Towards the next generation of image guidance for endoscopic procedures

CVPR Workshop on 3D Computer Vision in Medical Environments

June 16th 2019

Mathias Unberath, PhD

Assistant Research Professor

Department of Computer Science

Johns Hopkins University

Xingtong Liu
Graduate Student
Department of Computer Science

Ayushi Sinha, PhD
Assistant Research Scientist
Computational Sensing and Robotics

Russell H Taylor, PhD
John C. Malone Professor
Department of Computer Science

Gregory Hager, PhD
Mandell Bellmore Professor
Department of Computer Science

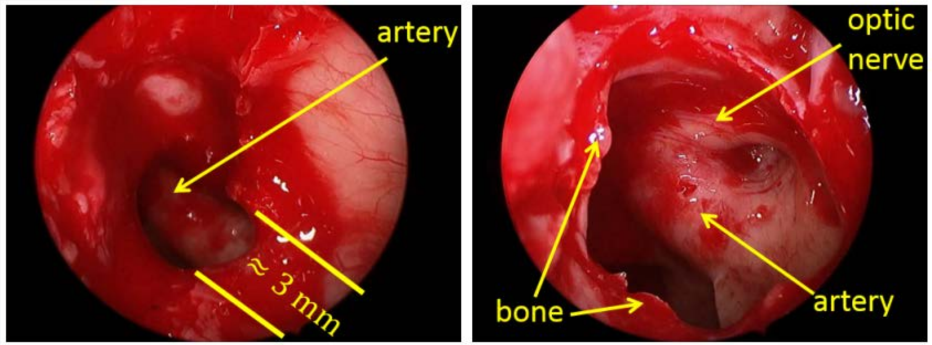
Masaru Ishii, MD
Associate Professor
Department of Otolaryngology

Some Background: Clinical and Technical

Navigating Sinus Surgery

Endoscopic Sinus Surgery

- Functional sinus surgery
  - Close proximity to critical structures
  - Surgical navigation desired



Challenges of Conventional Navigation

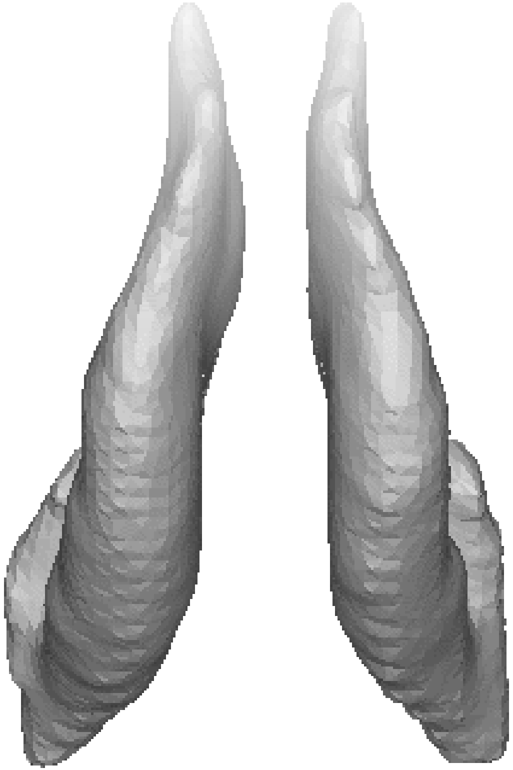
- Patient-specific 3D model of anatomy
  - Pre-operative (potentially outdated)
  - Obtained from CT scan (usually)

- Intra-operative registration: Optical tracking
  - CT to marker (via surface digitization)
  - Endoscope / tool to anatomy
  - → Line of sight constraints
  - → Visualization on model

Challenges of Conventional Navigation

- Patient-specific 3D model of anatomy
  - Pre-operative (potentially outdated)
  - Obtained from CT scan (usually)
- Intra-operative registration: Optical tracking
  - CT to marker (via surface digitization)
  - Endoscope / tool to anatomy
  - → Line of sight constraints
  - → Visualization on model
- Observations
  - Complex setups increase procedure time
  - Disruptive workflows promote frustration
  - → **Where to innovate?**

Step 1: Navigating in the Absence of CT



- Patient-specific 3D model of anatomy
  - ~~Pre-operative (potentially outdated)~~
  - ~~Obtained from CT scan (usually)~~
  - → **Population-derived atlas of sinus anatomy**
- Intra-operative registration: Optical tracking
  - ~~CT to marker (via surface digitization)~~
  - → **Model to video registration**
  - Endoscope / tool to anatomy
  - → Line of sight constraints
  - → Visualization on model

