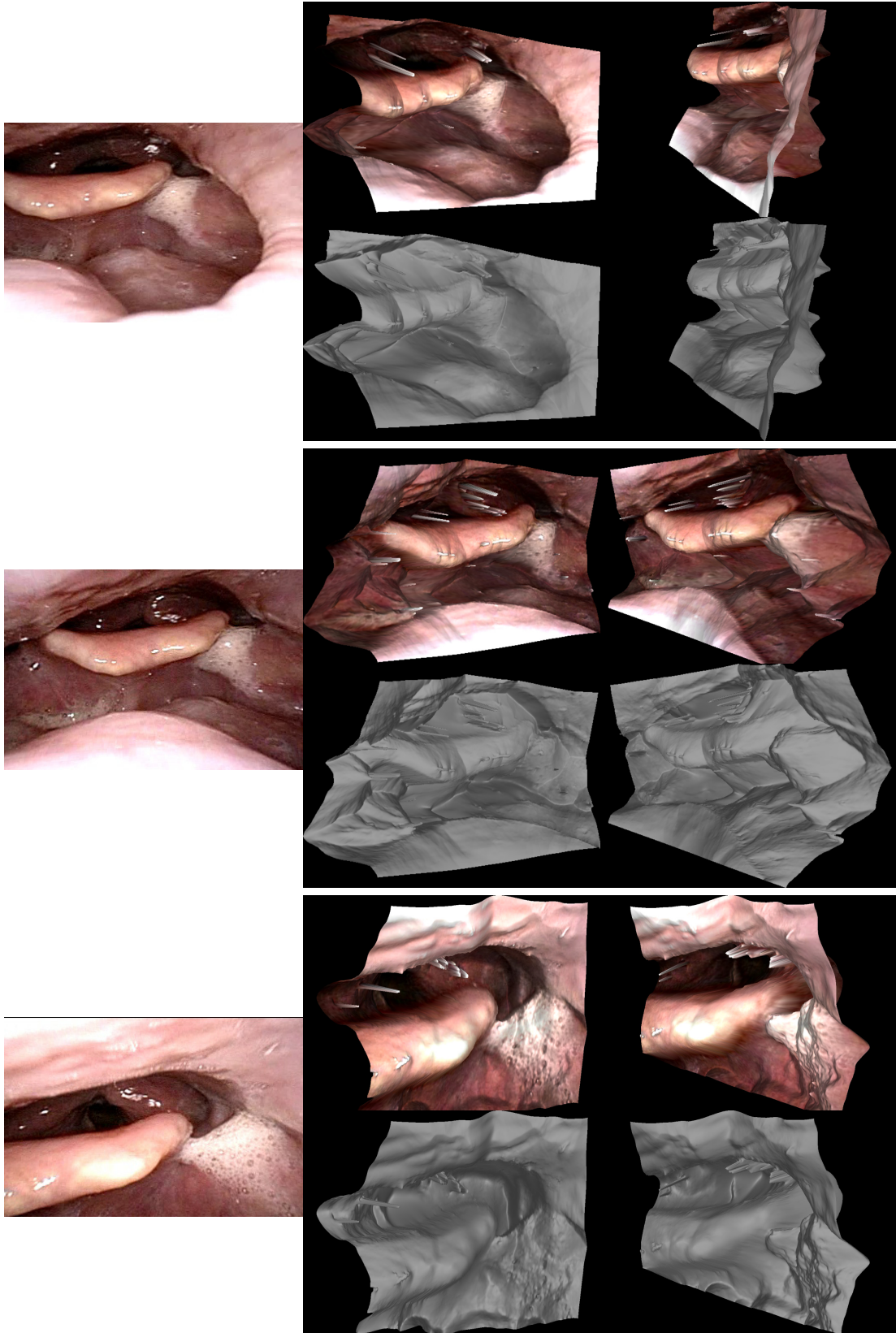


Figure 3.11: Example results for three images from a live endoscopic video. Left: Original image. Right: Surface estimated from the image using the proposed algorithm.

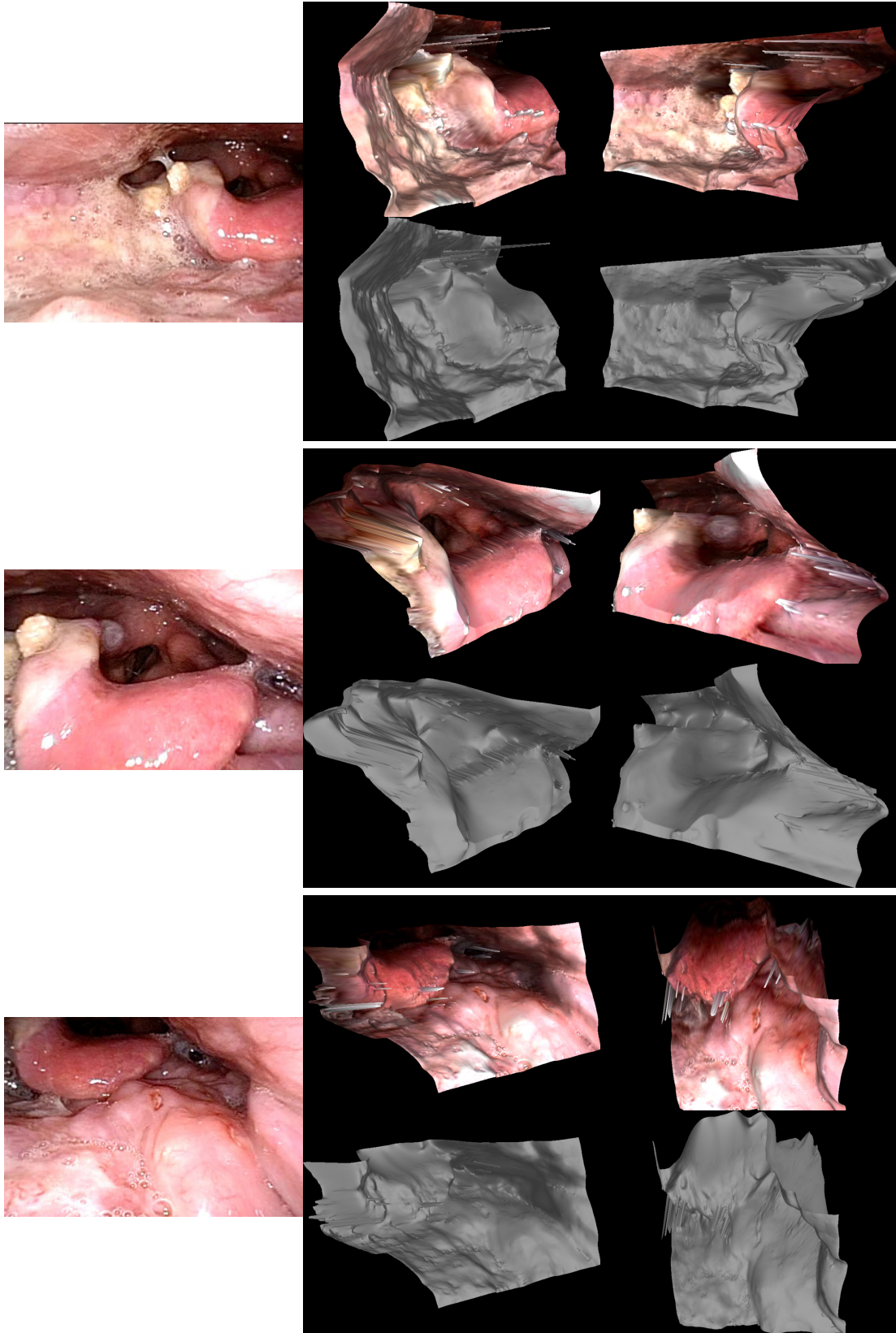


Figure 3.12: Example results for three images from a live endoscopic video. Left: Original image. Right: Surface estimated from the image using the proposed algorithm.

Limitation	Mean (Std. Dev.) Proportion of Pixels within X mm of GT			
	0.5mm	1mm	1.5mm	2mm
25 pts.	0.082 (0.042)	0.163 (0.074)	0.240 (0.101)	0.310 (0.121)
50 pts.	0.106 (0.050)	0.209 (0.087)	0.304 (0.109)	0.388 (0.122)
100 pts.	0.140 (0.051)	0.268 (0.088)	0.374 (0.109)	0.460 (0.119)
150 pts.	0.152 (0.049)	0.291 (0.082)	0.406 (0.100)	0.498 (0.105)
no deriv.	0.172 (0.048)	0.323 (0.075)	0.442 (0.084)	0.535 (0.086)
none	0.172 (0.053)	0.328 (0.082)	0.451 (0.092)	0.545 (0.090)

Table 3.3: Ablation analysis of the proposed method on ground-truth endoscopic data with $K = 2$.

Ref. Model	Mean (Std. Dev.) Proportion of Pixels within X mm of GT			
	0.5mm	1mm	1.5mm	2mm
Lamb.	0.057 (0.012)	0.114 (0.022)	0.169 (0.033)	0.220 (0.041)
$K = 1$	0.063 (0.015)	0.125 (0.026)	0.187 (0.037)	0.247 (0.047)
$K = 2$	0.057 (0.010)	0.113 (0.017)	0.166 (0.031)	0.218 (0.046)
$K = 3$	0.059 (0.011)	0.118 (0.024)	0.177 (0.040)	0.237 (0.057)
$K = 4$	0.060 (0.013)	0.122 (0.025)	0.190 (0.040)	0.262 (0.061)
$K = 5$	0.057 (0.011)	0.119 (0.023)	0.185 (0.038)	0.257 (0.058)

Table 3.4: Accuracy of the proposed SfM+SfS approach on real endoscopic video without accounting for surface interreflections.

CHAPTER 4: 3D RECONSTRUCTION OF TRANSIENT OBJECTS

Although structure-from-motion and multi-view stereo pipelines are able to achieve high-quality 3D reconstructions for many types of image collections, they are only able to reconstruct static structure, *i.e.*, surfaces that are stationary in all input images. In terms of creating a virtual representation of a *place*, this leaves a lot to be desired, since the virtual environment is completely devoid of context for how people exist within the space. If the 3D positions of dynamic objects like people and cars were able to be automatically represented within the 3D scene, this would open the door to a variety of interesting applications for both virtual reality and scene understanding. For example, 3D maps could be augmented with moving objects, VR environments of real places could include immersive human avatars, and city planners could assess large-scale motion flows for cars or pedestrians.

However, due to restrictions in spatial and temporal sampling, recovering 3D context for dynamic objects is a difficult task to accomplish in large-scale environments. This is intuitively true spatially: Assuming only visual data is available, a large number of cameras is required to observe all parts of the scene. For example, to accurately reconstruct the motion of an individual walking through a city, one would need multiple video cameras placed along every street where the person moves. Moreover, if one is interested in modeling general interactions with the environment – *i.e.* *object class* behavior, such as the typical locations where people stand when sightseeing, rather than *object instance* behavior, such as the path of a single individual – then large-scale temporal sampling is also required. Depending on the rate that objects are observed, imagery spanning hours, days, or even months may be necessary to robustly capture all possible placements of dynamic objects within the environment.

To tackle the problem of modeling *object class* behavior on a large scale, I propose to develop a new method for placing people into 3D reconstructions obtained from Internet photo-collections. Because these vast collections consist of many images taken from multiple positions across a long period of time, such imagery uniquely meets the spatial and temporal sampling requirements for large-scale scenes. However, using only still images also comes with a crucial drawback: Since images in the dataset are typically taken at least several minutes apart, we must generally assume that no two images in the collection capture the same person in the same place at the same time. As such, a successful modeling approach can leverage neither the typical triangulation methods used in rigid SfM nor the temporal correspondences used in non-rigid SfM. Complicating matters further, the scale of the scene (*e.g.*, in meters) and the ground surfaces are often difficult or impossible to recover in such imagery, which rules out the direct use of shape or ground-contact priors for 3D placement.

Considering this difficult scenario, I develop a method that takes an initial SfM model reconstructed from Internet images, plus 2D person detections in the individual images (Wei et al., 2016; Cao et al., 2017), and jointly outputs: 3D positions for the detected people, a gravity direction for the reconstruction, an estimated scale for the scene based on a height distribution prior, and a ground surface interpolated from the sparse set of 3D positions where the individuals are determined to stand. A key insight for this work is that, while exact triangulation is not possible for an individual, sufficiently large image sets are likely observe two people standing spatially *nearby*, albeit in different images and at different times. Leveraging this, I propose a new *approximate triangulation* approach that scores a scene scale hypotheses based on the number of nearby individuals found, with rough assumptions about body size, and considering visibility constraints effected by static structure. I further demonstrate how an initial scene scale estimate and individual height estimates can be refined using a height distribution prior, a local ground-plane prior, and visibility constraints. Finally, I demonstrate the potential for using these 3D person placements to recover a ground surface for the scene. To evaluate the accuracy of the approach, I quantitatively compare estimated scene scales to manually determined scales for objects in the scene with known dimensions.

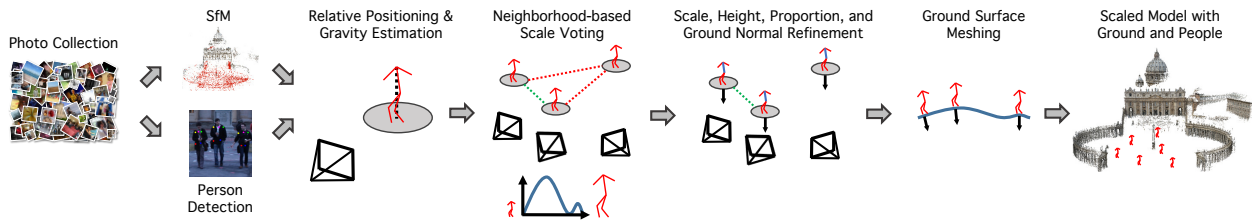


Figure 4.1: The pipeline of the proposed reconstruction system.

4.1 Approach

In this section, I present an approach for placing people, estimating scale, and recovering ground surface in a 3D scene. An overview of the pipeline is shown in Fig. 4.1. Starting from an initial set of photos of a scene, Structure-from-Motion (SfM) is first used to obtain camera parameters and sparse structure. Next, 2D torso points are detected for people in the images (Wei et al., 2016; Cao et al., 2017); these detections are used to estimate the distances and rotations of individuals relative to each camera, as well as a global scene gravity vector (Section 4.1.1). A range of possible scene scales are then tested and ranked using approximate semantic triangulation (Section 4.1.2). After this initial scale estimation, the scale and the 3D placement of the people are jointly refined using known human height statistics and encouraging a locally planar ground surface (Section 4.1.3). In the last stage, the ground surface is recovered using Poisson surface reconstruction (Kazhdan and Hoppe, 2013) (Section 4.1.4). For visualization, human avatars are placed into the 3D space with clothing colors sampled from the input images; the ground is also textured using image data and semantic pixel labelings (Yu and Koltun, 2016) (Section 4.1.5).

4.1.1 Person Detection and Gravity Estimation

The input to the algorithm consists of a set of photos of a scene, plus a sparse representation of the scene obtained from these images via SfM (Schönberger and Frahm, 2016). The first step is to detect people in the images and obtain an initial estimate of each person’s *absolute position* – that is, the real-world coordinates (in meters) of the person in the reference frame of the camera when the image was taken. These initial positions will subsequently be used for a coarse scene

scale estimation. The general approach taken here is to detect torso points in each image and, for each detection, fit a planar torso model to the detected points. The detected torsos are assumed to be aligned with the (initially unknown) gravity vector for the scene, which is a generally valid assumption given that most people stand upright (Lv et al., 2002; Krahnstoever and Mendonca, 2005; Micusik and Pajdla, 2010). A joint optimization is thus performed over three variable classes: 1) the global gravity vector, 2) the absolute position of each person’s neck point, and 3) the 1-DoF heading (rotation around the gravity vector) of the person. This optimization is done by minimizing the reprojection error of the posed torso models back into their original images.

Torso Detection: For detection, the method employs Convolutional Pose Machines (CPM) (Wei et al., 2016; Cao et al., 2017), a state-of-the-art joint detector specifically designed for real-time, multi-person pose estimation. Image-space joints on the torso are defined by taking the CPM detections for the neck, shoulders, and hips. In implementation, only joint detections having at least 30% confidence are considered, and individuals are excluded if they do not have confident detections in the neck and at least one of the hips.

Torso Model Fitting: As a coarse initialization that will later be refined, a fixed-size planar torso model is fit to each detection. This model is centered at the neck point with a width of 30cm and a height of 52cm (Fig. 4.2). By convention, gravity points in the positive y direction, so the model is defined in the xy plane.

The torso model is transformed to match the detected 2D joints for person i . Because we have obtained an initial SfM reconstruction of the scene, we know the pose $[R_i \mid \mathbf{t}_i]$ and the intrinsics of the observing camera. The camera location in the reconstruction space does not matter at this stage, but it is necessary to know the orientation of the camera relative to the gravity direction of the scene.

The model-to-camera transformation is applied in four steps. First, the model is rotated around the y axis by angle θ_i ; denote the associated rotation as $R(\theta_i)$. Second, the model is aligned to the scene gravity vector $\mathbf{g} \in \mathbb{R}^3$, with $\|\mathbf{g}\| = 1$, by calculating the rotation of the model gravity vector $[0 \ 1 \ 0]^T$ into \mathbf{g} . This rotation can be formulated as the unit quaternion $\mathbf{q}_{\mathbf{g}} = (\hat{v}_2, \hat{v}_3, 0, -\hat{v}_1)$, where $\hat{v} = \frac{v}{\|v\|}$ with $v = \mathbf{g} + [0 \ 1 \ 0]^T$; more generally, denote this model-to-world gravity alignment as

$R(\mathbf{g})$. When combined, these first two rotations represent the direction the person is facing in the gravity-aligned reconstruction space. Third, this result is placed in the coordinate frame of the observing camera by applying the extrinsic rotation matrix R_i . Finally, the model is translated relative to the camera based on the 3D position of the neck point $N_i = z_i[x_i \ y_i \ 1]^T$, where (x_i, y_i) is the 2D coordinate of the neck point in normalized camera coordinates, and z_i is the depth (in meters) of the person relative to the camera. Note, it is not required for (x_i, y_i) to exactly lie at the neck point detected by CPM.

For 3D joint J_m in the original torso model, we thus obtain a rotated, gravity-aligned, camera-aligned 3D joint:

$$J_{i,m} = R_i R(\mathbf{g}) R(\theta_i) J_m + N_i. \quad (4.1)$$

Optimization: The algorithm jointly optimizes \mathbf{g} and all individuals' poses $\Theta = \{(\theta_i, x_i, y_i, z_i)\}$ by minimizing the reprojection errors of the torso model into the original images:

$$\min_{\mathbf{g}, \Theta} \sum_i \phi \left(\sum_m \rho_{i,m}^2 \|\pi_i(J_{i,m}) - \mathbf{j}_{i,m}\|^2 \right), \quad (4.2)$$

where $\mathbf{j}_{i,m}$ is the 2D pixel location of detected joint m , $\pi_i(\cdot)$ is the projection function for camera i that converts 3D points relative to the camera into 2D pixel projections according to the camera intrinsics estimated in SfM, and $\rho_{i,m}$ is the joint detection confidence obtained from CPM. $\phi(\cdot)$ is a robust function that mitigates the effect of strong outlier detections; for implementation, the Huber loss function is employed with a threshold of 4 pixels (Huber, 1981).

The gravity vector is initialized to the geometric median of the individual camera down vectors. In order to obtain good initialization for depth, we perform a preliminary optimization of depths $\{z_i\}$ and gravity only, followed by a further optimization of all parameters. The depth and gravity optimization works as follows: Neck locations $\{(x_i, y_i)\}$ are fixed to the initially detected 2D locations, and depths are initialized to 1 meter. The rotation parameters $\{\theta_i\}$ are ignored; instead, a set of discrete rotations $\{\bar{\theta}_k\}$ is sampled at intervals of 10° . For each detection, the optimal rotation is taken as the angle in this set that minimizes the reprojection error. A modified version of Eq. (4.2)



Figure 4.2: To accurately localize 2D ground points for detected people, a planar torso model in 3D (left) is first fit to detected 2D neck, shoulder, and hip joints (middle-left). Right: Coordinate axes for the planar model.

is thus optimized:

$$\min_{\mathbf{g}, \{\theta_i\}} \sum_i \phi \left(\min_{\bar{\theta}_k} \sum_m \rho_{i,m}^2 \|\pi_i(J_{i,m}(\bar{\theta}_k)) - \mathbf{j}_{i,m}\|^2 \right), \quad (4.3)$$

where $J_{i,m}(\bar{\theta}_k) = R_i R(\mathbf{g}) R(\bar{\theta}_k) J_m + N_i$.

After this first optimization, $\{\theta_i\}$ values are initialized based on the value of $\bar{\theta}_k$ that minimizes the reprojection error for each person. The full set of parameters (\mathbf{g}, Θ) is then optimized using Eq. (4.2). Finally, the 3D reconstruction is re-oriented such that the estimated gravity vector is aligned with the positive y axis.

4.1.2 Voting-based Scale Estimation

At this point, we have obtained an initial absolute depth estimate for each person relative to the camera that observes them. Next, the method estimates an initial placement of the detections into the reconstruction space, while at the same time obtaining an initial absolute scale estimate for the scene. If the scene scale s (e.g. the length of 1 meter in the reconstruction space) were known, the

3D neck point of person i in the reconstruction space could be calculated as

$$P_i(s) = sR_i^T N_i + C_i, \quad (4.4)$$

where $N_i \in \mathbb{R}^3$ is the estimated 3D position of the neck point relative to the observing camera, $R_i \in \mathbb{R}^{3 \times 3}$ is the scene-to-camera rotation matrix, and $C_i \in \mathbb{R}^3$ is the 3D position of the camera in the reconstruction space.

