Endoscopic-CT: Learning-Based Photometric Reconstruction for Endoscopic Sinus Surgery

A. Reiter[1], S. Leonard[1], **A. Sinha**[1], M. Ishii[2], R. H. Taylor[1], and G. D. Hager[1]

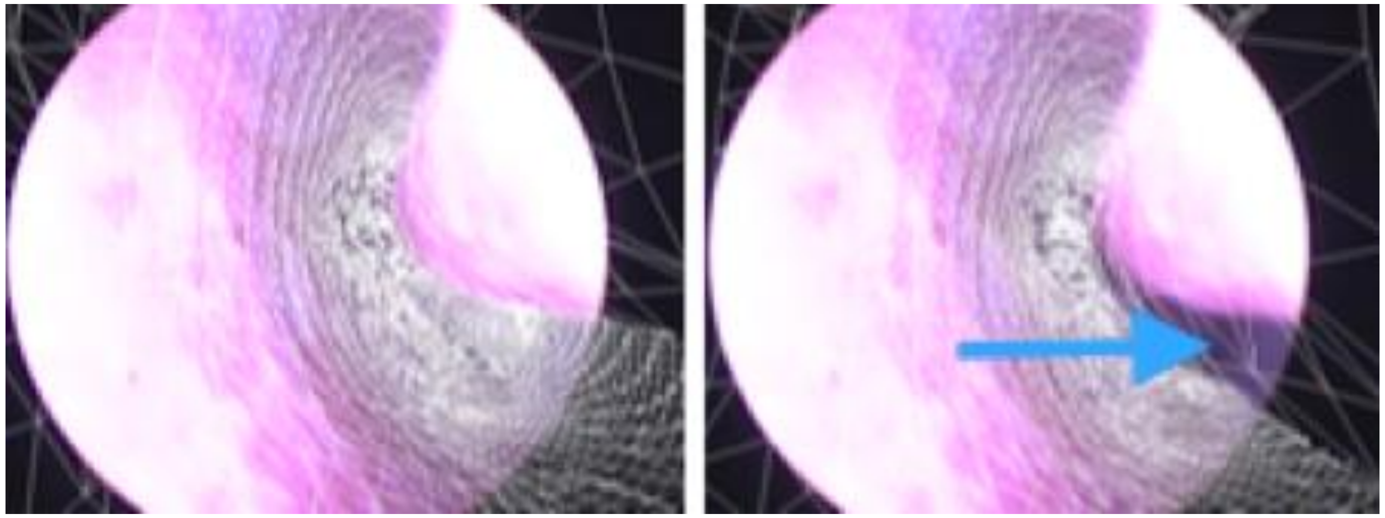
[1]Johns Hopkins University, Dept. of Computer Science, Baltimore, MD USA

[2]Johns Hopkins Medical Institutions, Dept. of Otolaryngology – Head and Neck Surgery, Baltimore, MD USA

SPIE Medical Imaging
Feb. 27 – Mar 3, 2016
San Diego, CA USA

Functional Endoscopic Sinus Surgery (FESS)

- Sinus surgery typically performed under endoscopic guidance
- Large percentage employ surgical navigation
- Very critical and delicate anatomy requires high precision
- We developed Video-CT registration that outperforms traditional navigation (~ 2mm ➔ ≤ 1.0mm)