Step 2: Navigating Without Prior Information

ent-specific 3D model of anatomy

~~re-operative (potentially outdated)~~

~~btained from CT scan (usually)~~

econstructed from endoscopy sequence

-operative registration: Optical tracking

~~T to marker (via surface digitization)~~

~~ndoscope / tool to anatomy~~

ine of sight constraints

~~isualization on model~~

verything relative to endoscopy

Towards Next-generation Image Guidance

Navigating in the Absence of CT

Building the Population-based Model

- Build statistical shape models
  - Principal component analysis
  - Capture anatomical variation

- Given shapes, $\mathbf{V} = [\mathbf{v}_1 \ \mathbf{v}_2 \ldots \mathbf{v}_{\mathbf{n}_v}]^\mathbf{T}$ with correspondences, we can compute:

Mean: $\bar{\mathbf{V}} = \frac{1}{n_\mathbf{s}} \sum_{i=1}^{n_\mathbf{s}} \mathbf{V}_i$

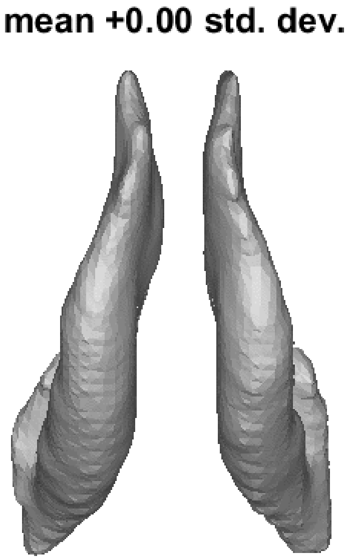
Variance: $\mathbf{\Sigma} = \frac{1}{n_\mathbf{s}} \sum_{i=1}^{n_\mathbf{s}} (\mathbf{V}_i - \bar{\mathbf{V}})(\mathbf{V}_i - \bar{\mathbf{V}})^\mathbf{T}$

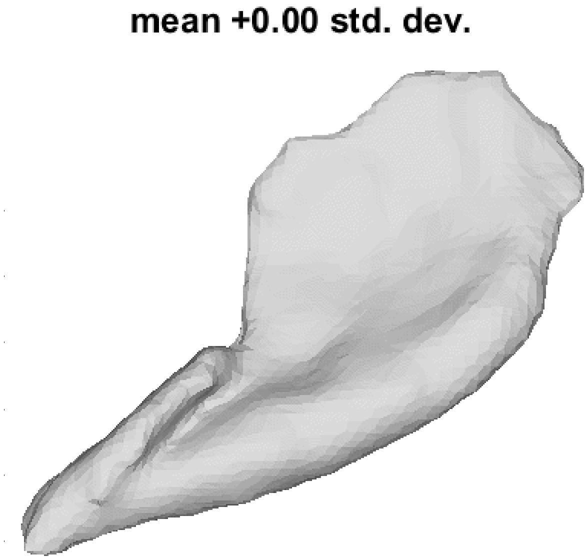
$$\mathbf{\Sigma} = [\mathbf{m}_1 \ldots \mathbf{m}_{n_\mathbf{s}}] \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{n_\mathbf{s}} \end{bmatrix} [\mathbf{m}_1 \ldots \mathbf{m}_{n_\mathbf{s}}]^\mathbf{T}$$

Sinha, A., Liu, X., Reiter, A., Ishii, M., Hager, G. D., & Taylor, R. H. (2018, September). Endoscopic navigation in the absence of CT imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 64-71). Springer, Cham.

Building the Population-based Model

- Build statistical shape models
  - Principal component analysis
  - **Capture anatomical variation (middle turbinate)**





Sinha, A., Liu, X., Reiter, A., Ishii, M., Hager, G. D., & Taylor, R. H. (2018, September). Endoscopic navigation in the absence of CT imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 64-71). Springer, Cham.

Estimating Patient Anatomy

- Deformable registration
  - **Optimize shape model model parameters**
  - Align with endoscopic video

- Given a new shape $\mathbf{V}^*$, we can compute:

Weights: $s_i = \mathbf{w}_i^{\mathrm{T}}(\mathbf{V}^* - \bar{\mathbf{V}})$

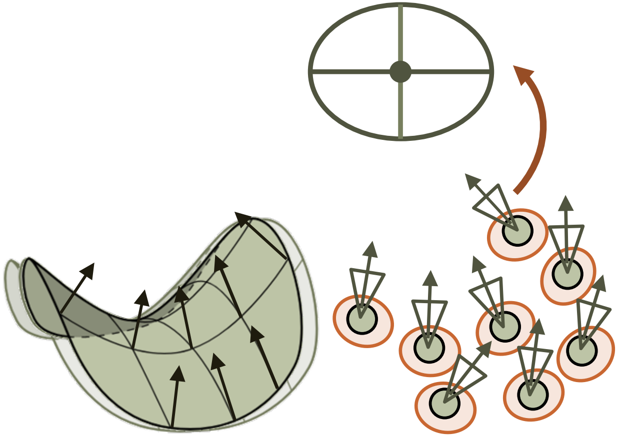
$$\mathbf{w}_i = \sqrt{\lambda_i}\mathbf{m}_i$$