In principle, s could be determined from a known absolute distance between two points in the reconstruction space, *e.g.*, the width of a building or the distance between two cameras. Alternatively, if the cameras were synchronized, an individual could be triangulated from detections in multiple views, and the scale could be chosen as that which best matches this 3D point. Lacking known distances, I propose to instead leverage *approximate semantic triangulation*. The idea here is that, given enough input images, and especially in well-traveled areas, there is a high probability that at least two individuals in different images will be observed in nearby locations, and at similar heights above the ground. The method samples a range of scale hypotheses for the 3D reconstruction and scores each based on the observed person correspondences.

Pairwise Approximate Triangulation: More explicitly, consider the 3D neck placements $P_i(s)$ and $P_j(s)$ (Eq. (4.4)) for two individuals at some scene scale s . Recall that, by convention, the y axis defines the vertical span of the scene, and the xz plane defines the horizontal space. Two individuals are identified as standing “nearby” if they are within some fixed absolute distance τ_{xz} in the horizontal space. In addition, say that the individuals are standing at similar heights if their neck points are within some fixed absolute distance τ_y in the vertical space. Taking $\Delta P_{ij}(s) = P_i(s) - P_j(s)$, let $M_{ij}(s)$ denote the binary indicator function that determines whether persons i and j are approximately triangulated at scale s :

$$M_{ij}(s) = (|\Delta P_{ij}^{xz}(s)| < s\tau_{xz}) \wedge (|\Delta P_{ij}^y(s)| < s\tau_y), \quad (4.5)$$

where $|\Delta P_{ij}^{xz}(s)|$ and $|\Delta P_{ij}^y(s)|$ denotes the horizontal and vertical distances between the neck points, respectively.

The value $M_{ij}(s)$ is computed for all pairs of detected people in separate images. An individual is successfully triangulated at scale s if any pairwise approximate triangulation was successful, and if they satisfy a visibility constraint ($V_i(s)$, explained below):

$$M_i(s) = V_i(s) \wedge \left(\bigvee_j (\mathcal{I}_i \neq \mathcal{I}_j) \wedge V_j(s) \wedge M_{ij}(s) \right), \quad (4.6)$$

where \mathcal{I}_i denotes the image in which person i was detected.

Visibility Constraint: An important constraint in the scale estimation is that the line segment from C_i to $P_i(s)$ should not intersect with structures such as walls. This constraint may be violated if s is too large, which pushes $P_i(s)$ further from the observing camera. Accordingly, $V_i(s)$ is an indicator function denoting whether the detection of person i is possible at scale s given the free space of the static parts of the scene. In practice, $V_i(s)$ is computed by voxelizing the SfM 3D point cloud with a fixed voxel size of one meter (s units in the reconstruction space). Ray-tracing is then performed from C_i along ray $R_i^T N_i$ to compute the first point of intersection with a filled voxel. Denote the distance from C_i to this voxel as $v_i(s)$. $V_i(s)$ is then defined as

$$V_i(s) = s \|N_i\| < v_i(s). \quad (4.7)$$

Scale Scoring: A hypothesized scale s is scored by taking a weighted aggregate of all $M_i(s)$:

$$S(s) = \sum_i w_i M_i(s). \quad (4.8)$$

Setting $w_i = 1$ is equivalent to counting the successfully triangulated individuals at scale s . I have experimentally found slightly better performance by weighting individuals by the number of detections in their associated image, *i.e.*, $w_i = 1/N_{\mathcal{I}_i}$, where $N_{\mathcal{I}_i}$ is the total number of detections

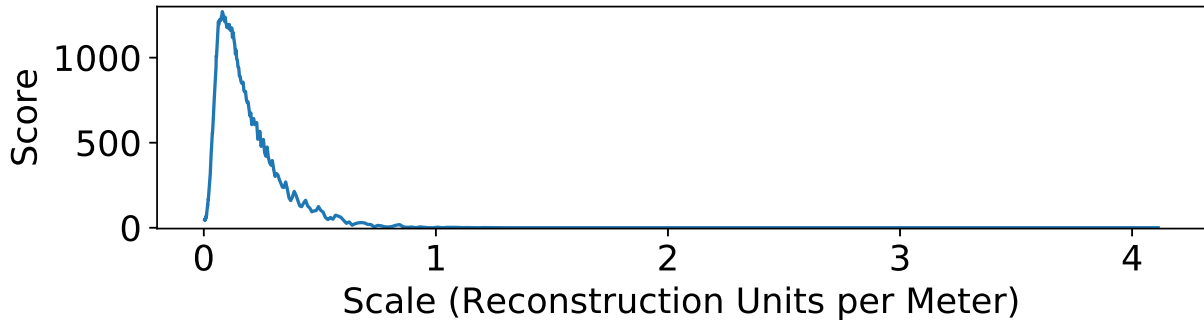


Figure 4.3: Scale scoring curve for a model of the Pantheon. The peak is chosen as the initial scale estimate.

in image \mathcal{I}_i . This weighting mitigates the ambiguity of person placement in crowded areas, where incorrect scales can still yield valid triangulations due to the overall person density.

Finally, an initial voting-based estimate of the scene scale is obtained by sampling a range of possible scales and selecting the scale hypothesis with the highest score $S(s)$. For purposes of implementation, this range is generated by assuming that the vertical span of the SfM point cloud is between 1 and 1000 meters. The method starts at the smallest possible scale and test all scales in the range, stepping at 2% increments in s . Also at this stage, the approach only considers individuals having all five torso joints detected with at least 30% confidence. I use absolute horizontal and vertical thresholds of $\tau_{xz} = 1.5\text{m}$ and $\tau_y = 0.1\text{m}$; an example scoring curve using these parameters is shown in Fig. 4.3. In practice, I have found that the voting approach is not too sensitive to the value of these parameters – the “nearby” and “similar height” heuristics should merely reflect how a pedestrian might characterize these terms for someone passing them on the street. Besides, the main point of this stage is to obtain an approximate initial scene scale, and I show in my experiments that the method can tolerate an initial scale error of at least 15%.

4.1.3 Scale Refinement, Height Estimation, and Ground Surface Estimation

Having obtained an initial scale estimate s , the algorithm next jointly refines this scale, estimates a height h_i in meters for each detected individual, and estimates a ground surface unit normal $\mathbf{n}_i \in S^2$ for the ground point at which each individual stands. As part of this optimization, a

torso height $t_i = \beta_i h_i$ is also estimated for each person, where β_i is the individual's torso-to-height proportion. In the following, I first formulate how to obtain a person's 3D position in the reconstruction space given s , h_i , and β_i . I then introduce the three terms of the joint optimization function and finally address the overall formulation.

Position as a Function of Height and Proportion: While Eq. (4.4) is convenient for an initial neck point placement, it relies on a fixed torso size. This can be generalized by allowing the torso height t_i to vary as a fraction β_i of the person's height h_i . The end result is that an increase or decrease in torso size accordingly affects the distance of the neck point N_i in Eq. (4.4) to the camera.

Let $\mathbf{r}_i = N_i / \|N_i\|$ denote the ray from the origin through the neck point of the fitted torso model in the reference frame of the camera. Moreover, let \mathbf{h}_i be the ray for the hip midpoint of the model. For every 3D point falling on \mathbf{r}_i , there is an associated point on \mathbf{h}_i that falls directly below it along the gravity direction (y axis). Again assuming that the torso aligns with the gravity vector, we can find such a neck/hip point pair for any torso height t_i . By similar triangles, we can determine a new neck point $N_i(t_i) = \rho_i t_i \mathbf{r}_i$ for any torso height, where ρ_i is the ratio between neck-point-to-camera distance and torso height.

In practice, an understanding of human proportions can be explicitly encoded in this formulation by expressing torso height as a percentage of total height, *i.e.*, $t_i = \beta_i h_i$. Eq. (4.4) can thus be updated to express a person's 3D neck point in the reconstruction space (at scale s) as a function of height and proportion:

$$P_i(s, h_i, \beta_i) = s R_i^T N_i(t_i) + C_i = \rho_i \beta_i h_i R_i^T \mathbf{r}_i + C_i, \quad (4.9)$$

Photographers can also be included in the optimization in Eq. (4.14). However, since the torsos for photographers are not directly observed, they must be treated slightly differently. Specifically, assume that the camera center is $h_c/8$ meters above the neck point for photographer c . Accordingly, $\mathbf{r}_c = [0 \ 1 \ 0]$, and fixed values $\rho_c = 1$ and $\beta_c = 1/8$ can be adopted.

Ground Point Position: The ground point $G_i(s, h_i)$ lies vertically below the neck point $P_i(s, h_i, \beta_i)$. With the neck height being a fraction η of the total height of the person, the ground point in reconstruction space is given as

$$G_i(s, h_i, \beta_i) = P_i(s, h_i, \beta_i) + [0 \ s\eta h_i \ 0]^T. \quad (4.10)$$

I propose to use a fixed value of $\eta = 5/6$, reasoning that the top of the sternum (that is, the assumed neck point) is slightly less than two head lengths from the top of a person, and that human head length is approximately one-eighth of total height (Bogin and Varela-Silva, 2010).

Optimization Overview: As previously mentioned, the scene scale s is optimized along with the set $\{(h_i, \beta_i, \mathbf{n}_i)\}$ of per-person heights, proportions, and ground normals. The objective function for this optimization has three terms: 1) a prior on height, 2) a local ground planarity term for pairs of nearby people, and 3) a visibility constraint.

Height Distribution Prior: I propose to leverage the known distribution of human heights as a prior on the estimated height h_i for each person. Here, I employ a Gaussian mixture model (GMM) for this distribution; in principle, any GMM or otherwise appropriate probability distribution could be used. The GMM probability function is given as the sum of probabilities for K separate Gaussians:

$$p(h_i) = \sum_{k=1}^K \frac{\alpha_k}{\sigma_k \sqrt{2\pi}} \exp\left(-\frac{(h_i - \mu_k)^2}{2\sigma_k^2}\right), \quad (4.11)$$

I use a general two-component GMM for male and female adult heights, respectively: $\{(\alpha_k, \mu_k, \sigma_k)\} = [(0.504, 1.768, 0.068), (0.496, 1.646, 0.060)]$, which was aggregated from several sources (Garcia and Quintana-Domeque, 2007; Subramanian et al., 2011; The World Bank Group, 2017). In principle, more detailed models could be used, such as a model that captures factors of age or ethnicity.

Local Planarity Prior: The second objective term encourages the ground surface between nearby people to be relatively smooth (but not necessarily horizontal). This is enforced by endowing each individual with a ground normal \mathbf{n}_i that defines a planar ground patch around the point at

which they stand. Nearby ground points that are far from this flat surface receive a penalty. For two individuals i and j , the point-to-plane distance in meters between $G_j(s, h_j, \beta_j)$ and the patch for person i is given as

$$d_{ij} = \frac{1}{s} \left| (G_j(s, h_j, \beta_j) - G_i(s, h_i, \beta_i))^T \mathbf{n}_i \right|. \quad (4.12)$$

This distance is penalized with a squared loss during optimization.

Visibility Constraint: The optimization again seeks to penalize scales and heights that push neck points into or beyond static parts of the scene. To do this, the static scene voxelization is computed at the initial scale s_0 and then used to compute $v_i(s_0)$. (Note that this is the distance in reconstruction units to the nearest static surface for the detection ray; it is expressed as a function of scale only to clarify that the voxelization occurs at a fixed scale.) Dropping the dependence on initial scale, these maximum distances $\{v_i\}$ are fixed during the optimization. The visibility penalty term is close to zero for neck-to-camera distances much less than v_i and close to one for values much greater:

$$\nu_i(s, h_i, \beta_i) = \frac{1}{\pi \tan^{-1}(2)} \tan^{-1} \left(\frac{2}{\tau_o} \left(\|N_i(t_i)\| - \frac{v_i}{s} \right) \right), \quad (4.13)$$

where $\|N_i(t_i)\| = \varrho_i \beta_i h_i$ is the neck-to-camera distance in meters, and τ_o is a value in meters such that an “overshooting” of $3\tau_o$ meters results in a penalty of approximately 0.95. In my experiments, I use $\tau_o = 0.2\text{m}$.

Optimization: Eqs. (4.11-4.13) are combined into a single objective function to be minimized:

$$E(s, \{(h_i, \beta_i, \mathbf{n}_i)\}) = -\frac{1}{D} \sum_{i=1}^D \log p_i(h_i) + \frac{1}{4|\mathcal{N}|\lambda^2} \sum_{(i,j) \in \mathcal{N}} (d_{ij}^2 + d_{ji}^2) + \frac{1}{D} \sum_{i=1}^D \nu_i(s, h_i, \beta_i), \quad (4.14)$$

where D is the total number of detected people, \mathcal{N} is a set of person neighbors to which the local planarity prior is applied bidirectionally, and λ is a weight for the planarity penalty. The first term is derived by taking the negative log-likelihood of the height probability. In my experiments, I set $\lambda = 0.02$, which roughly reflects an expected ground plane noise of 2cm. The neighborhood structure \mathcal{N} is defined based on the initial person placements at s_0 . The “nearby” constraint can

loosened in this stage, since we have at least a rough estimate of the scale of the scene. Specifically, nearby initial placements are identified as those having neck points within 3m of each other in the horizontal space and 0.242m in the vertical space. Under the adopted height distribution, this vertical limit is the 95% threshold for the height difference between randomly chosen height pairs – *i.e.*, when considering all pairs of people in a population, only 5% are expected to have a difference in height greater than this value.

The values for $\{\beta_i\}$ are constrained to the range $[0.25, 0.45]$, which reasonably captures the range of human torso proportions (Bogin and Varela-Silva, 2010), and these values are initialized to 0.3 for optimization. Individual heights are randomly initialized by sampling from the height distribution model. Normals are parameterized by spherical coordinates and are initialized with small random perturbations. At this stage, person detections having at least four detected joints are also included for optimization.

4.1.4 Ground Surface Reconstruction

Using the optimized 3D ground points and ground point normals, a ground surface is recovered using the Poisson surface reconstruction (PSR) implementation of Kazhdan and Hoppe (Kazhdan and Hoppe, 2013). PSR produces a high-quality mesh with adaptive resolution from an input set of oriented points, which in this case is defined by $\{(G_i(s, h_i, \beta_i), \mathbf{n}_i)\}$. Prior to running PSR, the input point cloud is filtered by removing individuals who are more than 40m from their observing camera or who fail the visibility constraint at the optimized scale s . Small, far-off groups of photographers are also removed.

4.1.5 Visualization

To demonstrate the potential of this method for scene completion, the recovered ground surface is textured, and a subset of all detected people are placed into the reconstruction space. The person visualization consists of a low-poly model for each detection, with the shirt and pants colored by sampling the original image. Each person model is scaled to match the estimated height for the

detection. To portray a realistic spatial distribution of pedestrians, a random subset of individuals is selected for visualization. This is achieved by treating the selection as a set cover problem and taking a greedy approach. Specifically, for each photographer c , denote \mathcal{O}_c as the set of people observed in the image taken by that cameraperson, and $\mathcal{V}_c \supseteq \mathcal{O}_c$ as the set of all individuals (including photographers) placed within the viewing frustum of the photographer’s camera, up to some maximum depth. The algorithm selects a random photographer and marks all individuals in \mathcal{V}_c as “visited.” At the same time, all individuals in \mathcal{O}_c who were not previously marked as visited are placed into the reconstruction; if any such person exists, the photographer is also placed into the scene. Photographers are randomly and iteratively selected in this fashion until all people are marked as visited.

The ground surface obtained from PSR has accurate geometry but lacks color. To texture this surface, each vertex on the ground surface mesh is projected into each individual image registered in the 3D reconstruction and, if the projection lies within the image boundaries, the color value is sampled at the pixel in which it falls. The sampled colors are aggregated over all images, and the median color is computed for each vertex. To avoid sampling non-ground pixels (caused by, *e.g.*, occluding scene geometry or pedestrians), the texturing process utilizes dense pixel-wise semantic labeling. For each image, the convolutional neural network of (Yu and Koltun, 2016), trained on the Cityscapes dataset (Cordts et al., 2016), is applied to obtain a most-probable class labeling for each pixel. When aggregating color values, sampled pixels are ignored if they are not identified as ground, sidewalk, or terrain.

4.2 Evaluation

The proposed method has been applied to several several scenes from large-scale image photo-collections (Li et al., 2010; Cao and Snavely, 2012; Wilson and Snavely, 2014; Heinly et al., 2015), as well as the well-known Cornell Arts Quad dataset (Crandall et al., 2011). Evaluation in the context of unordered Internet photo-collections is challenging for the tasks of placing people and estimating ground surfaces due to the lack of available ground truth. However, the estimate of the

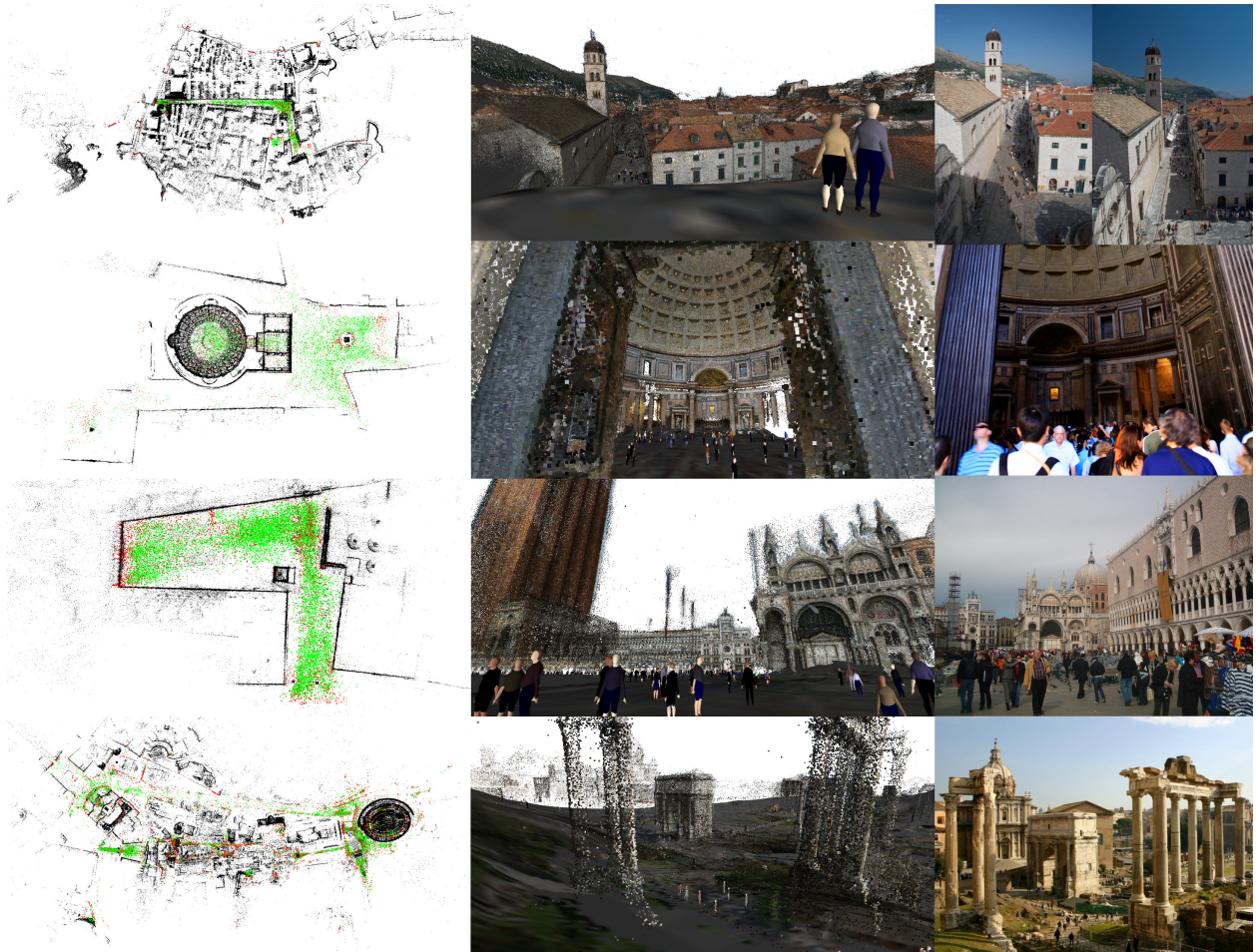


Figure 4.4: Overhead views (left) and sample renderings with ground and person avatars (middle) for the proposed method. Examples of real photos are shown on the right. The green dots in the overhead views show person placements, with cameras as red dots and detected people as green dots. Black dots show static structure. From top: Dubrovnik, Croatia; the Pantheon; San Marco Plaza, Venice; and the area around the Colosseum and Roman Forum in Rome.

scene scale, which is integral to the overall pipeline, can be evaluated quantitatively. To perform this assessment, the ground-truth scene scale was manually obtained for a number of reconstructions by taking the known sizes of structures in each scene and comparing them to their size in the reconstruction space. The scale evaluation results are shown in Table 4.1. It can be seen that the proposed scale estimation via a height distribution prior reliably determines the scene scale. Effectively, the method uses object semantics to overcome the inherent scale ambiguity of SfM reconstructions.