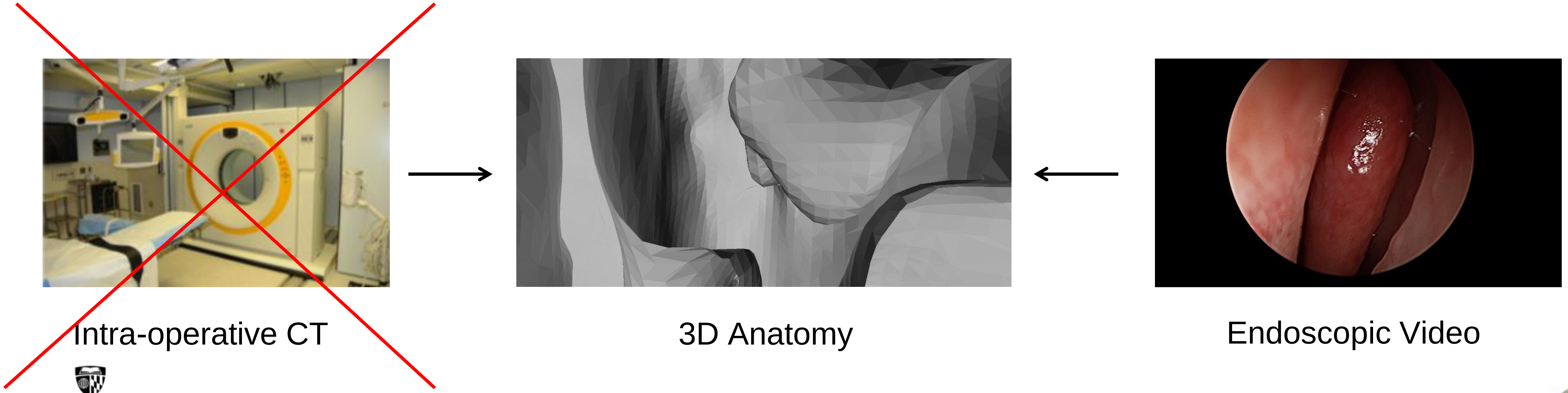


A comparison of our Video-CT registration (left) and traditional navigation using Optotrak* (right). The arrow indicates an obvious error in the latter.

*http://www.ndigital.com/msci/products/optotrak-certus/

Beyond Navigation

- Reconstruction also important for in-situ FESS
- Corresponding (surgically) disturbed anatomy to pre-op CT becomes challenging
- Can perform intra-op CT, but risks exposing patient to additional radiation (e.g., situational awareness, metrology, etc)
- This work presents ***Endoscopic-CT***: video-based dense reconstruction using video to take place of intra-op CT