Estimated shape: $\tilde{\mathbf{V}}^* = \bar{\mathbf{V}} + \sum_{i=1}^{n_{\mathbf{m}}} s_i \mathbf{w}_i$

Sinha, A., Liu, X., Reiter, A., Ishii, M., Hager, G. D., & Taylor, R. H. (2018, September). Endoscopic navigation in the absence of CT imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 64-71). Springer, Cham.

Estimating Patient Anatomy

- Deformable registration
  - **Optimize shape model model parameters**
  - **Align with endoscopic video**



- Simultaneously, align rigidly

  Can be solved with the
  Generalized Deformable Most Likely Oriented Point (GD-IMLOP) algorithm

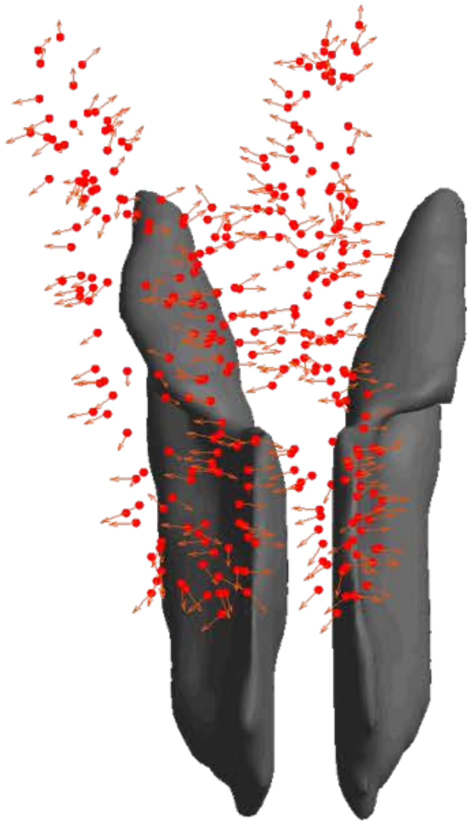
Sinha, A., Liu, X., Reiter, A., Ishii, M., Hager, G. D., & Taylor, R. H. (2018, September). Endoscopic navigation in the absence of CT imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 64-71). Springer, Cham.

Estimating Patient Anatomy

- Deformable registration
  - Optimize shape model model parameters
  - Align with endoscopic video

- Simultaneous deformable and rigid alignment to unseen shape $\mathbf{V}^*$

- Great!



Sinha, A., Liu, X., Reiter, A., Ishii, M., Hager, G. D., & Taylor, R. H. (2018, September). Endoscopic navigation in the absence of CT imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 64-71). Springer, Cham.

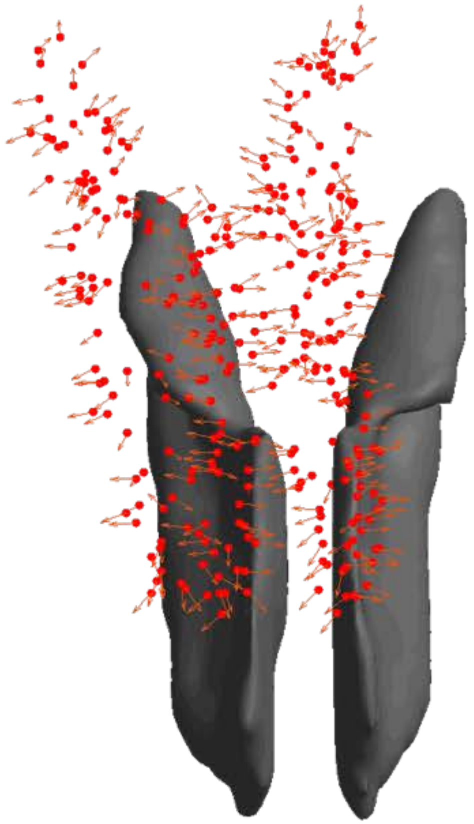
Estimating Patient Anatomy

- Deformable registration
  - Optimize shape model model parameters
  - Align with endoscopic video

- Simultaneous deformable and rigid alignment to unseen shape $\mathbf{V}^*$

- Great!