The gravity vector for each scene, which is estimated during the initial torso fitting stage, can also be quantitatively compared to other approaches. When compared to the implementation of Schönberger and Frahm (2016) that performs automatic gravity vector estimation from scene vanishing points, the torso-based approach has an average difference of 1.078° over the datasets of Wilson and Snavely (2014). This indicates that the gravity vector estimation of the proposed approach has very similar performance to other methods. One (slightly contrived) situation where the torso method might be preferred is when the dominant scene lines strongly deviate from Manhattan world assumptions, which would cause vanishing-point estimation methods to fail entirely. Of course, the input imagery would need to contain a sufficient number of people without too many instances where the torso was not relatively upright.

For qualitative analysis, sample overhead and ground-level visualizations are shown in Fig. 4.4 on four large-scale datasets: Dubrovnik, the Pantheon, San Marco Plaza, and the Campitelli in Rome. Fig. 4.5 shows additional overhead visualizations for several other scenes, along with a comparative aerial image from Google Earth. For the overhead visualizations, green dots show the placement of detected individuals, red dots show locations for photographers, and black dots show static scene structure. In general, the placements for detected people into the scene reflect the actual structures where people walk, particularly along sidewalks. Places where people do not walk (*e.g.*, the fountains in Trafalgar Square) contain low densities of (likely mis-detected) people. The accurate scale estimates presented in the paper and above provide additional evidence as to the correctness of these placements. There were failure cases on other scenes, such as the Statue

of Liberty (not shown), that were primarily caused by a large number of false person detections on human-like statues. These false detections are also visible in the water of the Trevi Fountain, below; however, the scene conditions in that case did not appear to negatively influence the result. I have also empirically found that the method’s accuracy is generally higher in scenes having 1) a larger number of person detections and 2) more complete static reconstructions obtained via Structure-from-Motion. The former condition provides greater support for approximate semantic triangulation, while the latter is important for enforcing visibility constraints, which are helpful in avoiding under-estimation of the length of one unit in the reconstruction space.

Scene	Error	n_p	n_c
Alamo	+0.3%	1940	699
Brandenburg Gate	-7.2%	5115	1131
British Museum	+0.3%	2925	507
Buckingham Palace	-5.9%	4972	1257
Campitelli	+1.9%	15836	16834
Cornell Quad	-4.0%	550	4773
Dubrovnik	-0.15%	5066	2714
Hōzōmon Temple, Tokyo	-1.5%	1768	230
Lincoln Memorial	+6.5%	875	183
New York City Library	-1.4%	466	480
Palace of Westminster	-8.8%	331	496
Pantheon	+4.3%	8656	3310
Piccadilly Square	-6.1%	7908	2453
Pike Place Market, Seattle	+8.5%	1081	312
Sacré Cœur, Paris	-0.3%	1705	782
San Marco	-0.3%	15712	4916
Taj Mahal	-1.1%	395	805
Tōdai-ji Temple, Nara	-2.1%	2419	733
Tower Bridge, London	-2.6%	213	125
Tower of London	-4.7%	551	381
Trafalgar Square	+3.2%	13306	4328
Trevi Fountain	-3.3%	4934	2343
Union Square Park, NYC	-4.5%	2833	1023

Table 4.1: Quantitative results on the proposed method for scale and placement. “% Error” gives the amount that the method over/under-estimated the distance of one unit in the reconstruction. n_p and n_c show the number of placed detected people and photographers, respectively, recovered by the method.

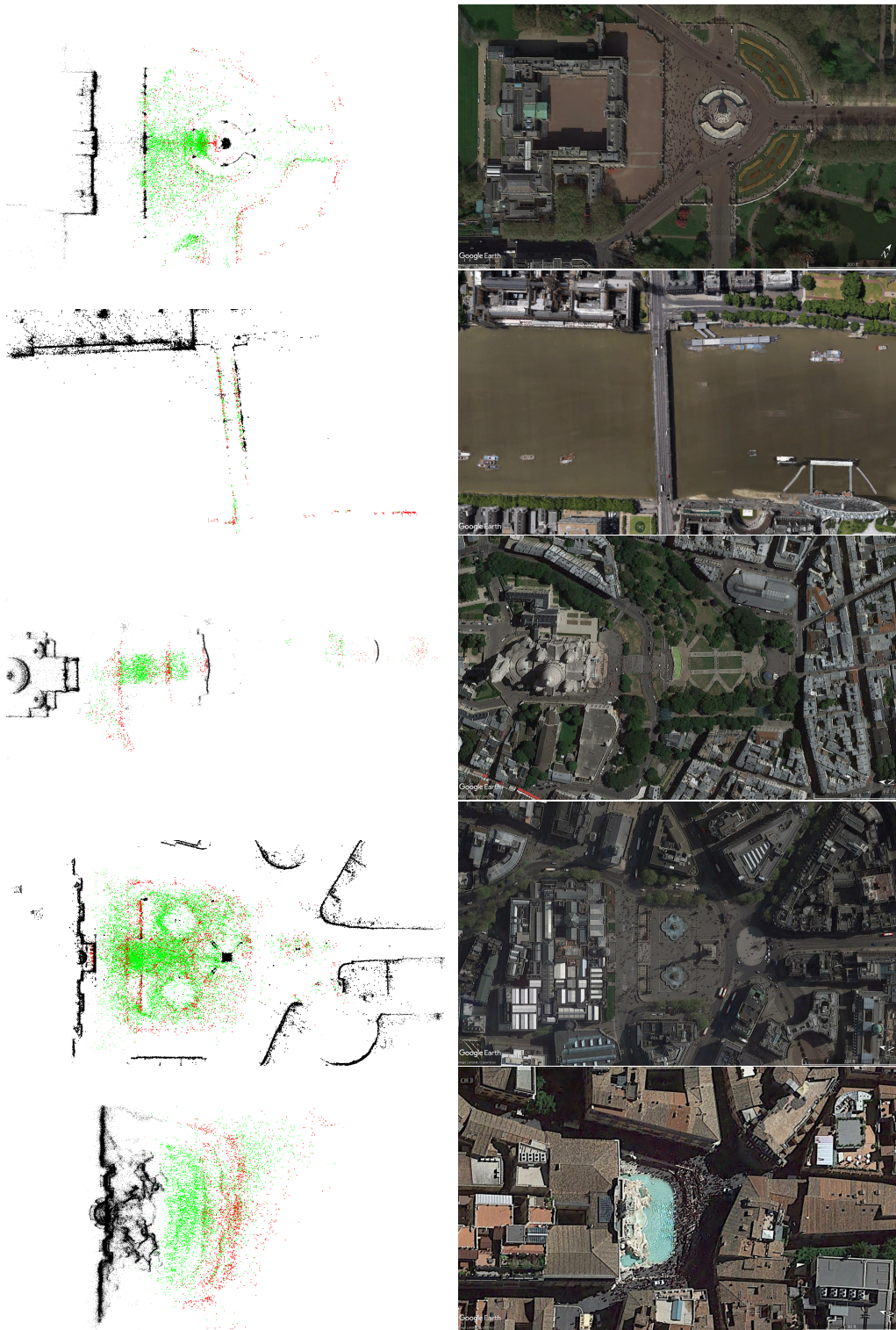


Figure 4.5: Overhead visualizations of person placements (left) versus aerial views from Google Earth (right). From top to bottom: Buckingham Palace, the Palace of Westminster, the Sacré Cœur in Paris, Trafalgar Square, and Trevi Fountain.

4.3 Ablative Analysis

This section provides additional analysis on the various parts of the reconstruction pipeline. Recall that the algorithm has two general stages: scale voting and scale refinement. The scale voting stage serves to initialize the subsequent refinement. The following subsections demonstrate that both stages are necessary to produce a satisfactory result, and it is also demonstrated how different parameter selections affect the end result in both stages.

4.3.1 Visibility Constraint During Voting

First, I analyze the effect of removing the visibility constraint (Eq. (4.7)) during the scale voting procedure. The visibility constraint is necessary at this stage, but using the constraint alone is not sufficient to obtain the scene scale. Fig. 4.6 shows the effect of turning off the constraint for the Campitelli model. Because the (model-space) neighborhood radius in Eq. (4.5) generally grows faster with respect to scale than pairwise person distances, using the neighborhood term alone will result in artificially high overlap at larger scales. The visibility constraint is thus important to rule out impossible person placements.

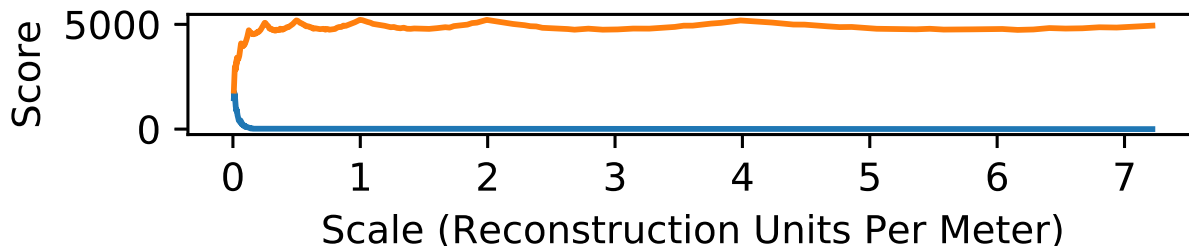


Figure 4.6: Result of the scale voting scheme with (blue) and without (orange) the visibility constraint. The ground-truth scale is near 0.01 reconstruction units per meter.

Fig. 4.7 demonstrates that the visibility constraint alone is not sufficient for determining the scene scale. For each detection in the Campitelli model, I compute the ratio of the estimated neck distance $s||N_i||$ to the visibility threshold $v_i(s)$ (cf. Eq. (4.7)) for the ground-truth scene scale, and for $\pm 10\%$ and $\pm 20\%$ of this scale. These ratios are sorted across all individuals and plotted. At the correct scale, individuals adjacent to static structures will have a ratio of ~ 1 . With perfect detections,

this principle could conceivably be used to estimate the correct scale. If all neck distances and detections were correct, and assuming at least one person stands exactly next to, *e.g.*, a wall, we could conceivably select the (approximately) correct scale based on this principle. It is clear from the figure, however, that false detections and mis-estimations of the neck distance (having ratios much greater than 1) make this threshold ambiguous. The proposed voting-based approximate triangulation approach is thus necessary to robustly obtain an initial scene scale.

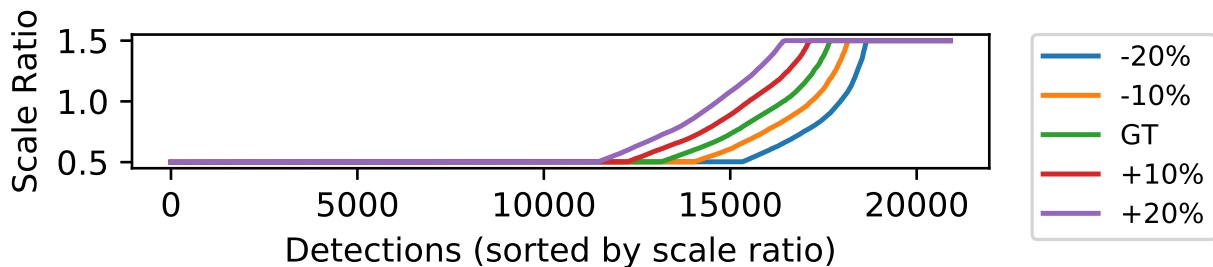


Figure 4.7: Ratio of the estimated neck distance $s||N_i||$ to the visibility threshold $v_i(s)$ for the ground-truth scale (GT), and for larger/smaller scales. Values are sorted and clipped to $[0.5, 1.5]$.

4.3.2 Effect of Scale Refinement Terms

There are three optimization terms in the scale refinement stage of the proposed algorithm (Eq. (4.14)): a height prior, a local planarity penalty, and a visibility constraint. The algorithm requires the local planarity term – without it, the optimal solution is to set the scale to an infinitesimal positive value (maximizing Eq. (4.13)) and each h_i to the most probable height according to the prior distribution. Table 4.2 shows the estimated scales with the height and visibility terms removed. The effect of the height prior varies between datasets, but in general better scale estimates are recovered when the constraint is included. The visibility constraint is intended for scenes with fewer individuals, to help prevent scale over-estimation caused by fewer well-supported neighborhoods.

4.3.3 Effect of Parameters during Refinement

To investigate the sensitivity of the algorithm to parameter changes, Table 4.2 further shows results after modifying the four major tunable parameters of the refinement (photographer camera

height β_c , “overshooting” threshold τ_o , planarity penalty λ , and the xz neighbor threshold) by $\pm 10\%$. The relative scale differences are generally small, and only minor changes are observed in the estimated 3D positions of the detected individuals.

4.3.4 Comparing Scale Voting and Scale Refinement

Finally, the 3rd and 4th columns of Table 4.2 show the scale improvement of the refinement stage versus the initial voting. For many datasets, the refined scale estimate is closer to the ground truth. Since the local planarity term is the driving factor in the refinement step, this result supports the notion that the person placement (including the initial 3D triangulation) is an important component of the approach.

4.4 Discussion

This chapter introduced a new approach for adding transient elements to large-scale static 3D reconstructions, operating under the difficult scenario of having minimal prior knowledge about the scene. Specifically, the method leverages recent advances in image-based person detection, along with population height distribution priors, to jointly place detected people into the scene, estimate the absolute scale of the reconstruction, recover the gravity vector of the scene, and recover the underlying ground surface. The method has been tested on a large collection of real-world datasets, and quantitative and qualitative results demonstrate the significant advances of approach in modeling hard-to-capture scene elements. One key insight of this work is that knowledge of object class properties, such as height distribution in humans, can provide adequate constraints on 3D placement even when exact correspondence is impossible.

Scene	GT	Initial	Final	No Height	No Visib.	-10%	+10%
Cornell Quad	0.0269	0.0259	0.0280	0.0278	0.0294	0.0282	0.0272
Dubrovnik	0.0200	0.0183	0.0200	0.0199	0.0195	0.0197	0.0198
Pantheon	0.0913	0.0799	0.0873	0.0912	0.0877	0.0877	0.0874
Campitelli	0.0104	0.0097	0.0102	0.0102	0.0103	0.0102	0.0102
San Marco	0.0379	0.0336	0.0380	0.0375	0.0367	0.0383	0.0385
Alamo	0.1350	0.1253	0.1346	0.1323	0.1363	0.1320	0.1351
NYC Library	0.1437	0.1262	0.1418	0.1553	0.1442	0.1429	0.1403
Piccadilly	0.1216	0.1263	0.1290	0.1442	0.1329	0.1289	0.1275
Brandenburg Gate	0.1266	0.1287	0.1365	0.1433	0.1369	0.1356	0.1356
British Museum	0.3913	0.2793	0.3900	0.3434	0.4014	0.3923	0.3877
Buckingham Palace	0.0629	0.0604	0.0668	0.0776	0.0663	0.0662	0.0658
Hōzōmon Temple	0.5651	0.5070	0.5739	0.4642	0.5941	0.5797	0.5689
Lincoln Memorial	0.1161	0.1086	0.1090	0.1217	0.1093	0.1100	0.1080
Palace of Westmin.	0.0259	0.0280	0.0284	0.0298	0.0287	0.0289	0.0296
Pike Place Market	0.1840	0.1314	0.1696	0.1462	0.1754	0.1678	0.1704
Sacré Cœur	0.0507	0.0477	0.0509	0.0512	0.0503	0.0499	0.0502
Taj Mahal	0.0475	0.0420	0.0481	0.0488	0.0497	0.0491	0.0475
Tōdai-ji Temple	0.1340	0.1251	0.1369	0.1563	0.1380	0.1369	0.1354
Tower Bridge	0.2166	0.2391	0.2223	0.2391	0.2238	0.2244	0.2205
Tower of London	0.0484	0.0479	0.0507	0.0497	0.0517	0.0513	0.0498
Trafalgar Square	0.0700	0.0628	0.0678	0.0679	0.0673	0.0700	0.0671
Trevi Fountain	0.3179	0.2538	0.3288	0.3571	0.3278	0.3335	0.3213
Union Square	0.1380	0.1276	0.1430	0.1568	0.1427	0.1422	0.1416

Table 4.2: Ablative analysis on the importance of different parts of the proposed algorithm. GT: Ground-truth scene scales (reconstruction units per meter). Initial/Final: Estimates from the voting and refinement stages. No Height/Visib.: Height/visibility terms removed from final optimization. ± 10 : With all parameters modified by ten percent. Red cells: Results where the estimated length of one unit in the reconstruction was incorrect by $>10\%$.

CHAPTER 5: LIVING 3D RECONSTRUCTIONS

In the previous chapter, I introduced a method for recovering the 3D positions of people in large-scale 3D reconstructions from unordered Internet photo-collections, as well as the scale of the scene. Given this augmented reconstruction, two questions come to mind: 1) How can we use this information to improve the overall reconstruction? 2) Can we visualize the reconstructed scene in ways that make it appear more true to life? In this chapter, I introduce a pipeline that tackles both questions.

Regarding the first question, a key reconstruction aspect that can be improved is proper modeling of ground surfaces. The ground surfaces used for visualization in the previous chapter were reconstructed using a straightforward meshing technique — Poisson surface reconstruction applied to the oriented point cloud of person ground points — that did not integrate with the static structure of the scene in any way. To achieve a better scene representation, a more holistic reconstruction of both the person ground points and MVS depthmaps is desired. This is still a difficult problem, since ground points will generally only be available for the areas of the scene where people actually walk. Surface reconstruction approaches that aim to model weakly supported surfaces can recover the ground to some degree, but they are still limited in the sense that they require *some* level of support. For example, Fig. 5.1 shows scene reconstruction results from images of the Castel Sant’Angelo. The left column shows the result of a surface reconstruction approach that applies Delaunay tetrahedralization to a point cloud fused from MVS depth maps; the mesh surface is extracted by labeling the tetrahedra as “inside” or “outside” using graph-cut optimization and visibility-ray-based cost terms (Labatut et al., 2009; Jancosek and Pajdla, 2011; Schönberger et al., 2016). In areas with insufficient ground coverage in the 3D point cloud, the optimization has no signal to drive a correct ground surface recovery, and significant holes can appear in the mesh.

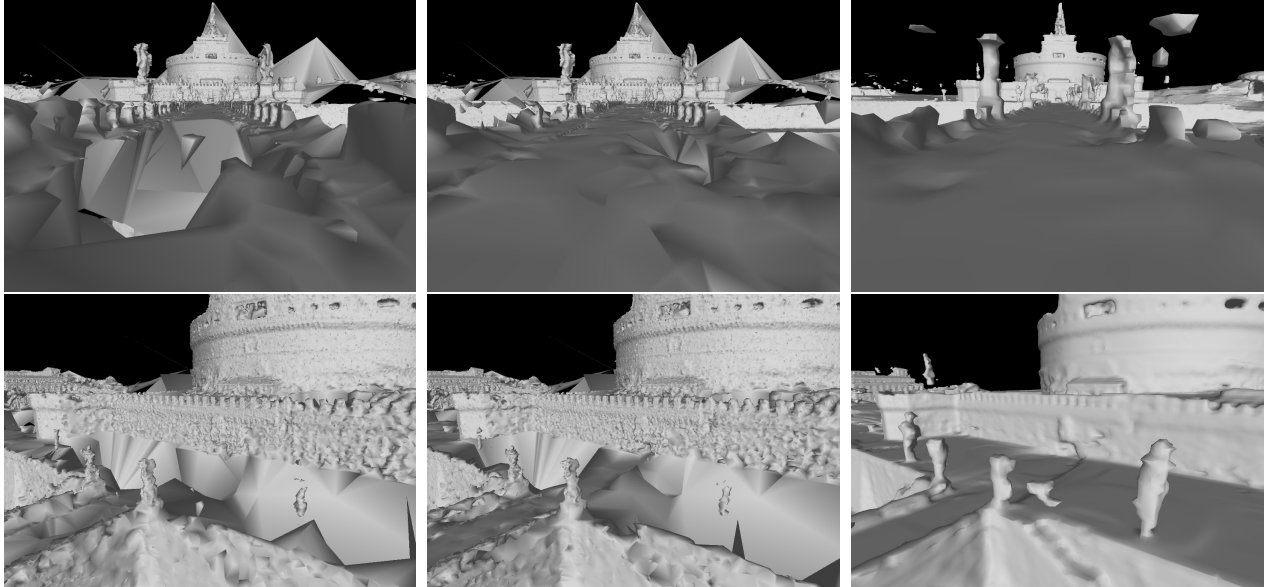


Figure 5.1: Left: A scene mesh generated by point cloud fusion from MVS depthmaps and Delaunay tetrahedralization with visibility optimization (Schönberger et al., 2016). Middle: Mesh for the same approach, with 3D ground points of detected people added to the point cloud. Right: Scene mesh generated by the proposed method.

In the middle column of Fig. 5.1, I have added the recovered 3D person ground points to the MVS-recovered point cloud. This helps the ground recovery for areas where people walk, but significant holes still exist outside of these regions (see especially the bottom image in the middle column, where the path from the bridge to the entrance is well covered, but the other ground areas near the building are not).

Obtaining the scale of the scene and gravity alignment provides an alternative route for recovering ground surfaces. The gravity direction provides constraints on vertical aspects of the scene (the bottom of the reconstruction should be below ground and the top of the reconstruction above it), and for regions where the ground is poorly recovered, we can assume that it is mainly flat, especially for areas where people might walk. Knowing the scale of the scene allows us for fiducial limits to be applied to the modeling problem. That is, an exact resolution for the reconstructed geometry can be specified, facilitating the use of volumetric approaches and removing the need for methods like Delaunay tetrahedralization that do not carry guarantees that the precise surface can necessarily be

extracted from the ad hoc topology. When scale and gravity are combined with 3D ground points for detected people, more complete scene representations can be obtained (Fig. 5.1, right).