~~Intra-operative CT~~

3D Anatomy

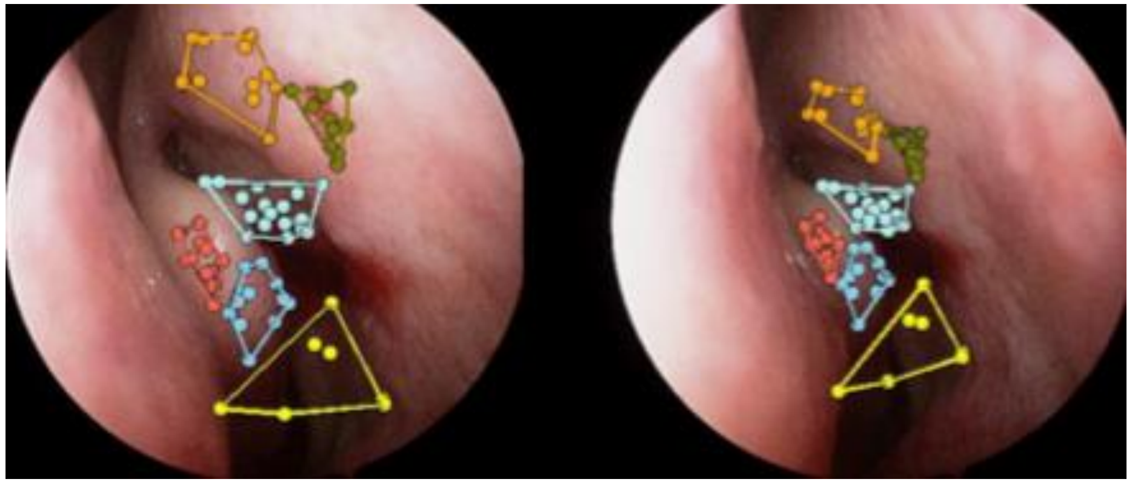
Endoscopic Video

Paper Overview

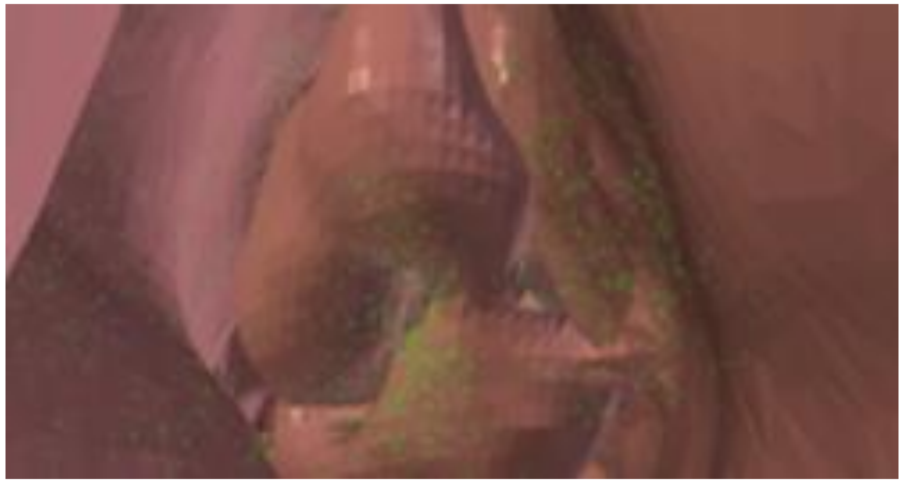
- Structure-from-Motion
- Light and Surface Geometry
- Training Process for Reconstruction
- Results
- Conclusions

Structure-from-Motion

- Our methodology relies on gathering data from Structure-from-Motion (SfM)
- Estimate 3-dimensional “structure” of a scene using a series of images
  - Also recover camera geometry (positions and orientations)
- Relate 3D scene points to colored 2D pixels across several images (important for training later on!)



Hierarchical Multi-Affine (HMA) Matching for Low-Textured, Robust Feature Matching



3D point cloud (green) generation + Trimmed-ICP yields Video-CT Registration

SEE OUR VIDEO-CT PAPER HERE AT SPIE 2016!

Light and Surface Geometry

- Due to low-texture, difficult to reconstruct *densely* (e.g., at all/most points) using traditional feature-based approaches
- Instead exploit light reflectance properties
- **Bidirectional Reflectance Distribution Function** (BRDF): relates incoming light, viewing direction, surface normal direction, and reflectance radiance
  - If modeled accurately, fully describes scene geometry from pixel values
- Most use *Lambertian* Assumption (light reflected equally in all directions)
  - Not really true for surgical data (e.g., tissue absorption, scattering, liquids, etc)

Light and Surface Geometry

The BRDF is a 4-dimensional function. Lambertian example:

Surface property

Measured from image

$$I_R = \frac{\rho}{\pi} L(\omega_i) \frac{\cos(\theta_i)}{r^2}$$

Encodes surface geometry

Scene Depth

where:

I

$\boldsymbol{I_R}$: reflectance
ρ: diffuse albedo
$L(\omega_i)$: light source radiance onto surface at x
θ_i: angle between surface normal $n(x)$ and light direction ω_i
r: distance between light source and surface point x

Training

- We note that SfM yields a set of 3D points on the anatomy and associated colored 2D pixel locations from several images
- Use this to *train* a general non-linear regressor to estimate the *Inverse-BRDF*. (Inverse lighting is an ill-formed problem; more unknowns than observations)
    - We assume a *fixed* lighting direction (b/c camera fixed to imaging source)
    - We assume a *fixed* surface albedo (not completely correct, but used as an approximation we will relax with future work)
    - All scene geometry defined w.r.t. camera coordinates
- Therefore we reduce the problem to regressing the following function using SfM as training data (we get multiple views of the same 3D points, which gives a sense of differences in shading w.r.t to camera, since light follows camera!):