- **But wait …**
  Where do we get the new shape from?
  How does this link to endoscopy?



Sinha, A., Liu, X., Reiter, A., Ishii, M., Hager, G. D., & Taylor, R. H. (2018, September). Endoscopic navigation in the absence of CT imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 64-71). Springer, Cham.

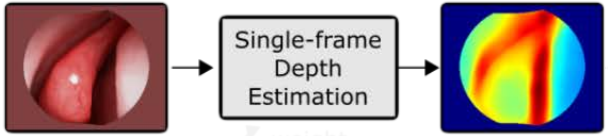
Estimating Patient Anatomy

- Deformable registration
  - Optimize shape model model parameters
  - Align with endoscopic video

- Estimating unseen shapes $\mathbf{V}^*$ from endoscopic video

… some AI maybe?

Training Phase



This is what we are after here

Endoscopic image in → Depth map out

ConvNets are trained via backpropagation

→ Need informative gradients

→ Consequently, need informative loss

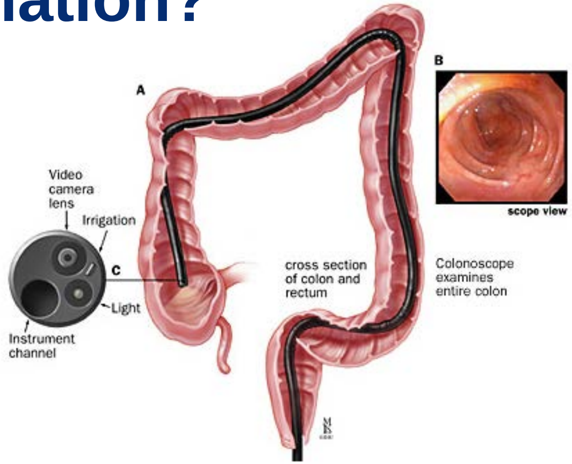
→ **How to supervise learning?**

How to supervise monocular depth estimation?

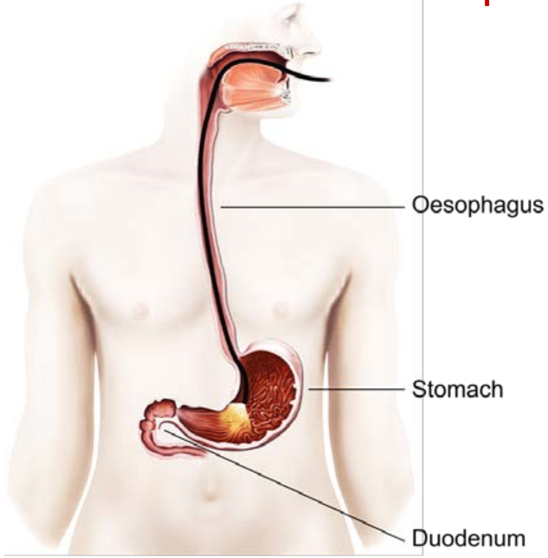
Remembering the application: **Endoscopy**

→ **Miniaturized equipment** to inspect difficult to access anatomy

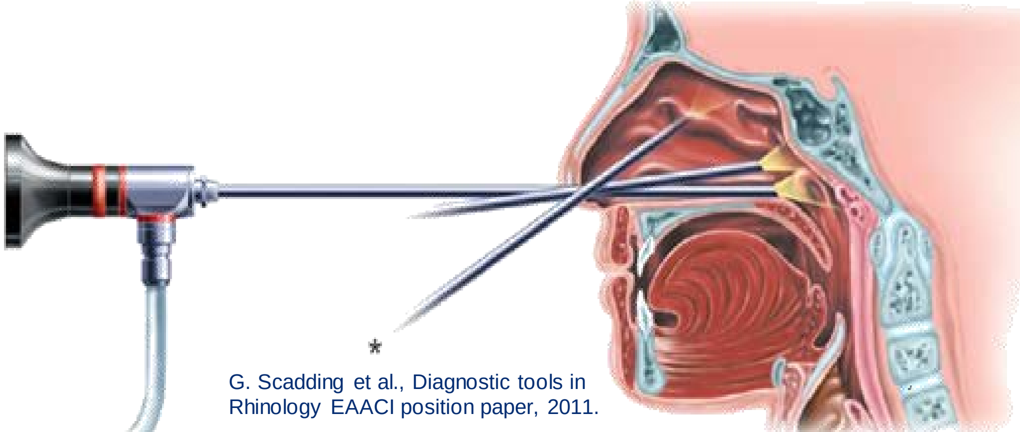
→ Prohibitively disruptive to install dedicated hardware, both stereo setup or depth sensing