Ground surface reconstruction also is a necessary component for the second problem of whether reconstructions can be visualized more as a *place* and less as (for want of a better term) a 3D model of a place. To this end, I introduce the concept of *living 3D reconstructions*: models of real places that include dynamic representations of transient objects, all recovered automatically from crowd-sourced imagery. To demonstrate this in the setting of modeling people, I introduce a way of bringing 3D reconstructions “to life” via crowd simulation. Given 3D positions of people in the input imagery and a reconstruction of the scene that includes a complete ground surface, I first extract “walkable” ground regions of the mesh, including sidewalks and other regions where people are commonly found. Walking virtual pedestrians are then rendered into the scene and then, using crowd simulation software (Curtis et al., 2016), are made to navigate between different positions in the ground mesh where actual people were determined to have stood. The resulting 3D scenes carry the immersion that comes with observing people moving around in an environment, including the context of the size of buildings and — importantly — how people dynamically exist within the space. These living models are a compelling first step towards new visualization approaches in large-scale virtual tourism.

5.1 Robust Surface Reconstruction

In this section, I outline a volumetric approach for recovering a complete surface mesh for a scene given a set of depthmaps obtained using MVS and 3D ground points for the people detected in the input imagery.

5.1.1 Truncated Signed Distance Function Aggregation

As mentioned previously, knowing the scale of the scene allows for reconstruction at a fixed fiducial level of geometric detail. A natural way to apply this principle is to divide the reconstruction space into a 3D grid of voxels, with each voxel storing information about the observed surface

points (for example, the 3D points of MVS depthmaps) that fall within or near its interior. These aggregated surface measurements can be used to derive a distance field $u(\mathbf{x})$ for the scene, that is, a value relating the distance to the nearest surface point at each 3D coordinate \mathbf{x} . The distance field can represent absolute distance, or it can be *signed*, with positive values denoting that a point is exterior to a surface and negative values denoting interiority (or vice versa). In either case, the distance function is an implicit representation, where the zero level set delineates the actual underlying surface. The voxel representation of $u(\mathbf{x})$ is convenient to optimize and can be used to extract a mesh for the zero level set of the surface (Lewiner et al., 2003).

When working with 2.5D surface representations like MVS depthmaps, one approach for aggregating distance-to-surface measurements is to consider distance values along each line of sight (Canelhas, 2017). Starting from the camera center, each pixel with known depth is marched through the voxel space along its ray, and the (signed) distance from each voxel center to the observed point is computed and stored; the distance can either be the direct 3D distance to the observed point or, if surface normal estimates are available, the distance to the local plane defined by the point and its normal. This raycasting approach works well near the surface but requires a few considerations when applied to real depthmaps. For one, measurement noise must be accounted for when computing distances, so it is better to (robustly) aggregate the observations rather than taking the minimum computed distance for each voxel as its $u(\mathbf{x})$ value. Second, the 2.5D representation only provides freespace information for the space between the surface and the camera, meaning that interior distances behind the surfaces can only be inferred. Third, depending on the spatial configuration of the source images, there may be rays that do not intersect with a voxel but whose surface points are closer to the voxel’s center than those that do pass through that voxel. This means that the exact distance for exterior surface measurements becomes less reliable for voxels further from the surface.

To address these shortcomings, Curless and Levoy (1996) introduced the truncated signed distance function (TSDF). Here, the aggregated distance values are clipped to a manually determined exterior maximum of $D_{near} > 0$ and an interior minimum of $D_{far} < 0$, which are usually set to a

small multiple of the voxel size. To account for the surface interior being unobserved, computed distances less than D_{near} are ignored. In a perfect scenario, D_{near} would be set to half the distance between the near surface point and the corresponding back surface point along the same ray. In practice, D_{near} should be large enough to accommodate measurement noise yet small enough so as to not negatively impact depth measurements for back surfaces.

Curless and Levoy (1996) suggest a weighted aggregation approach to account for measurement noise. Denoting $\mathcal{F}(\mathbf{x})$ as the set of TSDF distance measurements $\{f_k\}$ and associated weights $\{w_k\}$ for a given voxel, the aggregated TSDF value $T(\mathbf{x})$ is computed as

$$T(\mathbf{x}) = \frac{1}{W(\mathbf{x})} \sum_{(f_k, w_k) \in \mathcal{F}(\mathbf{x})} w_k f_k, \quad W(\mathbf{x}) = \sum_{(f_k, w_k) \in \mathcal{F}(\mathbf{x})} w_k. \quad (5.1)$$

The weight for each observation should reflect a level of confidence that the depth is correct. For example, to account for the larger depth uncertainty in oblique surface views, I use a fronto-parallel bias in my implementation, where the distance for a pixel with ray \mathbf{r}_k and MVS-estimated surface normal \mathbf{n}_k is weighted as $w_k = \mathbf{r}_k \cdot \mathbf{n}_k$.

One additional consideration when aggregating MVS depthmaps in a TSDF is that many spurious depth measurements can occur in general scenes, even when strong filtering is applied to rule out erroneous depths. This means that applying freespace constraints across the entire ray can lead to erroneous carving of the geometry. For example, in some situations, images can have noisy, incorrect geometry estimates for points on the ground that are much farther than the actual distance to the ground surface. Tracing the TSDF along the entire ray thus leads to votes of D_{near} for points on the ground, which ought to have a distance of zero. The crossing rays from many images lead to a false strong confidence that the “ground truth” is far away from the scene’s surface, even though the spurious surface estimates that cast the rays are themselves only weakly supported in the TSDF by a small number of images. A simple way around this is to only begin TSDF aggregation once a minimum distance to the surface (*e.g.*, D_{near}) is reached. The missing freespace values can be inferred via optimization.

5.1.2 Regularizing the Distance Field

The TSDF aggregation procedure can be interpreted as the weighted least-squares solution that minimizes the energy functional

$$E(u(\mathbf{x})) = \int_{\Omega} \frac{1}{2} \sum_{(f_k, w_k) \in \mathcal{F}(\mathbf{x})} w_k (u(\mathbf{x}) - f_k)^2 d\mathbf{x}, \quad (5.2)$$

where the integral is taken over all coordinates \mathbf{x} in the 3D volume Ω . In general, the set $\mathcal{F}(\mathbf{x})$ denotes all observed data points associated with a given point in space, for any choice of data association. In the case where the space is a voxelized TSDF, $\mathcal{F}(\mathbf{x})$ consists of all weighted distance values aggregated in the given voxel. Since the above energy functional can be evaluated point-wise, it is straightforward to see that $T(\mathbf{x})$ is the optimum at each point \mathbf{x} :

$$\begin{aligned} \frac{\partial E}{\partial u}(u(\mathbf{x})) &\stackrel{!}{=} 0 \\ \sum_{(f_k, w_k) \in \mathcal{F}(\mathbf{x})} w_k (u(\mathbf{x}) - f_k) &= 0 \\ u(\mathbf{x}) \sum_{(f_k, w_k) \in \mathcal{F}(\mathbf{x})} w_k - \sum_{(f_k, w_k) \in \mathcal{F}(\mathbf{x})} w_k f_k &= 0 \\ u(\mathbf{x}) &= T(\mathbf{x}). \end{aligned} \quad (5.3)$$

In practice, the surface extracted from a raw aggregated TSDF can be incomplete and, especially when derived from depthmap data obtained using MVS, noisy. To obtain a smooth geometry, Zach et al. (2007) proposed to minimize a total variation (TV) functional:

$$E(u(\mathbf{x})) = \int_{\Omega} |\nabla u(\mathbf{x})| + \lambda \Phi(u(\mathbf{x}), \mathcal{F}(\mathbf{x})) d\mathbf{x}. \quad (5.4)$$

The first term is the so-called total variation penalty that encourages a smooth zero-level set by selecting a distance field that undergoes minimal change. The second term, $\Phi(u, \mathcal{F})$, is a (potentially robust) data term weighted by some value λ . Applying a non-robust squared loss results in the

well-known Rudin-Osher-Fatemi (ROF) model (Rudin et al., 1992):

$$\Phi_{ROF}(u, \mathcal{F}) = \frac{1}{2} \sum_{(f_k, w_k) \in \mathcal{F}} w_k (u - f_k)^2, \quad (5.5)$$

which is the same as the integrand in Eq. (5.2). To make the data term robust to outlier observations, Zach et al. (2007) suggested a TV- $L1$ approach with

$$\Phi_{L1}(u, \mathcal{F}) = \sum_{(f_k, w_k) \in \mathcal{F}} w_k |u - f_k|. \quad (5.6)$$

Ummenhofer and Brox (2015) adopted a similar data term but use a small-threshold Huber model to maintain differentiability near zero.

Let us consider these data term options from the perspective of maximum likelihood estimation, *i.e.*, minimizing the negative log-likelihood of an assumed probability distribution $p_k(f)$ for each distance observation. In the squared (ROF) case, the error from each observation to the true distance value is assumed to follow a Gaussian distribution with variance $\sigma^2 = 1/(\lambda w_k)$. Without a very effective prior on each w_k , this loss is going to be quite sensitive to spurious surface measurements brought about by incorrect estimations in the MVS depthmaps. The $L1$ model assumes an underlying Laplace distribution that lends greater probability to outlier observations; a good confidence estimate in w_k can, of course, still help the approach. The result here is that the median observed distance is preferred, rather than the mean preferred in the ROF case. A Huber loss adopts the Gaussian up to a certain threshold, and then switches to the higher-probability tails of the Laplace distribution. Of course, all three of these data terms assume that non-outlier measurements follow some fixed distribution (Gaussian or Laplace) that may or may not well model the complex surface variations that can arise in MVS depth estimation for general image collections.

From a practical perspective, the squared loss has the nice property that only the aggregated value $T(\mathbf{x})$ and weight $W(\mathbf{x})$ need to be stored in order to compute its derivative (Eq. (5.2)). The $L1$ and Huber approaches require that *all* observations $\mathcal{F}(\mathbf{x})$ be kept in memory, which can be prohibitive when processing billions of points from a reconstruction with several thousand MVS depthmaps.

Histogram binning approaches can potentially avoid this overhead, albeit with quantization error in the target distances (Zach, 2008). When scaling to very large spaces like those found in large-scale Internet photo-collections, an accurate, low-memory solution is preferred. So, unless we have a reason to assume that an L1 noise model would better characterize the non-outlier noise distribution, it follows that a perhaps more scalable approach is to derive a robust loss that uses the squared data term.

An approach that I have found effective is simple truncation of the ROF model:

$$\Phi_{TRUNC}(u, \mathcal{F}) = \begin{cases} \frac{1}{2} \sum_{(f_k, w_k) \in \mathcal{F}} w_k (u - f_k)^2 & \text{if } |u - T| < \tau_T \wedge W > \tau_W \\ 0 & \text{otherwise,} \end{cases} \quad (5.7)$$

for TSDF error threshold τ_T and TSDF weight threshold τ_W . This approach is intended for reconstructions with relatively large voxels (say, greater than 0.1m^3) that aggregate many surface observations from the individual depthmap pixels. In this case, the TSDF computation for well-supported surfaces is quite robust — a small number of spurious point measurements passing through the voxel will not strongly affect the overall weighted average distance. Outliers in the TSDF mainly come from spurious surface estimates for points that are actually in the air or underground. These points are usually only supported by a small number of images, and their associated voxels thus typically have a relatively low aggregated TSDF weight $W(\mathbf{x})$.

Of course, it is also likely the case that valid surfaces exist whose computed $W(\mathbf{x})$ is also small, so the threshold of τ_W may be on its own too restrictive. Accordingly, I start with a value of τ_W that is relatively high and progressively decrease it each iteration. The threshold τ_T helps to restrict spurious points while allowing weakly observed points to become active. The idea here is that the observed points grow “from the ground up,” or to put it perhaps more accurately, outward from the existing surface. That is, only strongly supported surfaces are reconstructed in the early iterations; voxels away from these surfaces begin to adopt strong positive or negative distance values. When τ_W reaches a smaller value, spurious points that are, *e.g.*, floating in the air find themselves in a local distance field of larger positive values. The τ_T threshold and TV smoothness constraint work

together to inhibit a surface to form from this sporadic point. For weakly supported valid points, on the other hand, we have the constraint that no observable static object simply floats in the air or is buried in the ground — it must touch air and be connected to the ground. Weakly observed surfaces near to the current estimated surface are therefore more likely to pass the τ_T check, and thus there is a better chance that the missing structure evolves. Note that this approach will not always work if structures are missed, for example if a street sign is reconstructed in MVS but the pole it is attached to is not.

5.1.3 Optimizing the Distance Field

Zach (2008) proposed an algorithm for solving Eq. (5.4) that follows a framework from Chambolle (2004, 2005). I briefly review this formulation here and introduce an extension for applying a gravity-based prior in the next subsection. Chambolle (2004) noted that the gradient magnitude can be simply expressed as the inner product of the gradient with its associated direction vector, *i.e.*,

$$|\nabla u| = \max_{\mathbf{p}: \|\mathbf{p}\| \leq 1} \mathbf{p} \cdot \nabla u. \quad (5.8)$$

Thus, minimizing a TV- Φ energy can be equivalently expressed as a minimization/maximization of

$$\min_u \max_{\mathbf{p}} E(u, \mathbf{p}) = \min_u \max_{\mathbf{p}} \int_{\Omega} \mathbf{p} \cdot \nabla u + \lambda \Phi(u, \mathcal{F}) \, d\mathbf{x}, \quad (5.9)$$

for which Chambolle demonstrated a convergent optimization algorithm under the ROF model.

Since the data term is zero for voxels that lack observations, it is necessary to introduce a strictly convex relaxation using an auxiliary distance field, v (Zach, 2008):

$$\min_{u,v} \max_{\mathbf{p}} E(u, v, \mathbf{p}) = \min_{u,v} \max_{\mathbf{p}} \int_{\Omega} \mathbf{p} \cdot \nabla u + \frac{1}{2\theta} (u - v)^2 + \lambda \Phi(v, \mathcal{F}) \, d\mathbf{x}, \quad (5.10)$$

where θ is a small positive constant, with smaller values leading to higher faithfulness to the observed distances. This equation can be optimized using gradient ascent scheme of Chambolle

(2005), iterating between updates of \mathbf{p} , v , and u . Specifically, we have the following three update equations from iteration n to $n + 1$:

$$\begin{aligned} \frac{\partial E}{\partial \mathbf{p}} &= \nabla u \\ \implies \mathbf{p}^{(n+1)} &= \pi_{\|\cdot\| \leq 1} (\mathbf{p}^{(n)} + \alpha \nabla u^{(n)}) \end{aligned} \quad (5.11)$$

$$\begin{aligned} \frac{\partial E}{\partial v} &= -\frac{1}{\theta}(u - v) + \lambda \frac{\partial \Phi}{\partial v}(v, \mathcal{F}) \stackrel{!}{=} 0 \\ \implies v^{(n+1)} &= u^{(n)} - \lambda \theta \frac{\partial \Phi}{\partial v}(v^{(n+1)}, \mathcal{F}) \end{aligned} \quad (5.12)$$

$$\begin{aligned} \frac{\partial E}{\partial u} &= -\nabla \cdot p + \frac{1}{\theta}(u - v) \stackrel{!}{=} 0 \\ \implies u^{(n+1)} &= v^{(n+1)} + \theta (\nabla \cdot p^{(n+1)}), \end{aligned} \quad (5.13)$$

where $\alpha \leq 1/(6\theta)$ is the step size for the gradient ascent scheme, and $\pi_{\|\cdot\| \leq 1}$ projects any vector with greater than unit length onto the unit sphere. The exact update for $v^{(n+1)}$ depends on the choice of Φ , of course — for the (truncated) ROF model, it is straightforward to compute the derivative of Φ , whereas models such as the TV- $L1$ may require a more nuanced update calculation (Zach, 2008). In the discretized setting, the gradient computations must be dual, *e.g.*, using forward differences for the computation of ∇u and backward differences for the computation of $\nabla \cdot p$ (Zach, 2008).

5.1.4 Gravity-aligned Surface Prior

One drawback of total variation approaches is that they can unrealistically interpolate surfaces in weakly supported and unsupported regions. For example, if the images in a photo-collection frequently capture heavy occlusion near the base of a building, multi-view stereo is likely to fail to reconstruct any geometry for the lower facade and the ground. Total variation regularization provides us with a way to recover this missing geometry, but the minimal surface in the TV model consists of a smooth curve between the well-supported portions of the upper building facade and farther-away ground. This improves scene completeness, but the underlying geometry can be quite incorrect.

To more realistically capture such missing surfaces, one solution I have found works well is to assume a Manhattan-world-type property for the scene (Coughlan and Yuille, 1999), enforcing that scene surfaces in unknown regions are either flat or vertical relative to the gravity vector of the scene.¹ Assuming that the reconstruction space is gravity-aligned with the y -axis pointing downward, the surface normal $\mathbf{n} = (n_1, n_2, n_3)^T$ would thus have $|n_2| = 1$ for a flat surface (*e.g.*, the ground) and $n_2 = 0$ for a vertical surface (*e.g.*, the facade of a building). Since $\mathbf{p}(\mathbf{x}) = (p_1(\mathbf{x}), p_2(\mathbf{x}), p_3(\mathbf{x}))^T$ serves as a “decoupled” estimate of the surface normal during optimization, the TV energy equation can be updated to include a regularization of this vector. Considering the energy only w.r.t \mathbf{p} for simplicity, we now have

$$\max_{\mathbf{p}} E_M(\mathbf{p}) = \max_{\mathbf{p}} \int_{\Omega} \mathbf{p} \cdot \nabla u - \lambda_M (1 - p_2^2) p_2^2 d\mathbf{x}, \quad (5.14)$$

where λ_M is a regularization weight for the Manhattan-world prior. The new term can be interpreted as “the magnitude of the vertical component of \mathbf{p} is either zero or one.” Values of p_2 away from this are penalized.

Computing an update for \mathbf{p} , we have

$$\frac{\partial E_M(\mathbf{p})}{\partial \mathbf{p}} = \nabla u - \begin{pmatrix} 0 \\ 2\lambda_M p_2 - 4\lambda_M p_2^3 \\ 0 \end{pmatrix} \quad (5.15)$$

The values for the horizontal components of $\frac{\partial E_M}{\partial \mathbf{p}}$ are the same as before, *i.e.*, $\frac{\partial E_M}{\partial p_1} = \nabla_x u$ and $\frac{\partial E_M}{\partial p_3} = \nabla_z u$, and we again can use gradient ascent to compute a new value for p_1 and p_3 . However, the vertical component now follows a cubic equation that can be found analytically by evaluating $\frac{\partial E_M}{\partial p_2} \stackrel{!}{=} 0$, without the need for gradient ascent. Rewriting the equation for p_2 from above in standard cubic form, we have

$$p_2^3 - \frac{1}{2}p_2 + q = 0, \quad (5.16)$$

¹Manhattan-world priors may also enforce orthogonality for adjoining vertical surfaces like the walls of a building. I do not enforce that here, however.

where $q = \frac{1}{4\lambda_M} \nabla_y u$. If the discriminant $D = -4 \left(-\frac{1}{2}\right)^3 - 27q^2$ of Eq. (5.16) is non-negative (or equivalently, if $3|q|\sqrt{6} \leq 1$), then there exist three real solutions for p_2 given by

$$\hat{p}_2^{(k)} = 2\sqrt{\frac{1}{6}} \cos \left(\frac{1}{3} \left(\cos^{-1} \left(-3q\sqrt{6} \right) - 2\pi k \right) \right), \quad k = 0, 1, 2. \quad (5.17)$$

(see, for example, Weisstein (2019)). Otherwise, there exists only one real-valued solution:

$$\hat{p}_2 = -2\sqrt{\frac{1}{6}} \operatorname{sgn}(q) \cosh \left(\frac{1}{3} \cosh^{-1} \left(-3|q|\sqrt{6} \right) \right). \quad (5.18)$$

In the case of three real-valued solutions, the updated value \hat{p}_2 can be selected as that closest to $\frac{\nabla_y u}{|\nabla u|}$, or zero if the current gradient of u is zero. The final updated value is $\mathbf{p}^{(n+1)} = \left(p_1^{(n)} + \alpha \nabla_x u^{(n)}, \hat{p}_2, p_3^{(n)} + \alpha \nabla_z u^{(n)} \right)^T$, which is then reprojected down to S^2 if it is longer than unit length.

5.1.5 Scenario-specific Considerations and Implementation

To speed up processing times for large-scale datasets, I have created GPU implementations of the TSDF aggregation and optimization algorithms described above. In addition to the gravity-aligned surface prior, there are two specific implementation approaches that I have found effective in improving the overall result. The first such approach is to add the constraint that the bottom layer of voxels in the voxel space is always interior (*i.e.*, underground). After each iteration, this constraint is applied by simply enforcing that $u(\mathbf{x}) \leq -h$ for voxels along the bottom boundary, where h is the voxel size. A similar constraint is also applied to the top boundary, so that the distance function at the top of the volume is never interior. The second approach is to apply a vertical “sweep” initialization to better propagate interior and exterior values from the raw TSDF.