$$f(u,v,r,g,b)=[z,n_x,n_y,n_z]$$

Training

$$f(u, v, r, g, b) = [z, n_x, n_y, n_z]$$

(u, v): pixel position

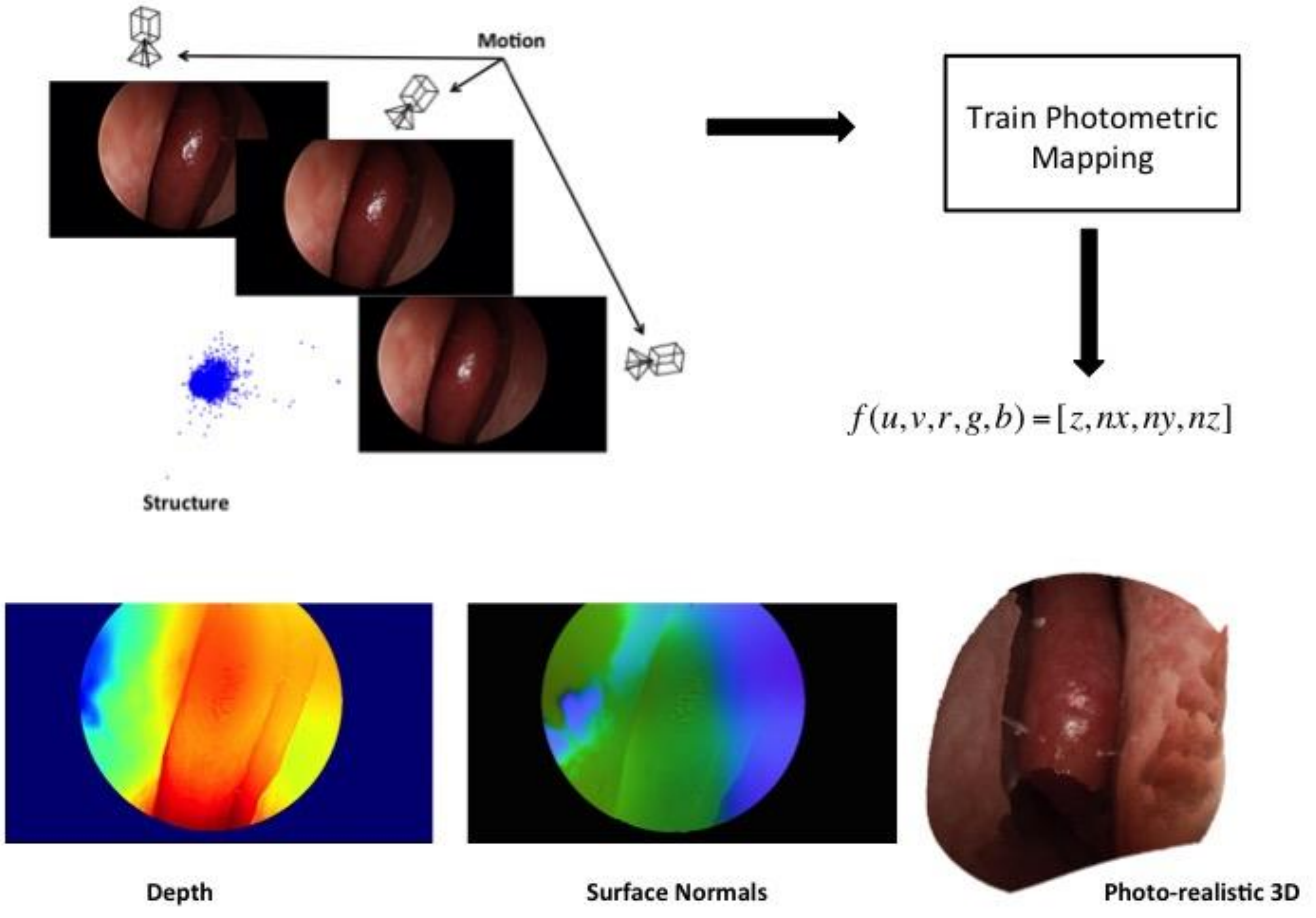
(r, g, b): red-green-blue color at pixel position (u, v)

z: depth of scene point corresponding to pixel position (u, v)

(n_x, n_y, n_z): unit surface normal vector corresponding to pixel position (u, v)

Because *f* is unknown, we train a 3-layer neural network to regress from the training data.