http://www.alfasurgerycenter.com/procedures.html



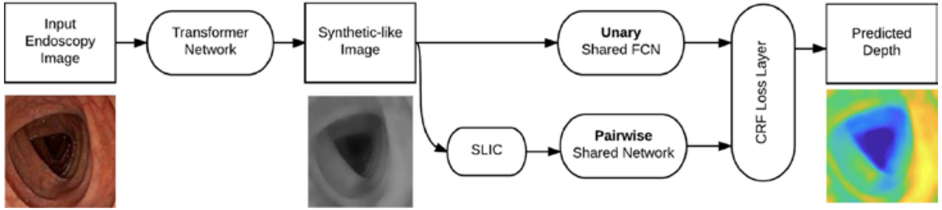
https://www.healthdirect.gov.au/surgery/upper-gi-endoscopy-and-colonoscopy



G. Scadding et al., Diagnostic tools in Rhinology EAACI position paper, 2011.

How to supervise monocular depth estimation?

Explicit style transfer



- Supervised training on simulated data from CT
- Real-to-synthetic conditional style transfer

→ Depth prediction on style-transferred images

Mahmood, F., & 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.

How to supervise monocular depth estimation?

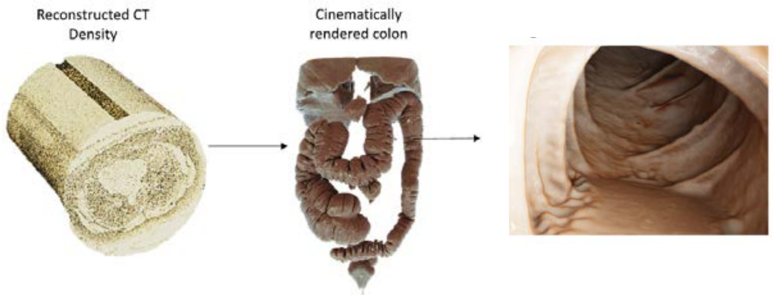
Explicit style transfer

- Supervised training on simulated data fro
- Real-to-synthetic conditional style transfe

→ Depth prediction on style-transferred imag

Mahmood, F., & Durr, N. J. (2018). Deep learning and conditional random fields estimation and topographical reconstruction from conventional endoscopy. *Med analysis*, *48*, 230-243.

Realistic simulation



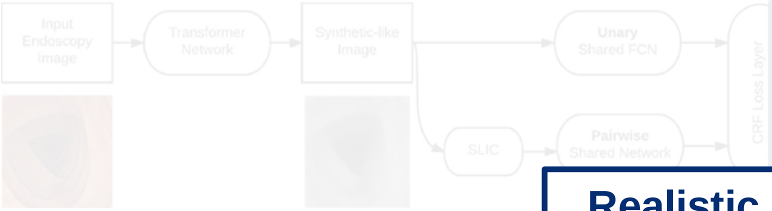
- Supervised training on simulated data from CT
- Cinematic (photorealistic) volume rendering

→ Depth prediction on acquired images

Mahmood, F., Chen, R., Sudarsky, S., Yu, D., & 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.

How to supervise monocular depth estimation?

Explicit style transfer



- Supervised training on simu
- Real-to-synthetic conditiona
- → Depth prediction on style-tra

Mahmood, F., & Durr, N. J. (2018). Deep learning and
estimation and topographical reconstruction from conve
analysis, *48*, 230-243.

Realistic simulation

Reconstructed CT Density

Cinematically rendered colon

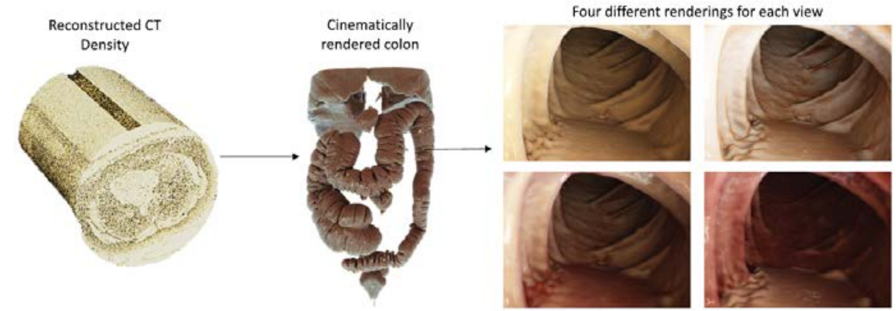
om CT

ing

ning with

edical images.