Starting from the top of the TSDF and sweeping down each column, each voxel is updated as

$$u'(\mathbf{x}) = \begin{cases} u(\mathbf{x}) & \text{if } W(\mathbf{x}) > \tau_W \vee u'_{prev} \geq 0 \\ \max(u'_{prev} - h, -2h) & \text{otherwise,} \end{cases} \quad (5.19)$$

where u'_{prev} is the value for the voxel immediately above. This approach increases the initial coverage of underground voxels, whose distance values should be negative. In a second pass, the same approach is applied in an upward sweep, filling in positive distances. This sweeping strategy is only applied once, before optimization. The effect of these two constraints are shown in the results section, below.

One final implementation detail concerns the addition of person ground points into the scene, which are obtained using the approach described in the previous chapter. Adding these ground points provides an important — and often, the only — cue for delineating where the ground surfaces lie in the scene. After the initial TSDF aggregation, the 3D point for each person is added to the TSDF simply by computing the nearest voxel centers above and below it. The TSDF is updated at these voxels based on the vertical distance to the ground point with a weight of $w_k = 1$. To increase the spatial effect of the ground point placement, I also apply this update to the 3×3 neighborhood in the XZ-plane around each of the two voxels.

5.2 Triangle Color Estimation and Walkable Area Extraction

Once optimization is complete, a final surface mesh is extracted using marching cubes (Lorenson and Cline, 1987; Lewiner et al., 2003). For visualization and for computing walkable ground regions of the reconstructed surface, I first compute an average color for each triangle in the mesh. This is done by rendering the mesh into each image using the camera intrinsics and extrinsics estimated via SfM. I use OpenGL vertex and fragment shaders to render the triangle indices of the mesh into each image, which provides the triangle associated with each pixel. The color for each projected triangle is computed as the average color of all relevant pixels in the image, and the final triangle color for

the mesh is computed as the average color over all images, with each image weighted by the number of pixels occupied by the triangle in that image. To avoid aggregating colors for transient objects such as people, birds, and cars, I exclude projected triangles that overlap with the bounding boxes of objects detected by an implementation of the Mask R-CNN neural network (He et al., 2017; Abdulla, 2017). Triangles with missing colors are filled using an iterative diffusion-type approach, where at each iteration, the new color is determined as the average color of its neighbors.

Having obtained triangle colors for the mesh, the next step is to compute the parts of the mesh on which pedestrians may actually walk. This information is available, in part, by the 3D ground points recovered for people detected in the individuals images — wherever a person was determined to be standing, the nearest mesh triangle can be considered as a ground surface. These triangles are outlined in white in Fig. 5.2. Depending on reconstructed crowd density, however, this surface is often incomplete, with many isolated patches of ground surfaces. To improve the coverage of walkable area, I adopt a simple color-based region-growing approach. Starting from an initial set of ground-labeled triangles, I iteratively check all ground adjacent triangles. If the color of the adjacent triangle is similar enough to the average color of its ground-labeled neighbors, that triangle also receives a ground label. This computation is iteratively applied until no adjacent triangle is relabeled as ground. To reduce sensitivity to mild color variations, I apply this computation after two initial color blur operations of the original colored mesh. To compute color similarity, I convert each red-green-blue vertex color into the L*a*b* color space with per-channel normalization to the range $[0, 1]$. A color c is considered similar to its average neighboring color μ if

$$\left(\prod_{i \in \{L^*, a^*, b^*\}} e^{-\frac{|c_i - \mu_i|}{2\sigma_i}} \right) < \tau_c. \quad (5.20)$$

I use $\sigma_{L^*} = 0.4$, $\sigma_{a^*} = \sigma_{b^*} = 0.35$, and $\tau_c = 0.05$. Example triangles added by this region-growing process are outlined in red in Fig. 5.2.

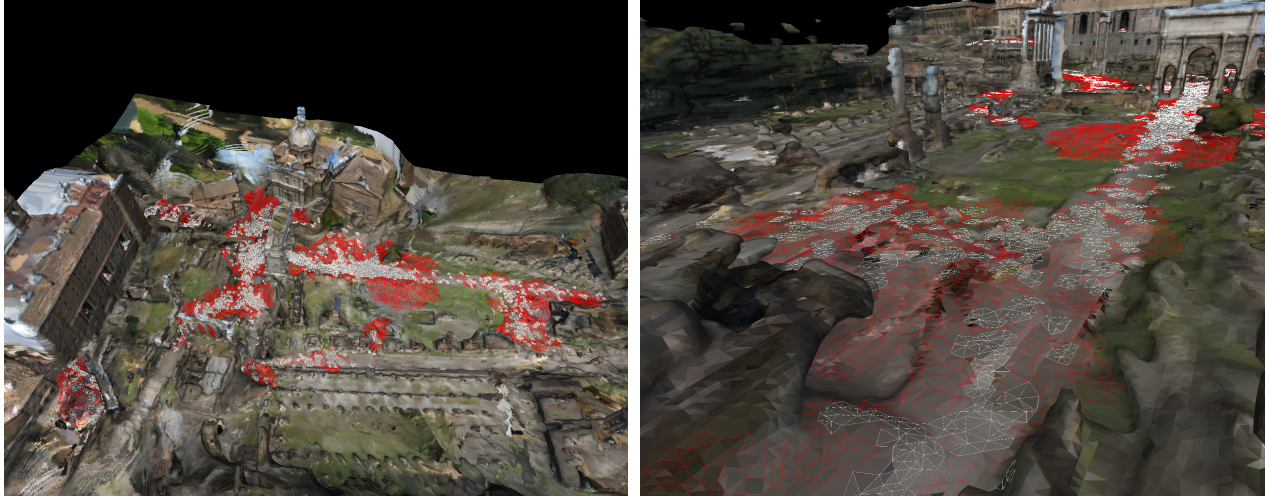


Figure 5.2: Initial labeled ground triangles (white) and additional triangles added after the proposed region-growing approach (red).

5.3 Crowd Simulation and Visualization

Once the “walkable” areas in the scene have been identified, a living 3D reconstruction can be achieved by adding animated pedestrians who walk over these surfaces. This is accomplished using Menge (Curtis et al., 2016), an the off-the-shelf tool that implements a variety of crowd simulation algorithms. The walkable ground surfaces form a navigation mesh, over which the virtual pedestrians plan paths to specified target destinations. For the visualizations I produce here, I randomly select starting and target destinations using the set of recovered ground positions for people detected in the input imagery. Once a virtual pedestrian reaches their target location, a new target location is randomly selected.

5.4 Results

I have applied the proposed pipeline to a number of large-scale community photo-collections from the MegaDepth dataset (Li and Snavely, 2018). Snapshots of the living 3D reconstructions with animated pedestrians are shown in Figs. 5.4-5.11. For these reconstructions, I apply a two-pass optimization of the distance field, first at a voxel size of 1m ($D_{near} = 4m$ and $D_{far} = -1m$) and then, using this result as initialization, at a resolution of 0.5m ($D_{near} = 3m$ and $D_{far} = -0.8m$).

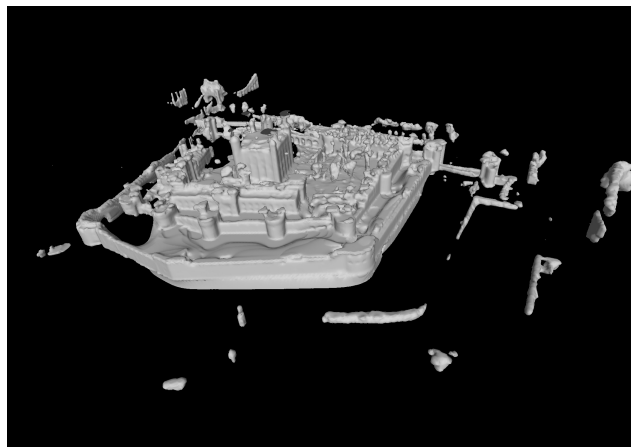
For voxel size h , I use $\lambda = \left(\frac{3}{2h}\right)^2$, $\lambda_M = 10$, $\theta = (0.01h)^2$, and $\tau_T = 2h$. The 1m resolution uses $\tau_W = 320$, which is decreased by 0.01% each iteration, and TSDF voxels with $W(\mathbf{x}) < 20$ are filtered out prior to the addition of 3D person ground points to the TSDF. Having been initialized from the 1m-resolution result, the 0.5m resolution uses $\tau_W = 0$, and TSDF voxels with $W(\mathbf{x}) < 10$ are filtered out, again prior to the addition of 3D person ground points to the TSDF. The 1m and 0.5m optimizations are run until the maximum percent change in any voxel is less than 0.5% and 1%, respectively.

Fig. 5.3 provides a qualitative ablative analysis for the surface reconstruction results of the Tower of London at a voxel resolution of 1m, considering the additional pipeline components of sweep initialization, the gravity-aligned prior, and the constraint that the bottom voxels of the reconstruction be interior. Of these, the interior constraint does the most for filling out the ground region of the scene; without it, the TV term merely enforces minimal surfaces around well-supported voxels. The sweep initialization generally leads to smoother surfaces when combined with the interior constraint, although TSDF noise and oversmoothing can still negatively affect the result. The gravity prior notably improves the overall local flatness and smoothness of the estimated surfaces.

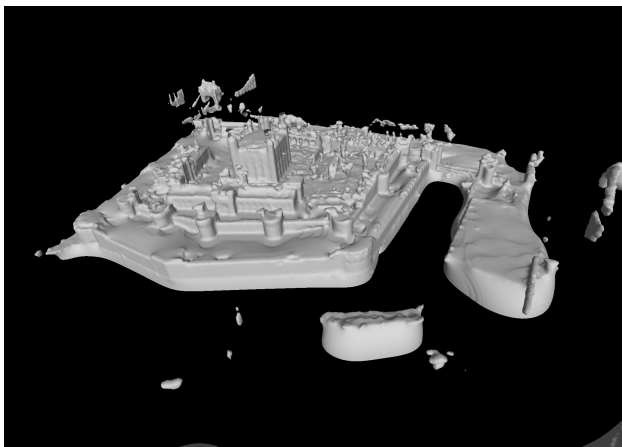
5.5 Discussion

This chapter introduced *living 3D reconstructions* — 3D reconstructions of real places that include dynamic representations of transient objects. As part of obtaining this representation, I introduced new considerations for ground surface modeling and general scene reconstruction that enhance the overall completeness of the scene versus alternative approaches. This is accomplished, in part, by leveraging the ground surface points for detected people, the scale of the scene, and the gravity direction of the scene. Crowd simulation additionally allows for a *resembling reconstruction* of the place, bridging the gap in realism by adding missing dynamic objects at scale. The resulting visualization moves beyond the static scene reconstruction of SfM and MVS, providing missing context for how the scene exists as a whole, complete with its individual buildings and ground surfaces, as well as representations of people that move about the space.

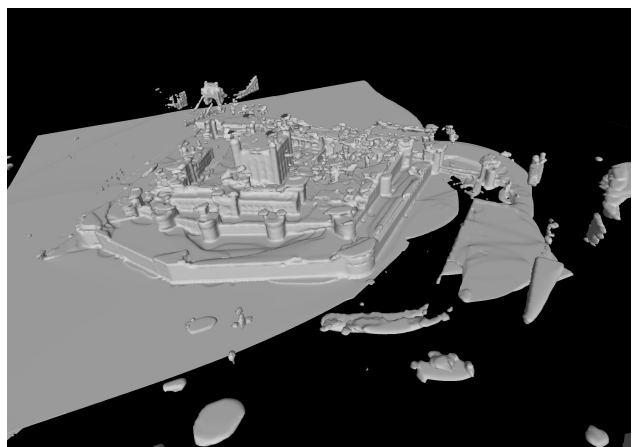
No additional settings



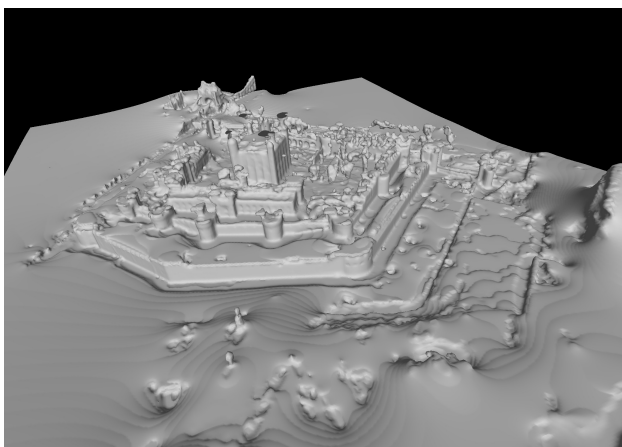
With sweep initialization



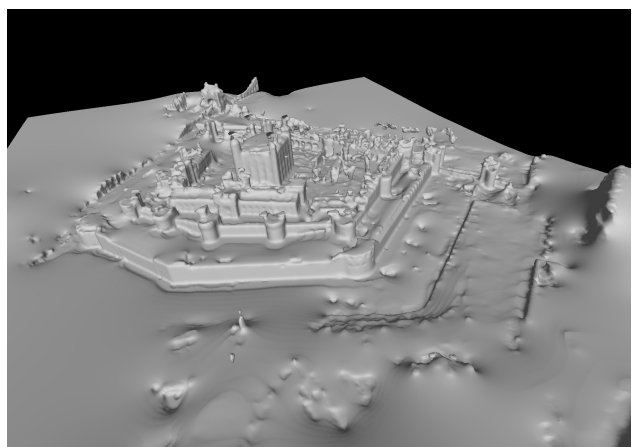
With gravity-aligned prior



With the bottom voxels constrained to be interior



With sweep initialization and interior constraint



With sweep, interior, and gravity prior

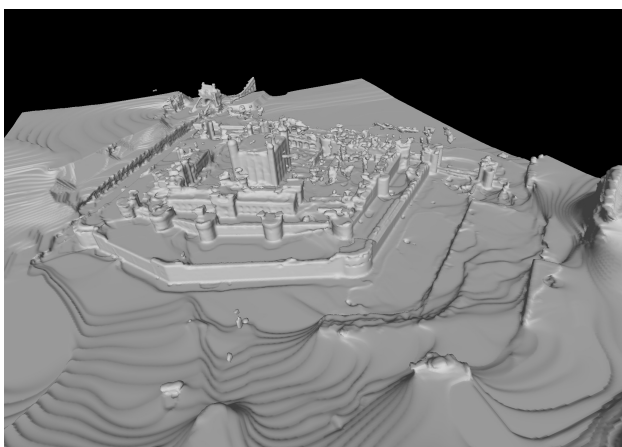


Figure 5.3: Qualitative, ablation comparison of scene reconstruction results for the Tower of London under the proposed implementation at a voxel resolution of 1m.

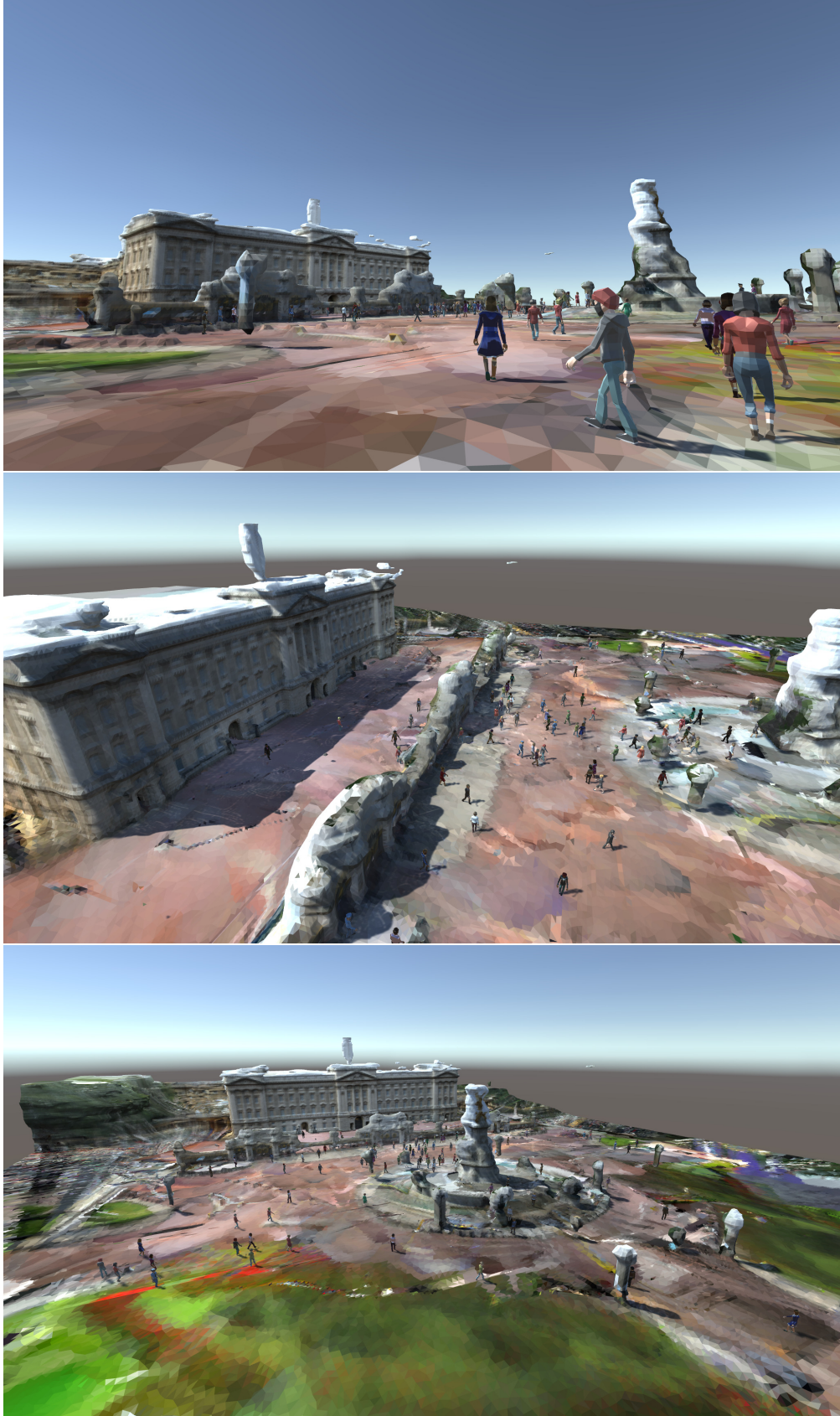


Figure 5.4: Reconstruction of Buckingham Palace.



Figure 5.5: Reconstruction of the Castel Sant'Angelo in Rome.

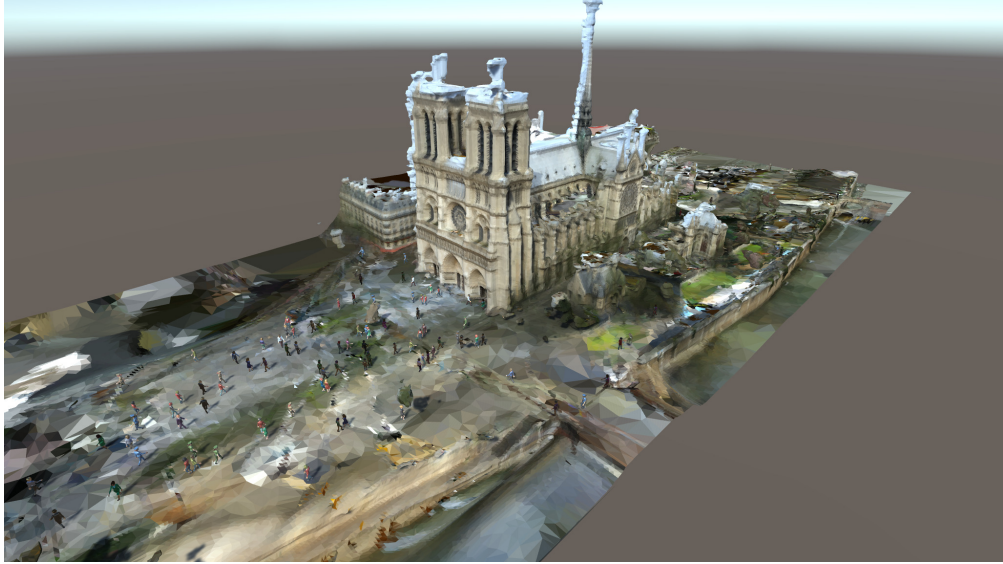


Figure 5.6: Reconstruction of the Notre Dame Cathedral in Paris.



Figure 5.7: Reconstruction of the Old Town Square in Prague.