System Overview



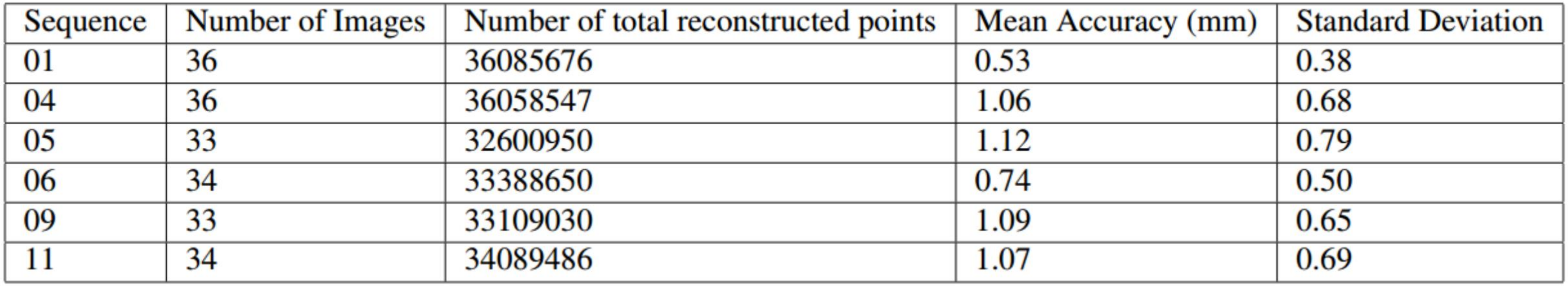
Train Photometric Mapping

$$f(u,v,r,g,b)=[z,nx,ny,nz]$$

Experiments and Results

- **Training**
  - 103,665 SfM points from 36 images to train the regressor
  - Image resolution 1920x1080
  - Train/Validate split: 77,748/25,917 (75%/25%)
  - Training validation error: 0.36mm in depth and 29.5° in surface normal error
- **Testing**:
  - 6 independent test sequences (separate areas of sinus anatomy from training, to demonstrate *local* robustness)
  - With “clean” anatomy (less liquids), obtain average depth error as low as 0.53mm.
  - With more liquids present, depth error increases to as high as 1.12mm