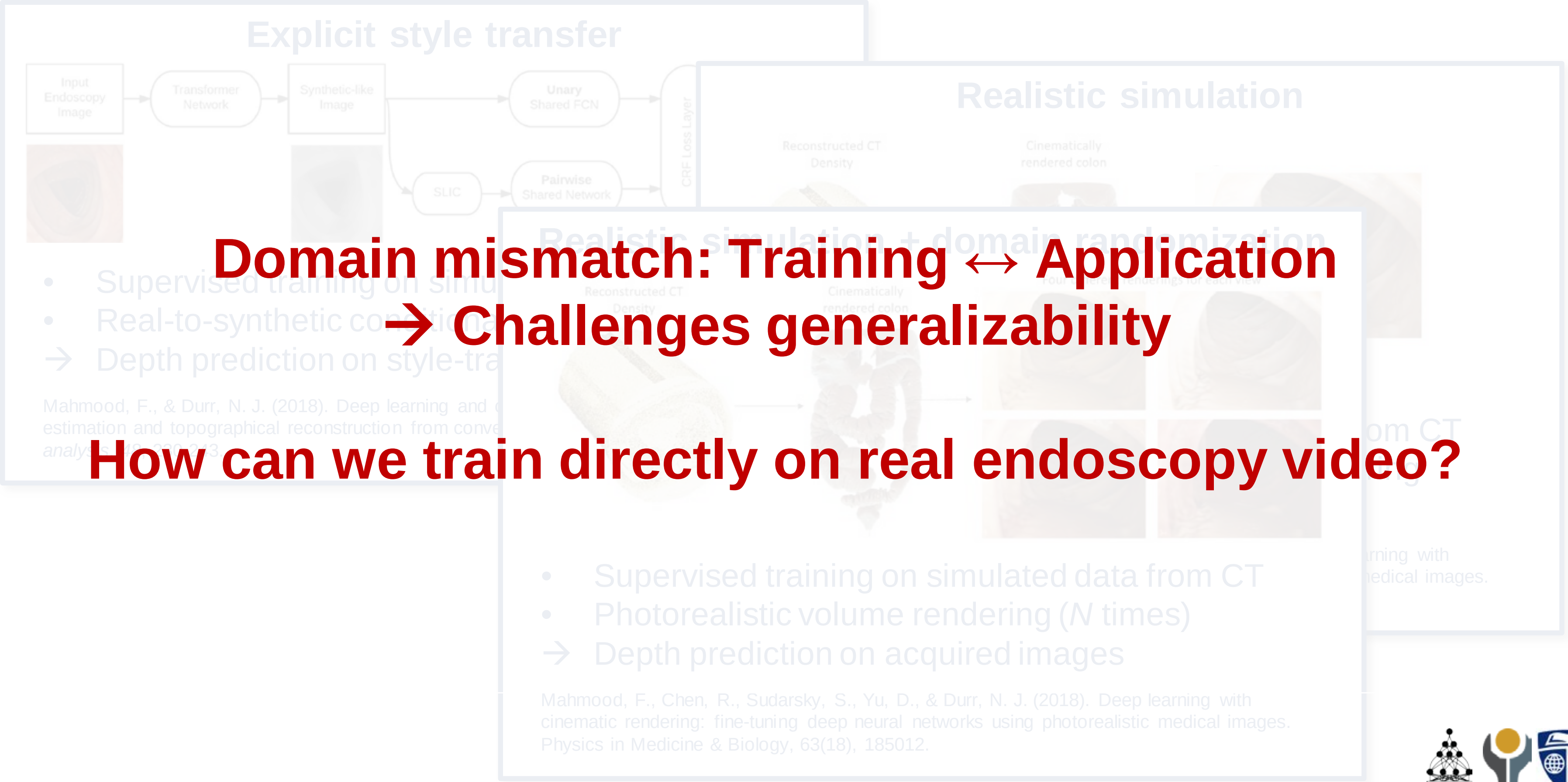
Realistic simulation + domain randomization



- Supervised training on simulated data from CT
- Photorealistic volume rendering (*N* times)
- → Depth prediction on acquired images

Mahmood, F., Chen, R., Sudarsky, S., Yu, D., & 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.

How to supervise monocular depth estimation?