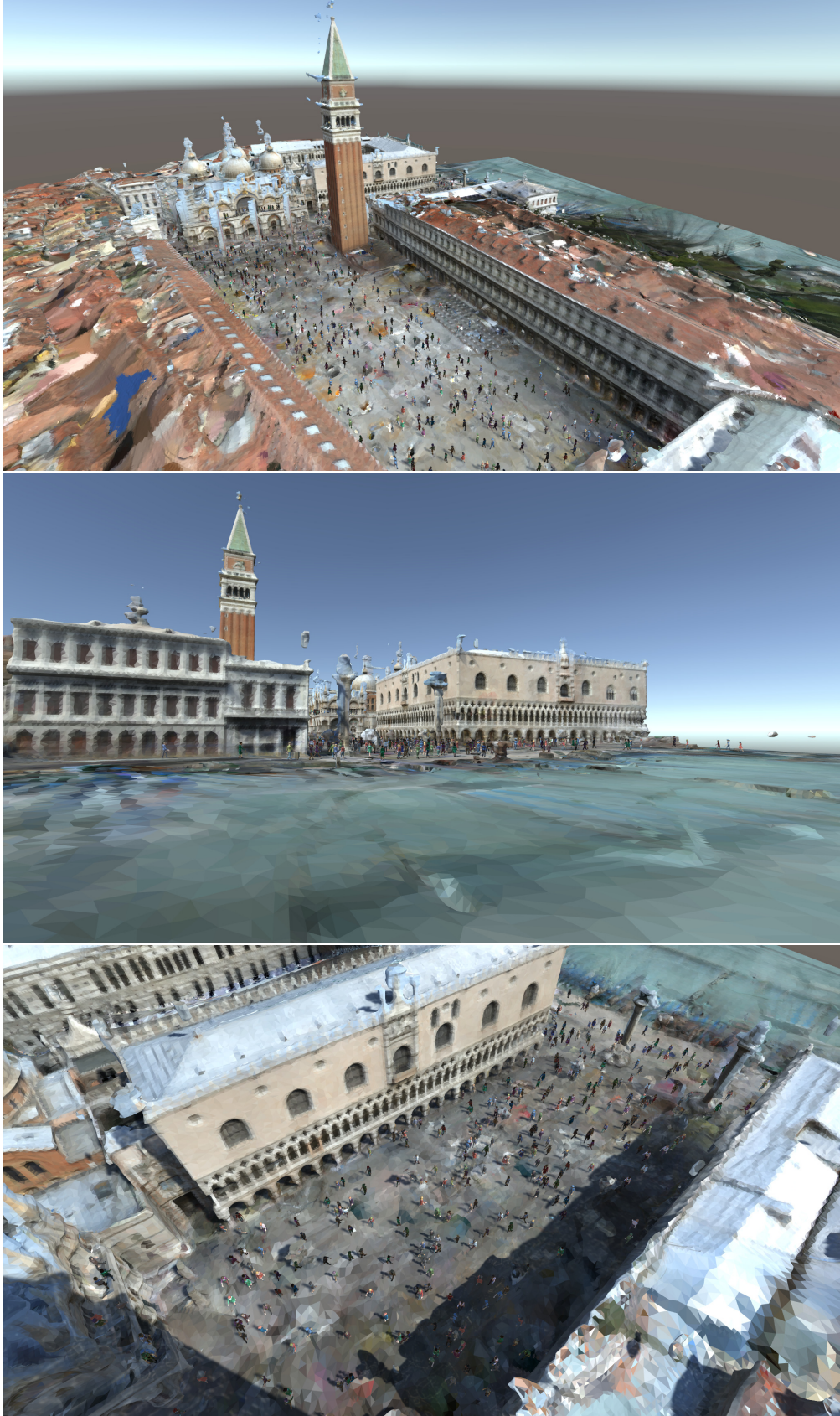


Figure 5.8: Reconstruction of the Piazza San Marco in Venice.



Figure 5.9: Reconstruction of the Sacré Cœur Basilica in Paris.



Figure 5.10: Reconstruction of the Tower of London.



Figure 5.11: Additional views of the reconstruction of the Tower of London.

CHAPTER 6: CONCLUSION AND FUTURE WORK

This dissertation presented a variety of solutions for extending 3D scene reconstruction beyond the gold-standard capabilities of Structure-from-Motion (SfM) and Multi-view Stereo (MVS), each tailored to a specific scenario. Chapter 3 addressed the problem of 3D reconstruction from endoscopic video. In this setting, SfM can generally provide sparse surface structure, but the lack of surface texture as well as complex, changing illumination conditions often causes MVS to produce incomplete or erroneous surfaces. To overcome these difficulties in dense surface reconstruction, I introduced a Shape-from-Motion-and-Shading (SfMS) method that utilizes an SfM-guided approach to Shape-from-Shading (SfS). In this context, SfM is used both to guide surface reflectance estimation and to regularize the SfS equation. I also introduced a 1D bidirectional reflectance distribution function to better model the illumination conditions of the endoscope, and I proposed a solution to account for light interreflections off the surface, which otherwise violate the 1D BRDF model.

In Chapter 4, I introduced an approach for augmenting 3D reconstructions from large-scale Internet photo-collections by recovering the 3D position of transient objects — specifically, people — in the input imagery. Since no two images can be assumed to capture the same person standing in the same location from two different angles, the typical triangulation constraints enjoyed by SfM and MVS cannot be directly leveraged to perform this reconstruction. I introduced an alternative method that leverages approximate semantic triangulation of objects of the same class type (in this case, pedestrians). The method is aided by constraints on the height distribution of people, as well as visibility and freespace constraints provided by the static reconstruction of the scene obtained via SfM. As a part of this reconstruction process, my approach additionally recovers the scale of the scene, its gravity direction, and an estimate for the ground surface normals for the point at which

each individual stands. Recovering scene scale is an especially important result in this process, as it cannot be otherwise automatically obtained using SfM.

Finally, I introduced in Chapter 5 the concept of using crowd-sourced imagery to create *living 3D reconstructions* — visualizations of real places that include dynamic representations of transient objects like pedestrians. This idea seeks to enhance 3D reconstructions such that they better resemble the place captured in the input imagery. As part of my approach for creating this visualization, a key difficulty to overcome is that ground surfaces are often poorly reconstructed using a typical SfM+MVS pipeline for Internet images. I leverage the 3D ground points obtained by the method in Chapter 4, along with the estimated scene scale, and introduced a tailored volumetric reconstruction approach that convincingly reconstructs ground surfaces in the scene. Crowd simulation (Curtis et al., 2016) is then employed to add virtual pedestrians to the space, who move between waypoints along the “walkable” surfaces of the scene.

6.1 Future Directions

There are a number of possible directions for future work on the topics presented in this thesis. I address a few ideas for extensions in the following subsections.

6.1.1 Extensions to Shading-based Endoscopic Reconstruction

The SfMS approach introduced in Chapter 3 provides a workable solution for joining SfM and SfS for endoscopic scenarios. However, the approach perhaps does not utilize multi-view information to its full potential. For one immediate example, the BRDF estimation process could actually be applied to all images simultaneously, rather than each image independently. This would provide many more SfM point observations to improve the result and make the reflectance more consistent on a frame-by-frame basis. A more elegant interreflection function is also highly desirable — the polynomial approximation I presented here really only demonstrates that such modeling is necessary. Due to changes in lighting over the course of the video, it is likely not possible to apply a single interreflection model to all frames simultaneously, but temporal constraints in adjacent video

frames could potentially be utilized. An approach that learns to generate interreflections, like the method of Li et al. (2018), is also compelling, since it could be used to abstract the complex models of reflectance, lighting, and shape estimation. A key open question with deep-learning-based approaches, however, is whether they can be truly generalized to the medical domain without requiring a large amount of high-quality real-world training data, which can be difficult to obtain and whose modalities may not offer complete supervision for the target application. Synthetic datasets offer an alternative path forward for large-scale network training; however, it is difficult to guarantee that synthetic imagery will have sufficient realism and diversity to allow the trained network to be successfully applied to real-world data.

Setting aside these potential challenges in obtaining training data, there are a number of intriguing avenues for further learning-based extensions. For instance, while it may be difficult to train a network to output monocular depth for each frame, it might be possible to learn to regress a surface normal, and perhaps a confidence in this prediction, at each pixel. This estimation would much better constrain the BRDF estimation of the method, since it would not rely on an existing surface to produce the per-point normal estimates.

In another direction, neural networks could potentially be used to replace the Lax-Friedrichs solution scheme for SfS, which introduces dissipative terms to maintain stability and, in the proposed implementation, only works by decreasing estimated depth values. Instead, it is theoretically possible to train a neural network to learn to perform this optimization automatically and to internally store an understanding of shape priors, analogous to the approach of Cherabier et al. (2018). The network could be trained either to generate updates for (log-)depth values given a current BRDF-prediction-versus-intensity error map, or it could alternatively completely optimize the solution given only a BRDF, an input image, and an initial surface, including depth values for SfM points. The network might be trained using synthetic renderings of objects; however, it is again unclear whether the approach would easily generalize to medical imagery if the network requires a color image as input.

Another extension of this method, still using a deep-learning approach, could be to perform monocular depth prediction guided by sparse SfM points, *e.g.*, taking as input the color image and

an associated sparse depthmap. The sparse constraints here would provide a signal for depth that neural networks could effectively use for depth disambiguation. This approach could provide an effective means for reconstructing weakly textured surfaces, such as uniformly colored walls. The approach could also be extended to include rendering-based refinement, similar to Li et al. (2018).

Returning to extensions of the method for endoscopy, one obvious target is to employ non-rigid SfM to improve the length and reliability of the reconstruction. This is a notoriously difficult task for endoscopic imagery (Münzer et al., 2018), but recent advances in dense non-rigid surface reconstruction (Innmann et al., 2019) might offer some paths forward if combined with shading constraints. It may also be possible to integrate multi-view constraints as part of the SfS algorithm, itself, to better leverage photometric consistency in a manner similar to Wu et al. (2010).

6.1.2 Extensions to 3D Reconstruction of Transient Objects and Living 3D Reconstructions

There are a number of exciting future directions for extending the work presented in Chapters 4 and 5. In particular, I am intrigued by the prospect of improving the immersion of virtual people into the scene: Instead of simply walking from waypoint to waypoint, can virtual pedestrians be rendered to interact with their environment in a convincing way? For example, if we detect a bench in the input imagery and place a model of it into the scene, pedestrians could be animated to go to the bench, sit, and perform some action, perhaps talking on the phone, eating a meal, or simply contemplating the world around them. The same sort of idea could be extended to doors (with animation of them opening and closing). Even more compelling is the idea that *scene understanding* is possible given the enhanced context of where people exist within the space. Given that SfM provides us with the location of registered images, it is already straightforward to answer the question, “Where is a good place to take a picture?” However, understanding where people exist in the scene and recognizing their actions potentially allows us to ask advanced questions, for example, “Where is a good spot to have a picnic?” or “Where is the line to enter the museum?” In the first scenario, we can detect people in the input imagery who are sitting in a grassy area and use this to answer the user’s question without external input. In the second scenario, we can detect lines of

people in the input images and use this in conjunction with maps of the scene to infer the likely location for the line to start.

Other fun examples emerge if object classes are modeled, in addition to people. For example, if we reconstruct canines, we could ask the question, “Where in the park am I likely to find the most dogs?” The scene could also be animated to include animals in a convincing way, for example having a person playing fetch with their dog in the same park, or birds eating breadcrumbs in front of a cafe. Other semantic classes that could be modeled include cars and bicyclists.

There are a number of reconstruction challenges that could be addressed in order to obtain more “true-to-life” scene representations. For example, small and/or thin objects like railings are often missed by my current implementation. Such objects are often difficult to reconstruct initially, during MVS. It may be best to separately model these objects and add them as details in the reconstruction, rather than try and directly model them in the voxelized space. Improving the texturing of the reconstruction is another open problem. In my implementation, I simply used per-face coloring to portray a rough visualization of the scene. In practice, texturing methods like that of Waechter et al. (2014) produce quite nice and highly realistic results for certain scene elements, but significant artifacts and poorly textured surfaces still frequently occur, especially for ground surfaces. Texturing from aerial imagery is one solution, except for parts of the scene that are not visible from above. Synthetic texture generation is another possibility; one interesting approach might be to train a neural network to generate realistic ground textures for ground-level imagery given a training set that contains registered aerial views.

Another research direction involves obtaining photorealistic ground-level visualizations. Can I look at a 3D reconstruction of a far-away place in, say, VR and feel like I am actually there? In this aspect, the “raw” 3D models obtained via 3D reconstruction often lack sufficient completeness and detail, and it is difficult to remove artifacts from the reconstruction output completely. Even as the corner cases of 3D reconstruction continue to be solved, it may be impractical to expect that a convincing 3D ground-level visualization can be obtained for general scenes and diverse image collections solely by rendering reconstructed 3D meshes. However, given the recent success

of deep learning approaches for reconstruction-driven image-based rendering (IBR) in controlled imagery (for example, Hedman et al. (2018)), it may be possible to apply similar principles to general Internet-based 3D reconstructions. One key difficulty here is that the input imagery must be normalized to have a consistent appearance for view blending; however, recent advances in neural rendering (Meshry et al., 2019) show much promise in realizing such an approach. Dynamic neural rendering is another very interesting idea in this space: Given that we can now render virtual pedestrians into the scene, could we add them to an IBR visualization and re-render them to appear lifelike? The possibilities for the future have only just begun.

APPENDIX A: DERIVATION OF ARTIFICIAL VISCOSITY VALUES IN SFS PDE SOLUTION

Eq. (3.23) outlined an approach for computing acceptable values of $\sigma_{i,j}^x$ and $\sigma_{i,j}^y$ in the Lax-Friedrichs Hamiltonian for use in Shape-from-Shading. The proposed value for $\sigma_{i,j}^x$ (and similarly for $\sigma_{i,j}^y$) is

$$\tilde{\sigma}_{i,j}^x = \max_{\cos(\theta) \in (0, T_{i,j}^x]} \left| \frac{\partial \tilde{\eta}}{\partial \cos(\theta)}(\cos(\theta)) \right| \max_{p,q} \left| \frac{\partial \cos(\theta)}{\partial p}(x_i, y_j, p, q) \right| \quad (\text{A.1})$$

$$\tilde{\sigma}_{i,j}^y = \max_{\cos(\theta) \in (0, T_{i,j}^y]} \left| \frac{\partial \tilde{\eta}}{\partial \cos(\theta)}(\cos(\theta)) \right| \max_{p,q} \left| \frac{\partial \cos(\theta)}{\partial q}(x_i, y_j, p, q) \right|, \quad (\text{A.2})$$

where $T_{i,j}^x$ is the largest possible value of $\hat{\mathbf{n}}(x_i, y_j) \cdot \hat{\mathbf{1}}(x_i, y_j) = \cos(\theta)$ given $p_{i,j}^+$ and $p_{i,j}^-$, for any value of q , and similarly for $T_{i,j}^y$.

The value of $T_{i,j}^x$ is relatively straightforward to compute. Recall from Eqs. (3.7) and (3.8) that $\cos(\theta)$ can be expressed as a function of $x = x_i$, $y = y_j$, $p = v_x$, and $q = v_y$:

$$\cos(\theta) = \frac{1}{\sqrt{(x^2 + y^2 + 1)(p^2 + q^2 + (xp + yq + 1)^2)}}, \quad (\text{A.3})$$

and therefore

$$\frac{\partial \cos(\theta)}{\partial q} = -\frac{q + x(xp + yq + 1)}{\sqrt{(x^2 + y^2 + 1)(p^2 + q^2 + (xp + yq + 1)^2)^3}}. \quad (\text{A.4})$$

Since $\cos(\theta)$ is maximized w.r.t q when $\frac{\partial \cos(\theta)}{\partial q} = 0$ (note that it is minimized only in the limit, as $|q|$ approaches infinity), we can set the numerator in the above equation to zero and arrive at $\hat{q} = -\frac{y(xp+1)}{y^2+1}$. Plugging this into the equation for $\cos(\theta)$, it follows that

$$T_{i,j}^x = \max_{p \in \{p_{i,j}^+, p_{i,j}^-\}} \sqrt{\frac{y^2 + 1}{(x^2 + y^2 + 1)(p^2(x^2 + y^2 + 1) + 2xp + 1)}}, \quad (\text{A.5})$$

and for the y case (where $\hat{p} = -\frac{x(yq+1)}{x^2+1}$):

$$T_{i,j}^y = \max_{q \in \{q_{i,j}^+, q_{i,j}^-\}} \sqrt{\frac{x^2 + 1}{(x^2 + y^2 + 1)(q^2(x^2 + y^2 + 1) + 2yq + 1)}}. \quad (\text{A.6})$$

Since the second maximum in both Eq. (A.1) and Eq. (A.2) is taken without bounds on p and q , the value of this maximum can be computed at any position ($x = x_i, y = y_j$). Considering $\sigma_{i,j}^y$ first, we can find the extrema of $\frac{\partial \cos(\theta)}{\partial p}$ by evaluating $\frac{\partial}{\partial p} \frac{\partial \cos(\theta)}{\partial p} \stackrel{!}{=} 0$ and $\frac{\partial}{\partial q} \frac{\partial \cos(\theta)}{\partial p} \stackrel{!}{=} 0$. Setting these to equality and solving for q , it works out that the value for q that minimizes and/or maximizes $\frac{\partial \cos(\theta)}{\partial p}$ is given simply as

$$\hat{q} = -\frac{y}{x^2 + y^2 + 1}. \quad (\text{A.7})$$

Evaluating $\frac{\partial}{\partial p} \frac{\partial \cos(\theta)}{\partial p} \stackrel{!}{=} 0$ at this value of q , we find two possible values of p :

$$\hat{p} = \frac{-2x^5 - 2x^3(y^2 + 2) - 2x(y^2 + 1) \pm \sqrt{2(x^2 + 1)(x^2 + y^2 + 1)^3}}{2(x^2 + 1)(x^2 + y^2 + 1)^2}. \quad (\text{A.8})$$

It turns out that, for any value of (x, y) , $\frac{\partial \cos(\theta)}{\partial p}$ is maximized if the plus is taken and minimized if the minus is taken. Moreover, $|\frac{\partial \cos(\theta)}{\partial p}|$ is maximized at (\hat{p}, \hat{q}) regardless of whether the plus or minus is taken.

The derivation is similar for $\sigma_{i,j}^x$. Specifically, $|\frac{\partial \cos(\theta)}{\partial q}|$ is maximized by

$$\hat{p} = -\frac{x}{x^2 + y^2 + 1} \quad (\text{A.9})$$

$$\hat{q} = \frac{-2y^5 - 2y^3(x^2 + 2) - 2y(x^2 + 1) \pm \sqrt{2(y^2 + 1)(x^2 + y^2 + 1)^3}}{2(y^2 + 1)(x^2 + y^2 + 1)^2}. \quad (\text{A.10})$$

REFERENCES

- Abdulla, W. (2017). Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN.
- Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S. M., and Szeliski, R. (2011). Building rome in a day. *Communications of the ACM*, 54(10):105–112.
- Ahmed, A. and Farag, A. (2007). Shape from shading for hybrid surfaces. In *2007 IEEE International Conference on Image Processing*, volume 2, pages II–525. IEEE.
- Ahmed, A. H. and Farag, A. A. (2006). A new formulation for shape from shading for non-lambertian surfaces. In *Computer Vision and Pattern Recognition (CVPR)*.
- Akhter, I., Sheikh, Y., Khan, S., and Kanade, T. (2009). Nonrigid structure from motion in trajectory space. In *Advances in Neural Information Processing Systems*, pages 41–48.
- Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., and Savarese, S. (2016). Social lstm: Human trajectory prediction in crowded spaces. In *Computer Vision and Pattern Recognition (CVPR)*, pages 961–971.
- Albrecht, T., Baumann, I., Plinkert, P. K., Simon, C., and Sertel, S. (2016). Three-dimensional endoscopic visualization in functional endoscopic sinus surgery. *European Archives of Otorhino-Laryngology*, 273(11):3753–3758.
- Avidan, S. and Shashua, A. (2000). Trajectory triangulation: 3d reconstruction of moving points from a monocular image sequence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):348–357.
- Badiqué, E., Komiya, Y., Ohyama, N., Honda, T., and Tsujiuchi, J. (1988). Use of color image correlation in the retrieval of gastric surface topography by endoscopic stereopair matching. *Applied Optics*, 27(5):941–948.
- Baiget, P., Fernández, C., Roca, X., and González, J. (2009). Generation of augmented video sequences combining behavioral animation and multi-object tracking. *Computer Animation and Virtual Worlds*, 20(4):473–489.
- Barron, J. T. and Malik, J. (2014). Shape, illumination, and reflectance from shading. *Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1670–1687.
- Bartoli, A., Gérard, Y., Chadebecq, F., Collins, T., and Pizarro, D. (2015). Shape-from-template. *Pattern Analysis and Machine Intelligence (PAMI)*, 37(10):2099–2118.
- Bernhardt, S., Nicolau, S. A., Bartoli, A., Agnus, V., Soler, L., and Doignon, C. (2015). Using shading to register an intraoperative ct scan to a laparoscopic image. In *Computer-Assisted and Robotic Endoscopy*, pages 59–68. Springer.

- Bernhardt, S., Nicolau, S. A., Soler, L., and Doignon, C. (2017). The status of augmented reality in laparoscopic surgery as of 2016. *Medical Image Analysis*, 37:66–90.
- Best, J. (2019). How virtual reality is changing medical practice: “doctors want to use this to give better patient outcomes”. *BMJ*, 364:k5419.
- Bickerton, R., Nassimizadeh, A.-K., and Ahmed, S. (2019). Three-dimensional endoscopy: The future of nasoendoscopic training. *The Laryngoscope*.
- Billings, S. and Taylor, R. (2015). Generalized iterative most likely oriented-point (G-IMLOP) registration. *International Journal of Computer Assisted Radiology and Surgery*, 10(8):1213–1226.
- Black, J., Ellis, T., and Rosin, P. (2002). Multi view image surveillance and tracking. In *Workshop on Motion and Video Computing*, pages 169–174. IEEE.
- Bleyer, M., Rhemann, C., and Rother, C. (2011). Patchmatch stereo-stereo matching with slanted support windows. In *British Machine Vision Conference*, volume 11, pages 1–11.
- Bogin, B. and Varela-Silva, M. I. (2010). Leg length, body proportion, and health: a review with a note on beauty. *International journal of environmental research and public health*, 7(3):1047–1075.
- Bregler, C., Hertzmann, A., and Biermann, H. (2000). Recovering non-rigid 3d shape from image streams. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 690–696. IEEE.
- Bulbul, A. and Dahyot, R. (2016). Populating virtual cities using social media. *Computer Animation and Virtual Worlds*.
- Burschka, D., Li, M., Ishii, M., Taylor, R. H., and Hager, G. D. (2005). Scale-invariant registration of monocular endoscopic images to CT-scans for sinus surgery. *Medical Image Analysis*, 9(5):413–426.
- Canelhas, D. R. (2017). *Truncated Signed Distance Fields Applied To Robotics*. PhD thesis, Örebro University.
- Cao, S. and Snavely, N. (2012). Learning to match images in large-scale collections. In *European Conference on Computer Vision (ECCV), Workshops and Demonstrations*, pages 259–270. Springer.
- Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *Computer Vision and Pattern Recognition (CVPR)*.
- Chambolle, A. (2004). An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20(1-2):89–97.
- Chambolle, A. (2005). Total variation minimization and a class of binary mrf models. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 136–152. Springer.