Experiments and Results



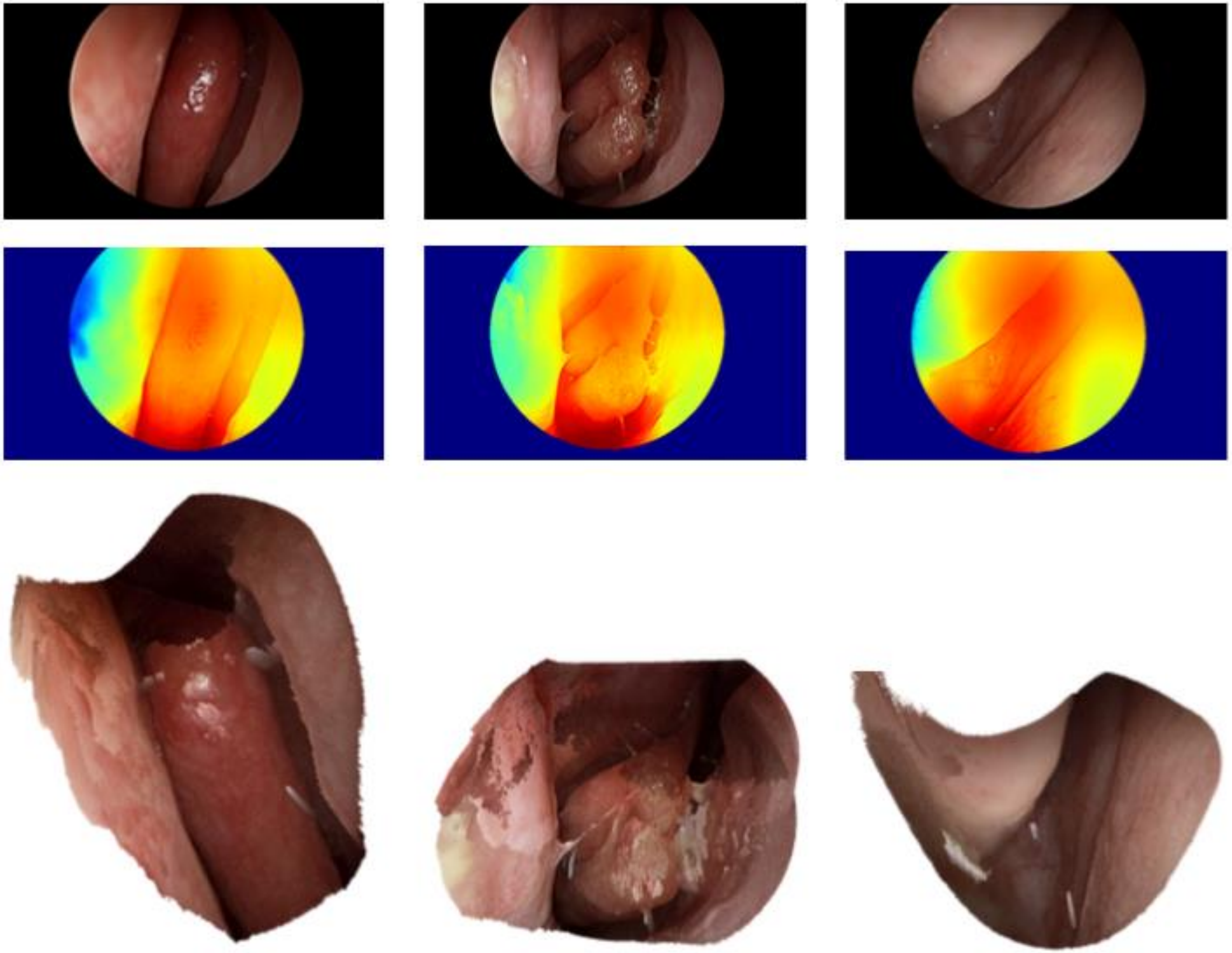
Sequence	Number of Images	Number of total reconstructed points	Mean Accuracy (mm)	Standard Deviation
01	36	36085676	0.53	0.38
04	36	36058547	1.06	0.68
05	33	32600950	1.12	0.79
06	34	33388650	0.74	0.50
09	33	33109030	1.09	0.65
11	34	34089486	1.07	0.69

Table 1. Table of mean accuracy and associated standard deviation in 3D position across several sequences

206 Total Images across 6 different “sequences” (each sequence focuses on a different non-overlapping part of the Sinus anatomy)

For each sequence, total points reconstructed *per-image*, registered to CT through SfM+ICP for evaluation (average distance of predicted 3D point to closest triangle in CT mesh)

Experiments and Results



Color

Depth

Photorealistic 3D

Conclusions

- Presented method for estimating inverse lighting model *per-patient* (meant to be re-trained for each patient individually, on-the-fly)
- Though constant albedo assumption is not correct, results show the variation in albedo is minimal across tissue
- High accuracy 3D reconstruction that matches CT accurately.
- Future Work:
  - Relax albedo assumption
  - Improve surface normal accuracy
  - Learn a *prior* model from a collection of patients to improve per-patient regression

THANK YOU!

Questions/Comments?

Work funded by NIH 5R01EB015530