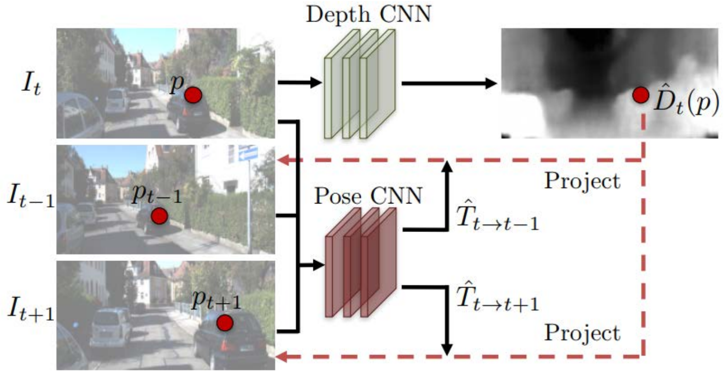
Domain mismatch: Training ↔ Application

→ Challenges generalizability

How can we train directly on real endoscopy video?

How to supervise monocular depth estimation?

Self-supervision



- Predict depth on target, synthesize neighbor views
- Photometric reconstruction loss for training

→ Self-supervision, directly on acquired video

Zhou, T., Brown, M., Snavely, N., & Lowe, D. G. (2017). Unsupervised learning of depth and ego-motion from video. In Proceedings of the IEEE CVPR (pp. 1851-1858).

Does this work for endoscopy?

How to supervise monocular depth estimation?

Classical – Structure from Motion

- Feature matching
- Triangulation and bundle adjustment

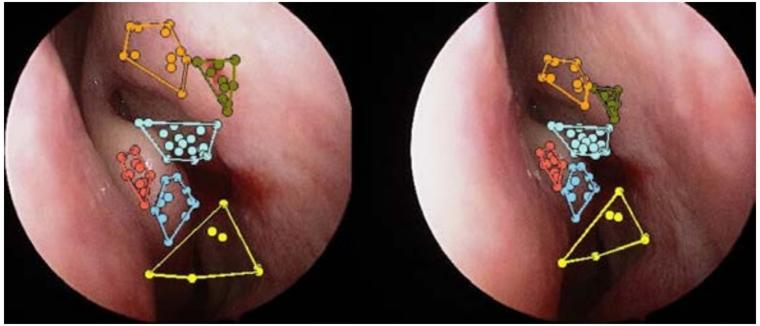
→ Reconstruction from acquired images

Snavely, N., Seitz, S. M., & Szeliski, R. (2006, July). Photo tourism: exploring photo collections in 3D. In ACM transactions on graphics (TOG) (Vol. 25, No. 3, pp. 835-846). ACM.

Does this work for endoscopy?

How to supervise monocular depth estimation?

Classical – Structure from Motion



- SURF feature matching, hierarchical refinement
- Triangulation and bundle adjustment

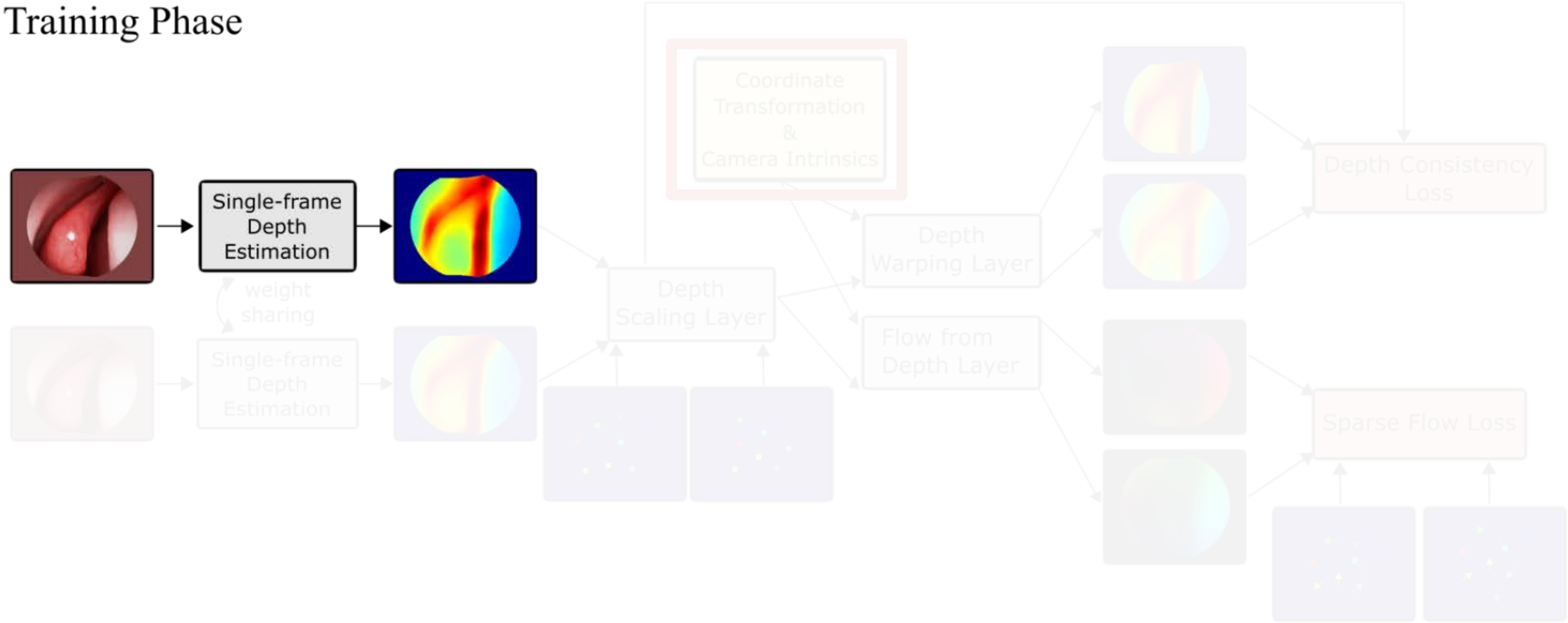
→ Reconstruction from acquired images (sparse)