- Chandraker, M. (2014a). On shape and material recovery from motion. In *European Conference on Computer Vision*, pages 202–217. Springer.
- Chandraker, M. (2014b). What camera motion reveals about shape with unknown brdf. In *Conference on Computer Vision and Pattern Recognition*, pages 2171–2178.
- Chandraker, M. (2015). The information available to a moving observer on shape with unknown, isotropic brdfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(7):1283–1297.
- Chang, P.-L., Handa, A., Davison, A. J., Stoyanov, D., et al. (2014). Robust real-time visual odometry for stereo endoscopy using dense quadrifocal tracking. In *International Conference on Information Processing in Computer-Assisted Interventions*, pages 11–20. Springer.
- Chang, P.-L., Stoyanov, D., Davison, A. J., et al. (2013). Real-time dense stereo reconstruction using convex optimisation with a cost-volume for image-guided robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 42–49. Springer.
- Chen, L., Tang, W., and John, N. W. (2017). Real-time geometry-aware augmented reality in minimally invasive surgery. *Healthcare Technology Letters*, 4(5):163–167.
- Chen, L., Tang, W., John, N. W., Wan, T. R., and Zhang, J. J. (2018). Slam-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality. *Computer Methods and Programs in Biomedicine*, 158:135–146.
- Cherabier, I., Schonberger, J. L., Oswald, M. R., Pollefeys, M., and Geiger, A. (2018). Learning priors for semantic 3d reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 314–330.
- Chung, A. J., Deligianni, F., Shah, P., Wells, A., and Yang, G.-Z. (2004). Enhancement of visual realism with brdf for patient specific bronchoscopy simulation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 486–493. Springer.
- Chung, A. J., Deligianni, F., Shah, P., Wells, A., and Yang, G.-Z. (2006). Patient-specific bronchoscopy visualization through brdf estimation and disocclusion correction. *IEEE transactions on medical imaging*, 25(4):503–513.
- Collins, T., Pizarro, D., Bartoli, A., Canis, M., and Bourdel, N. (2014). Computer-assisted laparoscopic myomectomy by augmenting the uterus with pre-operative mri data. In *International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 243–248. IEEE.
- Cook, R. L. and Torrance, K. E. (1982). A reflectance model for computer graphics. *ACM Transactions on Graphics*, 1(1):7–24.
- Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Computer Vision and Pattern Recognition (CVPR)*.

- Coughlan, J. M. and Yuille, A. L. (1999). Manhattan world: Compass direction from a single image by bayesian inference. In *International Conference on Computer Vision*, volume 2, pages 941–947. IEEE.
- Courty, N. and Corpetti, T. (2007). Crowd motion capture. *Computer Animation and Virtual Worlds*, 18(4-5):361–370.
- Crandall, D., Owens, A., Snavely, N., and Huttenlocher, D. (2011). Discrete-continuous optimization for large-scale structure from motion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3001–3008. IEEE.
- Curless, B. and Levoy, M. (1996). A volumetric method for building complex models from range images. In *Proceedings of SIGGRAPH '96*, pages 303–312. ACM.
- Curtis, S., Best, A., and Manocha, D. (2016). Menge: A modular framework for simulating crowd movement. *Collective Dynamics*, 1:1–40.
- Deguchi, K. (1996). Shape reconstruction from endoscope image by its shadings. In *International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 321–328. IEEE.
- Deguchi, K., Sasano, T., Arai, H., and Yoshikawa, Y. (1994). 3-d shape reconstruction from endoscope image sequences by the factorization method. In *IAPR Workshop on Machine Vision Applications (MVA)*, pages 455–459.
- Deligianni, F., Chung, A., and Yang, G.-Z. (2004). Patient-specific bronchoscope simulation with pq-space-based 2d/3d registration. *Computer Aided Surgery*, 9(5):215–226.
- Deligianni, F., Chung, A. J., and Yang, G.-Z. (2006). Nonrigid 2-d/3-d registration for patient specific bronchoscopy simulation with statistical shape modeling: Phantom validation. *IEEE Transactions on Medical Imaging*, 25(11):1462–1471.
- Doğan, Y., Demirci, S., Güdükbay, U., and Dibeklioğlu, H. (2018). Augmentation of virtual agents in real crowd videos. *Signal, Image and Video Processing*, pages 1–8.
- dos Santos, T. R., Seitel, A., Kilgus, T., Suwelack, S., Wekerle, A.-L., Kenngott, H., Speidel, S., Schlemmer, H.-P., Meinzer, H.-P., Heimann, T., et al. (2014). Pose-independent surface matching for intra-operative soft-tissue marker-less registration. *Medical Image Analysis*, 18(7):1101–1114.
- Durou, J.-D., Falcone, M., and Sagona, M. (2008). Numerical methods for shape-from-shading: A new survey with benchmarks. *Computer Vision and Image Understanding*, 109(1):22–43.
- Edgcumbe, P., Pratt, P., Yang, G.-Z., Nguan, C., and Rohling, R. (2015). Pico lantern: surface reconstruction and augmented reality in laparoscopic surgery using a pick-up laser projector. *Medical Image Analysis*, 25(1):95–102.
- Egi, H., Hattori, M., Suzuki, T., Sawada, H., Kurita, Y., and Ohdan, H. (2016). The usefulness of 3-dimensional endoscope systems in endoscopic surgery. *Surgical Endoscopy*, 30(10):4562–4568.

- Eigen, D., Puhrsch, C., and Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, pages 2366–2374.
- Fernández, C., Baiget, P., Roca, F. X., and González, J. (2011). Augmenting video surveillance footage with virtual agents for incremental event evaluation. *Pattern Recognition Letters*, 32(6):878–889.
- Fleuret, F., Berclaz, J., Lengagne, R., and Fua, P. (2008). Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282.
- Focken, D. and Stiefelhagen, R. (2002). Towards vision-based 3-d people tracking in a smart room. In *International Conference on Multimodal Interfaces*, pages 400–405. IEEE.
- Forsyth, D. and Zisserman, A. (1991). Reflections on shading. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):671–679.
- Frahm, J.-M., Fite-Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.-H., Dunn, E., Clipp, B., Lazebnik, S., et al. (2010). Building rome on a cloudless day. In *European Conference on Computer Vision (ECCV)*, pages 368–381. Springer.
- Frueh, C. and Zakhor, A. (2003). Constructing 3d city models by merging ground-based and airborne views. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–562. IEEE.
- Furukawa, Y., Curless, B., Seitz, S. M., and Szeliski, R. (2010). Towards internet-scale multi-view stereo. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1434–1441. IEEE.
- Furukawa, Y., Hernández, C., et al. (2015). Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148.
- Furukawa, Y. and Ponce, J. (2010). Accurate, dense, and robust multiview stereopsis. *Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376.
- Gallardo, M., Collins, T., and Bartoli, A. (2016). Using shading and a 3d template to reconstruct complex surface deformations. In *BMVC*.
- Gallardo, M., Collins, T., and Bartoli, A. (2017). Dense non-rigid structure-from-motion and shading with unknown albedos. In *International Conference on Computer Vision (ICCV)*, pages 3884–3892. IEEE.
- Garbey, M., Nguyen, T. B., Huang, A. Y., Fikfak, V., and Dunkin, B. J. (2018). A method for going from 2d laparoscope to 3d acquisition of surface landmarks by a novel computer vision approach. *International Journal of Computer Assisted Radiology and Surgery*, 13(2):267–280.
- Garcia, J. and Quintana-Domeque, C. (2007). The evolution of adult height in europe: a brief note. *Economics & Human Biology*, 5(2):340–349.

- Garg, R., BG, V. K., Carneiro, G., and Reid, I. (2016). Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756. Springer.
- Garg, R., Roussos, A., and Agapito, L. (2013). Dense variational reconstruction of non-rigid surfaces from monocular video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1272–1279. IEEE.
- Garg, R., Seitz, S. M., Ramanan, D., and Snavely, N. (2011). Where’s waldo: Matching people in images of crowds. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1793–1800. IEEE.
- Ge, W. and Collins, R. T. (2010). Crowd detection with a multiview sampler. In *European Conference on Computer Vision (ECCV)*, pages 324–337. Springer.
- Godard, C., Mac Aodha, O., and Brostow, G. J. (2017). Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7.
- Grasa, O. G., Bernal, E., Casado, S., Gil, I., and Montiel, J. (2013). Visual slam for handheld monocular endoscope. *Transactions on Medical Imaging*, 33(1):135–146.
- Grasa, O. G., Civera, J., and Montiel, J. (2011). EKF monocular slam with relocalization for laparoscopic sequences. In *International Conference on Robotics and Automation*, pages 4816–4821. IEEE.
- Gu, Z., Su, X., Liu, Y., and Zhang, Q. (2008). Local stereo matching with adaptive support-weight, rank transform and disparity calibration. *Pattern Recognition Letters*, 29(9):1230–1235.
- Guan, J., Deboeverie, F., Slembrouck, M., Van Haerenborgh, D., Van Cauwelaert, D., Veelaert, P., and Philips, W. (2016). Extrinsic calibration of camera networks based on pedestrians. *Sensors*, 16(5):654.
- Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., and Alahi, A. (2018). Social gan: Socially acceptable trajectories with generative adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)*.
- Haase, S., Forman, C., Kilgus, T., Bammer, R., Maier-Hein, L., and Hornegger, J. (2013). Tof/rgb sensor fusion for 3-d endoscopy. *Current Medical Imaging Reviews*, 9(2):113–119.
- Han, Y., Lee, J.-Y., and Kweon, I. S. (2013). High quality shape from a single rgb-d image under uncalibrated natural illumination. In *International Conference on Computer Vision (ICCV)*.
- Häne, C., Zach, C., Cohen, A., Angst, R., and Pollefeys, M. (2013). Joint 3d scene reconstruction and class segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 97–104.
- Häne, C., Zach, C., Cohen, A., and Pollefeys, M. (2016). Dense semantic 3d reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

- Haouchine, N., Dequidt, J., Peterlik, I., Kerrien, E., Berger, M.-O., and Cotin, S. (2014). Towards an accurate tracking of liver tumors for augmented reality in robotic assisted surgery. In *International Conference on Robotics and Automation (ICRA)*, pages 4121–4126. IEEE.
- Hartley, R. and Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge University Press.
- He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.
- Hedman, P., Philip, J., Price, T., Frahm, J.-M., Drettakis, G., and Brostow, G. (2018). Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 37(6).
- Heinly, J., Schönberger, J. L., Dunn, E., and Frahm, J.-M. (2015). Reconstructing the world* in six days*(as captured by the yahoo 100 million image dataset). In *Computer Vision and Pattern Recognition (CVPR)*.
- Helferty, J. and Higgins, W. (2002). Combined endoscopic video tracking and virtual 3d ct registration for surgical guidance. In *International Conference on Image Processing*, volume 2, pages II–II. IEEE.
- Helferty, J. P., Sherbondy, A. J., Kiraly, A. P., and Higgins, W. E. (2007). Computer-based system for the virtual-endoscopic guidance of bronchoscopy. *Computer Vision and Image Understanding*, 108(1-2):171–187.
- Hödlmoser, M., Micusik, B., and Kampel, M. (2011). Camera auto-calibration using pedestrians and zebra-crossings. In *International Conference on Computer Vision (ICCV) Workshops*, pages 1697–1704. IEEE.
- Hong, D., Tavanapong, W., Wong, J., Oh, J., and De Groen, P. C. (2009). 3d reconstruction of colon segments from colonoscopy images. In *2009 Ninth IEEE International Conference on Bioinformatics and BioEngineering*, pages 53–60. IEEE.
- Hong, D., Tavanapong, W., Wong, J., Oh, J., and De Groen, P. C. (2014). 3d reconstruction of virtual colon structures from colonoscopy images. *Computerized Medical Imaging and Graphics*, 38(1):22–33.
- Horn, B. K. (1970). *Shape from shading: A method for obtaining the shape of a smooth opaque object from one view*. Dissertation, Massachusetts Institute of Technology.
- Horowitz, I. and Kiryati, N. (2004). Depth from gradient fields and control points: Bias correction in photometric stereo. *Image and Vision Computing*, 22(9):681–694.
- Hu, M., Penney, G., Figl, M., Edwards, P., Bello, F., Casula, R., Rueckert, D., and Hawkes, D. (2012). Reconstruction of a 3d surface from video that is robust to missing data and outliers: Application to minimally invasive surgery using stereo and mono endoscopes. *Medical Image Analysis*, 16(3):597–611.

- Huang, R. and Smith, W. A. (2011). Shape-from-shading under complex natural illumination. In *International Conference on Image Processing*, pages 13–16. IEEE.
- Huber, P. J. (1981). *Robust Statistics*. John Wiley and Sons.
- Innmann, M., Kim, K., Gu, J., Niessner, M., Loop, C., Stamminger, M., and Kautz, J. (2019). Nrmvs: Non-rigid multi-view stereo. *arXiv preprint arXiv:1901.03910*.
- Jancosek, M. and Pajdla, T. (2011). Multi-view reconstruction preserving weakly-supported surfaces. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3121–3128. IEEE.
- Ji, D., Dunn, E., and Frahm, J.-M. (2014). 3d reconstruction of dynamic textures in crowd sourced data. In *European Conference on Computer Vision (ECCV)*, pages 143–158. Springer.
- Ji, D., Dunn, E., and Frahm, J.-M. (2016). Spatio-temporally consistent correspondence for dense dynamic scene modeling. In *European Conference on Computer Vision (ECCV)*, pages 3–18. Springer.
- Jin, G., Lee, S.-J., Hahn, J. K., Bielałowicz, S., Mittal, R., and Walsh, R. (2007). Active illumination based 3d surface reconstruction and registration for image guided medialization laryngoplasty. In *Medical Imaging 2007: Visualization and Image-Guided Procedures*, volume 6509, page 650908. International Society for Optics and Photonics.
- Johnson, M. K. and Adelson, E. H. (2011). Shape estimation in natural illumination. In *Computer Vision and Pattern Recognition*, pages 2553–2560. IEEE.
- Junejo, I. and Foroosh, H. (2006). Robust auto-calibration from pedestrians. In *International Conference on Video and Signal Based Surveillance (AVSS)*, pages 92–92. IEEE.
- Kao, C. Y., Osher, S., and Qian, J. (2004). Lax–friedrichs sweeping scheme for static hamilton–jacobi equations. *Journal of Computational Physics*, 196(1):367–391.
- Kaufman, A. and Wang, J. (2008). 3d surface reconstruction from endoscopic videos. In *Visualization in Medicine and Life Sciences*, pages 61–74. Springer.
- Kazhdan, M. and Hoppe, H. (2013). Screened poisson surface reconstruction. *ACM Transactions on Graphics (TOG)*, 32(3):29.
- Kitoh, S.-i., Obi, T., Yamaguchi, M., and Ohyama, N. (1997). Compensation of color shift due to the multiple reflection of illumination in ccd endoscopic color measurement:(1) principle and basic experiment. *Optics Communications*, 143(1-3):102–108.
- Kitoh, S.-i., Tokuoka, H., Hasegawa, J., Nonami, T., Obi, T., Yamaguchi, M., and Ohyama, N. (1998). Topographic measurement of internal surfaces using a sequence of stereo charge-coupled device endoscopic images:(1) method. *Optical Review*, 5(6):352–357.
- Koenderink, J. J., Van Doorn, A. J., and Stavridi, M. (1996). Bidirectional reflection distribution function expressed in terms of surface scattering modes. In *European Conference on Computer Vision (ECCV)*.

- Koppel, D., Chen, C.-I., Wang, Y.-F., Lee, H., Gu, J., Poirson, A., and Wolters, R. (2007). Toward automated model building from video in computer-assisted diagnoses in colonoscopy. In *Medical Imaging 2007: Visualization and Image-Guided Procedures*, volume 6509, page 65091L. International Society for Optics and Photonics.
- Krahnstoeber, N. and Mendonca, P. R. (2005). Bayesian autocalibration for surveillance. In *International Conference on Computer Vision (ICCV)*, volume 2, pages 1858–1865. IEEE.
- Kuhn, A., Price, T., Frahm, J.-M., and Mayer, H. (2017). Down to earth: Using semantics for robust hypothesis selection for the five-point algorithm. In *German Conference on Pattern Recognition (GCPR)*, pages 389–400. Springer, Cham.
- Kusakunniran, W., Li, H., and Zhang, J. (2009). A direct method to self-calibrate a surveillance camera by observing a walking pedestrian. In *Digital Image Computing: Techniques and Applications*, pages 250–255. IEEE.
- Labatut, P., Pons, J.-P., and Keriven, R. (2009). Robust and efficient surface reconstruction from range data. In *Computer graphics forum*, volume 28, pages 2275–2290. Wiley Online Library.
- Lau, W. W., Ramey, N. A., Corso, J. J., Thakor, N. V., and Hager, G. D. (2004). Stereo-based endoscopic tracking of cardiac surface deformation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 494–501. Springer.
- Lerner, A., Chrysanthou, Y., and Lischinski, D. (2007). Crowds by example. In *Computer Graphics Forum*, volume 26, pages 655–664. Wiley Online Library.
- Lewiner, T., Lopes, H., Vieira, A. W., and Tavares, G. (2003). Efficient implementation of marching cubes’ cases with topological guarantees. *Journal of Graphics Tools*, 8(2):1–15.
- Li, Y., Snavely, N., and Huttenlocher, D. P. (2010). Location recognition using prioritized feature matching. In *European Conference on Computer Vision (ECCV)*, pages 791–804. Springer.
- Li, Z. and Snavely, N. (2018). Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2041–2050.
- Li, Z., Xu, Z., Ramamoorthi, R., Sunkavalli, K., and Chandraker, M. (2018). Learning to reconstruct shape and spatially-varying reflectance from a single image. In *SIGGRAPH Asia 2018 Technical Papers*, page 269. ACM.
- Lin, B., Johnson, A., Qian, X., Sanchez, J., and Sun, Y. (2013). Simultaneous tracking, 3d reconstruction and deforming point detection for stereoscope guided surgery. In *Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions*, pages 35–44. Springer.
- Lin, B., Sun, Y., Qian, X., Goldgof, D., Gitlin, R., and You, Y. (2016). Video-based 3d reconstruction, laparoscope localization and deformation recovery for abdominal minimally invasive surgery: a survey. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 12(2):158–178.

- Liu, J., Collins, R. T., and Liu, Y. (2011). Surveillance camera autocalibration based on pedestrian height distributions. In *British Machine Vision Conference (BMVC)*.
- Liu, J., Collins, R. T., and Liu, Y. (2013). Robust autocalibration for a surveillance camera network. In *Workshop on Applications of Computer Vision (WACV)*, pages 433–440. IEEE.
- Lorensen, W. E. and Cline, H. E. (1987). Marching cubes: A high resolution 3d surface construction algorithm. In *Proceedings of SIGGRAPH '87*, volume 21, pages 163–169. ACM.
- Lurie, K. L., Angst, R., Zlatev, D. V., Liao, J. C., and Bowden, A. K. E. (2017). 3d reconstruction of cystoscopy videos for comprehensive bladder records. *Biomedical optics express*, 8(4):2106–2123.
- Lv, F., Zhao, T., and Nevatia, R. (2002). Self-calibration of a camera from video of a walking human. In *International Conference on Pattern Recognition (ICPR)*, volume 1, pages 562–567. IEEE.
- Lv, F., Zhao, T., and Nevatia, R. (2006). Camera calibration from video of a walking human. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1513–1518.
- Maes, P., Darrell, T., Blumberg, B., and Pentland, A. (1995). The ALIVE system: Full-body interaction with autonomous agents. In *Computer Animation '95.*, pages 11–18. IEEE.
- Mahmood, F., Chen, R., Sudarsky, S., Yu, D., and Durr, N. J. (2018). Deep learning with cinematic rendering: fine-tuning deep neural networks using photorealistic medical images. *Physics in Medicine & Biology*, 63(18):185012.
- Mahmood, F. and Durr, N. J. (2018). Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. *Medical Image Analysis*, 48:230–243.
- Mahmoud, N., Collins, T., Hostettler, A., Soler, L., Doignon, C., and Montiel, J. M. M. (2019). Live tracking and dense reconstruction for handheld monocular endoscopy. *Transactions on Medical Imaging*, 38(1):79–89.
- Maier-Hein, L., Groch, A., Bartoli, A., Bodenstedt, S., Boissonnat, G., Chang, P.-L., Clancy, N., Elson, D. S., Haase, S., Heim, E., et al. (2014). Comparative validation of single-shot optical techniques for laparoscopic 3-d surface reconstruction. *IEEE Transactions on Medical Imaging*, 33(10):1913–1930.
- Maier-Hein, L., Mountney, P., Bartoli, A., Elhawary, H., Elson, D., Groch, A., Kolb, A., Rodrigues, M., Sorger, J., Speidel, S., et al. (2013). Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery. *Medical Image Analysis*, 17(8):974–996.
- Malti, A. and Bartoli, A. (2014). Combining conformal deformation and cook–torrance shading for 3-d reconstruction in laparoscopy. *Biomedical Engineering, IEEE Transactions on*, 61(6):1684–1692.

- Malti, A., Bartoli, A., and Collins, T. (2011). Template-based conformal shape-from-motion from registered laparoscopic images. In *Medical Image Understanding and Analysis (MIUA)*.
- Malti, A., Bartoli, A., and Collins, T. (2012). Template-based conformal shape-from-motion-and-shading for laparoscopy. In *Information Processing in Computer-Assisted Interventions (IPCAI)*.
- Marcinczak, J. M. and Grigat, R.-R. (2014). Total variation based 3d reconstruction from monocular laparoscopic sequences. In *International MICCAI Workshop on Computational and Clinical Challenges in Abdominal Imaging*, pages 239–247. Springer.
- Marmol, A., Banach, A., and Peynot, T. (2019). Dense-arthroslam: Dense intra-articular 3-d reconstruction with robust localization prior for arthroscopy. *Robotics and Automation Letters*, 4(2):918–925.
- Marmol, A., Corke, P., and Peynot, T. (2018). Arthroslam: Multi-sensor robust visual localization for minimally invasive orthopedic surgery. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 3882–3889. IEEE.
- Martin-Brualla, R., He, Y., Russell, B. C., and Seitz, S. M. (2014). The 3d jigsaw puzzle: Mapping large indoor spaces. In *European Conference on Computer Vision (ECCV)*, pages 1–16. Springer.
- Matusik, W., Pfister, H., Brand, M., and McMillan, L. (2003). A data-driven reflectance model. *ACM Transactions on Graphics (TOG)*, 22(3):759–769.
- Meshry, M., Goldman, D. B., Khamis, S., Hoppe, H., Pandey, R., Snavely, N., and Martin-Brualla, R. (2019). Neural rerendering in the wild. *Conference on Computer Vision and Pattern Recognition*.
- Micusik, B. and Pajdla, T. (2010). Simultaneous surveillance camera calibration and foot-head homology estimation from human detections. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1562–1569. IEEE.
- Mirota, D., Taylor, R. H., Ishii, M., and Hager, G. D. (2009). Direct endoscopic video registration for sinus surgery. In *Medical Imaging 2009: Visualization, Image-Guided Procedures, and Modeling*, volume 7261, page 72612K. International Society for Optics and Photonics.
- Mori, K., Deguchi, D., Sugiyama, J., Suenaga, Y., Toriwaki, J.-i., Maurer Jr, C. R., Takabatake, H., and Natori, H. (2002). Tracking of a bronchoscope using epipolar geometry analysis and intensity-based image registration of real and virtual endoscopic images. *Medical Image Analysis*, 6(3):321–336.
- Mountney, P., Stoyanov, D., Davison, A., and Yang, G.-Z. (2006). Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 347–354. Springer.
- Mountney, P., Stoyanov, D., and Yang, G.-Z. (2010). Three-dimensional tissue deformation recovery and tracking. *IEEE Signal Processing Magazine*, 27(4):14–24.

- Mountney, P. and Yang, G.-Z. (2009). Dynamic view expansion for minimally invasive surgery using simultaneous localization and mapping. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1184–1187. IEEE.
- Mountney, P. and Yang, G.-Z. (2010). Motion compensated slam for image guided surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 496–504. Springer.
- Münzer, B., Schoeffmann, K., and Böszörményi, L. (2018). Content-based processing and analysis of endoscopic images and videos: A survey. *Multimedia Tools and Applications*, 77(1):1323–1362.
- Musse, S. R., Jung, C. R., Jacques Jr, J. C., and Braun, A. (2007). Using computer vision to simulate the motion of virtual agents. *Computer Animation and Virtual Worlds*, 18(2):83–93.
- Myung, Y.-S., Lee, C.-H., and Tcha, D.-W. (1995). On the generalized minimum spanning tree problem. *Networks*, 26(4):231–241.
- Nayar, S. K., Ikeuchi, K., and Kanade, T. (1991). Shape from interreflections. *International Journal of Computer Vision*, 6(3):173–195.
- Newcombe, R. A., Lovegrove, S. J., and Davison, A. J. (2011). Dtam: Dense tracking and mapping in real-time. In *International Conference on Computer Vision*, pages 2320–2327. IEEE.
- Nunes, A., Maciel, A., Cavazzola, L., and Walter, M. (2017). A laparoscopy-based method for brdf estimation from in vivo human liver. *Medical Image Analysis*, 35:620–632.
- Oda, N., Hasegawa, J., Nonami, T., Yamaguchi, M., and Ohyama, N. (1994). Estimation of the surface topography from monocular endoscopic images. *Optics Communications*, 109(3-4):215–221.
- Oda, N., Hasegawa, J., Nonami, T., Yamaguchi, M., and Ohyama, N. (1995a). Reconstruction of the gastric surface structure using a monocular ccd endoscope. *Optical Review*, 2(2):110–114.
- Oda, N., Hasegawa, J., Nonami, T., Yamaguchi, M., and Ohyama, N. (1995b). Shape measurement from endoscopic images: Determination of dimensional scale factor by a photometric method. *Optical Review*, 2(3):194–198.
- Okatani, T. and Deguchi, K. (1997). Shape reconstruction from an endoscope image by shape from shading technique for a point light source at the projection center. *Computer vision and image understanding*, 66(2):119–131.
- Oren, M. and Nayar, S. K. (1994). Generalization of lambert’s reflectance model. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 239–246. ACM.
- Osher, S. and Fedkiw, R. (2003). *Level set methods and dynamic implicit surfaces*. Springer Science & Business Media.

- Osher, S. and Shu, C.-W. (1991). High-order essentially nonoscillatory schemes for hamilton-jacobi equations. *SIAM Journal on numerical analysis*, 28(4):907–922.
- Otsuka, K. and Mukawa, N. (2004). Multiview occlusion analysis for tracking densely populated objects based on 2-d visual angles. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–I. IEEE.
- Oxholm, G. and Nishino, K. (2015). Shape and reflectance estimation in the wild. *Transactions on Pattern Analysis and Machine Intelligence*, 38(2):376–389.
- Parchami, M. and Mariottini, G.-L. (2014). A comparative study on 3-d stereo reconstruction from endoscopic images. In *Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments*, page 25. ACM.
- Park, H. S., Shiratori, T., Matthews, I., and Sheikh, Y. (2010). 3d reconstruction of a moving point from a series of 2d projections. In *European Conference on Computer Vision (ECCV)*, pages 158–171. Springer.
- Parot, V., Lim, D., González, G., Traverso, G., Nishioka, N. S., Vakoc, B. J., and Durr, N. J. (2013). Photometric stereo endoscopy. *Journal of Biomedical Optics*, 18(7):076017.
- Penne, J., Höller, K., Stürmer, M., Schrauder, T., Schneider, A., Engelbrecht, R., Feußner, H., Schmauss, B., and Hornegger, J. (2009). Time-of-flight 3-d endoscopy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 467–474. Springer.
- Phong, B. T. (1975). Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317.
- Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J., and Koch, R. (2004). Visual modeling with a hand-held camera. *International Journal on Computer Vision*, 59(3):207–232.
- Prados, E. and Faugeras, O. (2005). Shape from shading: a well-posed problem? In *Computer Vision and Pattern Recognition (CVPR)*.
- Prados, E. and Faugeras, O. (2006). Shape from shading. In Paragios, N., Chen, Y., and Faugeras, O. D., editors, *Handbook of mathematical models in computer vision*, pages 375–388. Springer.
- Prados, E. and Soatto, S. (2005). Fast marching method for generic shape from shading. In *Variational, Geometric, and Level Set Methods in Computer Vision (VLSM)*.
- Price, T., Schönberger, J. L., Wei, Z., Pollefeys, M., and Frahm, J.-M. (2018). Augmenting crowd-sourced 3D reconstructions using semantic detections. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1926–1935.
- Qiu, L. and Ren, H. (2018). Endoscope navigation and 3d reconstruction of oral cavity by visual slam with mitigated data scarcity. In *Computer Vision and Pattern Recognition Workshops*, pages 2197–2204.

- Quéau, Y., Mérou, J., Castan, F., Cremers, D., and Durou, J.-D. (2017). A variational approach to shape-from-shading under natural illumination. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 342–357. Springer.
- Rai, L. and Higgins, W. E. (2006). Image-based rendering method for mapping endoscopic video onto ct-based endoluminal views. In *Medical Imaging 2006: Visualization, Image-Guided Procedures, and Display*, volume 6141, page 614103. International Society for Optics and Photonics.
- Reichard, D., Bodenstedt, S., Suwelack, S., Mayer, B., Preukschas, A., Wagner, M., Kenngott, H., Müller-Stich, B., Dillmann, R., and Speidel, S. (2015). Intraoperative on-the-fly organ-mosaicking for laparoscopic surgery. *Journal of Medical Imaging*, 2(4):045001.
- Reichard, D., Bodenstedt, S., Suwelack, S., Wagner, M., Kenngott, H., Müller-Stich, B. P., Dillmann, R., and Speidel, S. (2016). Robust endoscopic pose estimation for intraoperative organ-mosaicking. In *Medical Imaging 2016: Image Processing*, volume 9784, page 97841Q. International Society for Optics and Photonics.
- Reiter, A., Léonard, S., Sinha, A., Ishii, M., Taylor, R. H., and Hager, G. D. (2016). Endoscopic-ct: learning-based photometric reconstruction for endoscopic sinus surgery. In *Medical Imaging 2016: Image Processing*, volume 9784, page 978418. International Society for Optics and Photonics.
- Röhl, S., Bodenstedt, S., Suwelack, S., Kenngott, H., Mueller-Stich, B. P., Dillmann, R., and Speidel, S. (2011). Real-time surface reconstruction from stereo endoscopic images for intraoperative registration. In *Medical Imaging 2011: Visualization, Image-Guided Procedures, and Modeling*, volume 7964, page 796414. International Society for Optics and Photonics.
- Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268.
- Rusinkiewicz, S. and Levoy, M. (2001). Efficient variants of the icp algorithm. In *3-D Digital Imaging and Modeling*.
- Russell, C., Yu, R., and Agapito, L. (2014). Video pop-up: Monocular 3d reconstruction of dynamic scenes. In *European conference on computer vision*, pages 583–598. Springer.
- Salzmann, M. and Fua, P. (2010). Deformable surface 3d reconstruction from monocular images. *Synthesis Lectures on Computer Vision*, 2(1):1–113.
- Schönberger, J. L. and Frahm, J.-M. (2016). Structure-from-motion revisited. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113.
- Schönberger, J. L., Zheng, E., Pollefeys, M., and Frahm, J.-M. (2016). Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*.
- Shan, Q., Adams, R., Curless, B., Furukawa, Y., and Seitz, S. M. (2013). The visual turing test for scene reconstruction. In *3D Vision (3DV)*, pages 25–32. IEEE.

- Shoji, H., Mori, K., Sugiyama, J., Suenaga, Y., Toriwaki, J.-i., Takabatake, H., and Natori, H. (2001). Camera motion tracking of real endoscope by using virtual endoscopy system and texture information. In *Medical Imaging 2001: Physiology and Function from Multidimensional Images*, volume 4321, pages 122–134. International Society for Optics and Photonics.
- Shu, C.-W. (2007). High order numerical methods for time dependent hamilton-jacobi equations. In *Mathematics and computation in imaging science and information processing*, pages 47–91. World Scientific.
- Sinha, A., Liu, X., Reiter, A., Ishii, M., Hager, G. D., and Taylor, R. H. (2018). Endoscopic navigation in the absence of ct imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 64–71. Springer.
- Snavely, N., Seitz, S. M., and Szeliski, R. (2006). Photo tourism: exploring photo collections in 3d. In *ACM transactions on graphics (TOG)*, volume 25, pages 835–846. ACM.
- Snavely, N., Seitz, S. M., and Szeliski, R. (2008). Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80(2):189–210.
- Song, J., Wang, J., Zhao, L., Huang, S., and Dissanayake, G. (2016). 3d shape recovery of deformable soft-tissue with computed tomography and depth scan. In *Australasian Conference on Robotic Automation*, pages 117–126.
- Song, J., Wang, J., Zhao, L., Huang, S., and Dissanayake, G. (2018). Mis-slam: Real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing. *Robotics and Automation Letters*, 3(4):4068–4075.
- Stoyanov, D. (2012). Surgical vision. *Annals of biomedical engineering*, 40(2):332–345.
- Stoyanov, D., Scarzanella, M. V., Pratt, P., and Yang, G.-Z. (2010). Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 275–282. Springer.
- Subramanian, S., Özaltın, E., and Finlay, J. E. (2011). Height of nations: a socioeconomic analysis of cohort differences and patterns among women in 54 low-to middle-income countries. *PLoS One*, 6(4):e18962.
- Sun, D., Liu, J., Linte, C. A., Duan, H., and Robb, R. A. (2013). Surface reconstruction from tracked endoscopic video using the structure from motion approach. In *Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions*, pages 127–135. Springer.
- Tankus, A., Sochen, N., and Yeshurun, Y. (2005). Shape-from-shading under perspective projection. *International Journal of Computer Vision*, 63(1):21–43.
- Thalmann, N. M. and Thalmann, D. (1997). Animating virtual actors in real environments. *Multimedia Systems*, 5(2):113–125.

- The World Bank Group (2017). Population, female (% of total). <https://data.worldbank.org/indicator/SP.POP.TOTL.FE.ZS>. Accessed: 2017-11-15.
- Thormahlen, T., Broszio, H., and Meier, P. N. (2002). Three-dimensional endoscopy. In *Falk Symposium*, pages 199–214. Kluwer Academic Publishers.
- Tokgozoglu, H. N., Meisner, E. M., Kazhdan, M., and Hager, G. D. (2012). Color-based hybrid reconstruction for endoscopy. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*.
- Totz, J., Thompson, S., Stoyanov, D., Gurusamy, K., Davidson, B. R., Hawkes, D. J., and Clarkson, M. J. (2014). Fast semi-dense surface reconstruction from stereoscopic video in laparoscopic surgery. In *International Conference on Information Processing in Computer-Assisted Interventions*, pages 206–215. Springer.
- Trocoli, T. and Oliveira, L. (2016). Using the scene to calibrate the camera. In *SIBGRAPI Conference on Graphics, Patterns and Images*, pages 455–461. IEEE.
- Tsai, P.-S. and Shah, M. (1994). Shape from shading using linear approximation. *Image and Vision Computing*, 12(8):487–498.
- Turan, M., Almalioglu, Y., Araujo, H., Konukoglu, E., and Sitti, M. (2017). A non-rigid map fusion-based direct slam method for endoscopic capsule robots. *International Journal of Intelligent Robotics and Applications*, 1(4):399–409.
- Turan, M., Almalioglu, Y., Araujo, H., Konukoglu, E., and Sitti, M. (2018). Deep EndoVO: A recurrent convolutional neural network (rcnn) based visual odometry approach for endoscopic capsule robots. *Neurocomputing*, 275:1861–1870.
- Ummenhofer, B. and Brox, T. (2015). Global, dense multiscale reconstruction for a billion points. In *International Conference on Computer Vision (ICCV)*, pages 1341–1349.
- Vagvolgyi, B., Su, L.-M., Taylor, R., and Hager, G. D. (2008). Video to ct registration for image overlay on solid organs. *Proc. Augmented Reality in Medical Imaging and Augmented Reality in Computer-Aided Surgery (AMIARCS)*, pages 78–86.
- Visentini-Scarzanella, M., Hanayama, T., Masutani, R., Yoshida, S., Kominami, Y., Sanomura, Y., Tanaka, S., Furukawa, R., and Kawasaki, H. (2015). Tissue shape acquisition with a hybrid structured light and photometric stereo endoscopic system. In *Computer-Assisted and Robotic Endoscopy*, pages 46–58. Springer.
- Visentini-Scarzanella, M., Stoyanov, D., and Yang, G.-Z. (2012). Metric depth recovery from monocular images using shape-from-shading and specularities. In *International Conference on Image Processing (ICIP)*.
- Visentini-Scarzanella, M., Sugiura, T., Kaneko, T., and Koto, S. (2017). Deep monocular 3d reconstruction for assisted navigation in bronchoscopy. *International Journal of Computer Assisted Radiology and Surgery*, 12(7):1089–1099.

- Vogel, O., Breuß, M., Leichtweis, T., and Weickert, J. (2009). Fast shape from shading for phong-type surfaces. In *Scale Space and Variational Methods in Computer Vision (SSVM)*.
- Waechter, M., Moehrle, N., and Goesele, M. (2014). Let there be color! large-scale texturing of 3d reconstructions. In *European Conference on Computer Vision*, pages 836–850. Springer.
- Wang, C., Cheikh, F. A., Kaaniche, M., and Elle, O. J. (2018). Liver surface reconstruction for image guided surgery. In *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10576, page 105762H. International Society for Optics and Photonics.
- Wang, H., Mirota, D., Ishii, M., and Hager, G. D. (2008). Robust motion estimation and structure recovery from endoscopic image sequences with an adaptive scale kernel consensus estimator. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7. IEEE.
- Wang, R., Pizer, S. M., and Frahm, J.-M. (2019). Recurrent neural network for (un-)supervised learning of monocular video visual odometry and depth. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5555–5564.
- Wang, R., Price, T., Zhao, Q., Frahm, J.-M., Rosenman, J., and Pizer, S. (2017). Improving 3D surface reconstruction from endoscopic video via fusion and refined reflectance modeling. In *SPIE Medical Imaging*, pages 101330B–101330B. International Society for Optics and Photonics.
- Wang, X. (2013). Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters*, 34(1):3–19.
- Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional pose machines. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Weisstein, E. W. (2019). Cubic Formula. From MathWorld—A Wolfram Web Resource. <http://mathworld.wolfram.com/CubicFormula.html>. Accessed: 2019-05-12.
- Wilson, K. and Snavely, N. (2014). Robust global translations with ldsfm. In *European Conference on Computer Vision (ECCV)*.
- Wu, C. (2013). Towards linear-time incremental structure from motion. In *3D Vision (3DV)*, pages 127–134. IEEE.
- Wu, C., Narasimhan, S. G., and Jaramaz, B. (2010). A multi-image shape-from-shading framework for near-lighting perspective endoscopes. *International Journal of Computer Vision*, 86(2-3):211–228.
- Wu, C., Varanasi, K., Liu, Y., Seidel, H.-P., and Theobalt, C. (2011). Shading-based dynamic shape refinement from multi-view video under general illumination. In *International Conference on Computer Vision (ICCV)*.
- Xiao, J., Chai, J.-x., and Kanade, T. (2004). A closed-form solution to non-rigid shape and motion recovery. In *European conference on computer vision*, pages 573–587. Springer.

- Yeung, S., Tsui, H.-T., and Yim, A. (1999). Global shape from shading for an endoscope image. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 318–327. Springer.
- Yoon, K.-J. and Kweon, I. S. (2006). Adaptive support-weight approach for correspondence search. *Pattern Analysis and Machine Intelligence*, 28(4):650–656.
- Yu, F. and Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. In *ICLR*.
- Zach, C. (2008). Fast and high quality fusion of depth maps. In *International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, volume 1.
- Zach, C., Pock, T., and Bischof, H. (2007). A globally optimal algorithm for robust TV-L1 range image integration. In *International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE.
- Zampokas, G., Tsiolis, K., Peleka, G., Mariolis, I., Malasiotis, S., and Tzovaras, D. (2018). Real-time 3d reconstruction in minimally invasive surgery with quasi-dense matching. In *International Conference on Imaging Systems and Techniques (IST)*, pages 1–6. IEEE.
- Zhang, R., Tsai, P.-S., Cryer, J. E., and Shah, M. (1999). Shape-from-shading: a survey. *Pattern Analysis and Machine Intelligence*, 21(8):690–706.
- Zhao, Q. (2017). *Surface Registration for Pharyngeal Radiation Treatment Planning*. PhD thesis, The University of North Carolina at Chapel Hill.
- Zhao, Q., Price, T., Pizer, S., Niethammer, M., Alterovitz, R., and Rosenman, J. (2015). Surface registration in the presence of missing patches and topology change. *Medical Image Understanding and Analysis (MIUA)*.
- Zhao, Q., Price, T., Pizer, S., Niethammer, M., Alterovitz, R., and Rosenman, J. (2016). The endoscopogram: a 3D model reconstructed from endoscopic video frames. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer International Publishing.
- Zheng, E., Ji, D., Dunn, E., and Frahm, J.-M. (2015). Sparse dynamic 3d reconstruction from unsynchronized videos. In *International Conference on Computer Vision (ICCV)*, pages 4435–4443.
- Zheng, E., Wang, K., Dunn, E., and Frahm, J.-M. (2014). Joint object class sequencing and trajectory triangulation (JOST). In *European Conference on Computer Vision (ECCV)*.
- Zheng, Q. and Chellappa, R. (1991). Estimation of illuminant direction, albedo, and shape from shading. In *Computer Vision and Pattern Recognition (CVPR)*, pages 540–545. IEEE.
- Zhou, T., Brown, M., Snavely, N., and Lowe, D. G. (2017). Unsupervised learning of depth and ego-motion from video. In *CVPR*, volume 2, page 7.
- Zollhöfer, M., Dai, A., Innmann, M., Wu, C., Stamminger, M., Theobalt, C., and Nießner, M. (2015). Shading-based refinement on volumetric signed distance functions. *ACM Transactions on Graphics (TOG)*, 34(4).