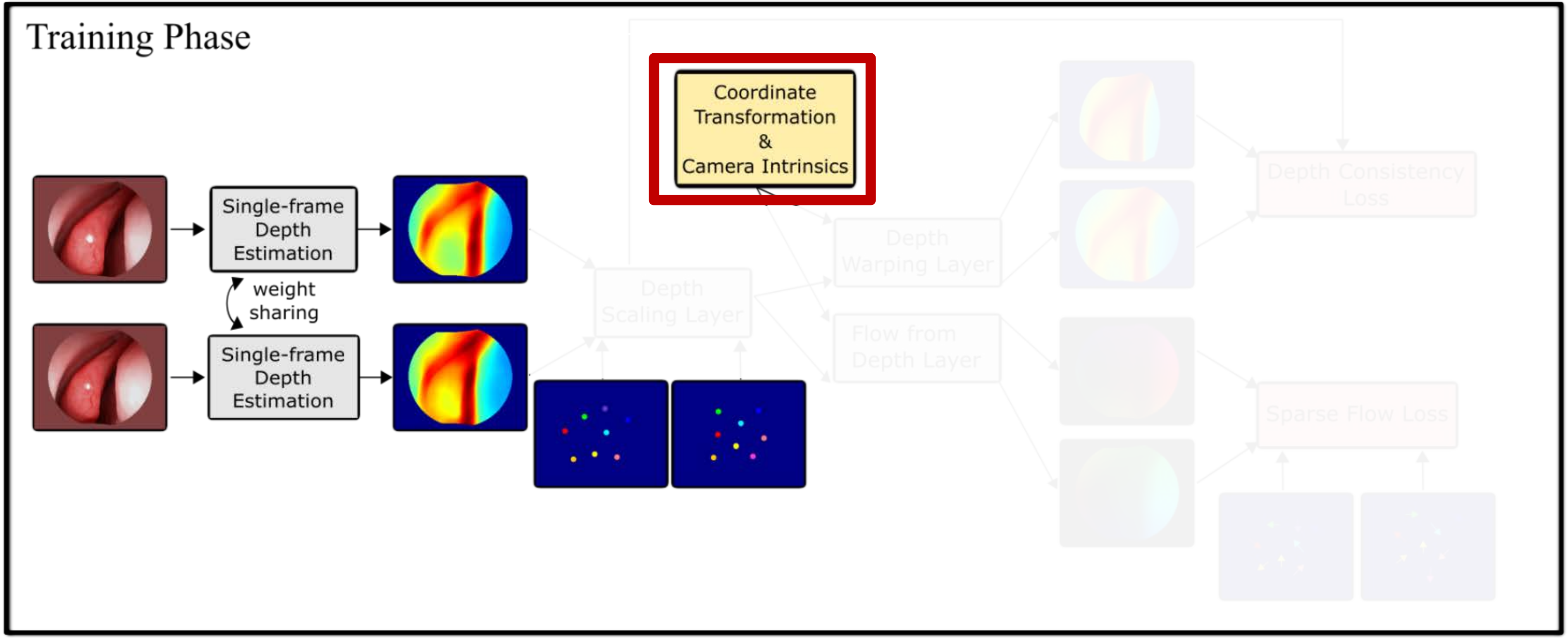
Leonard, S., Reiter, A., Sinha, A., Ishii, M., Taylor, R. H., & Hager, G. D. (2016, March). Image-based navigation for functional endoscopic sinus surgery using structure from motion. In Medical Imaging 2016: Image Processing (Vol. 9784, p. 97840V).

Yes(-ish).

So let's use this, then!

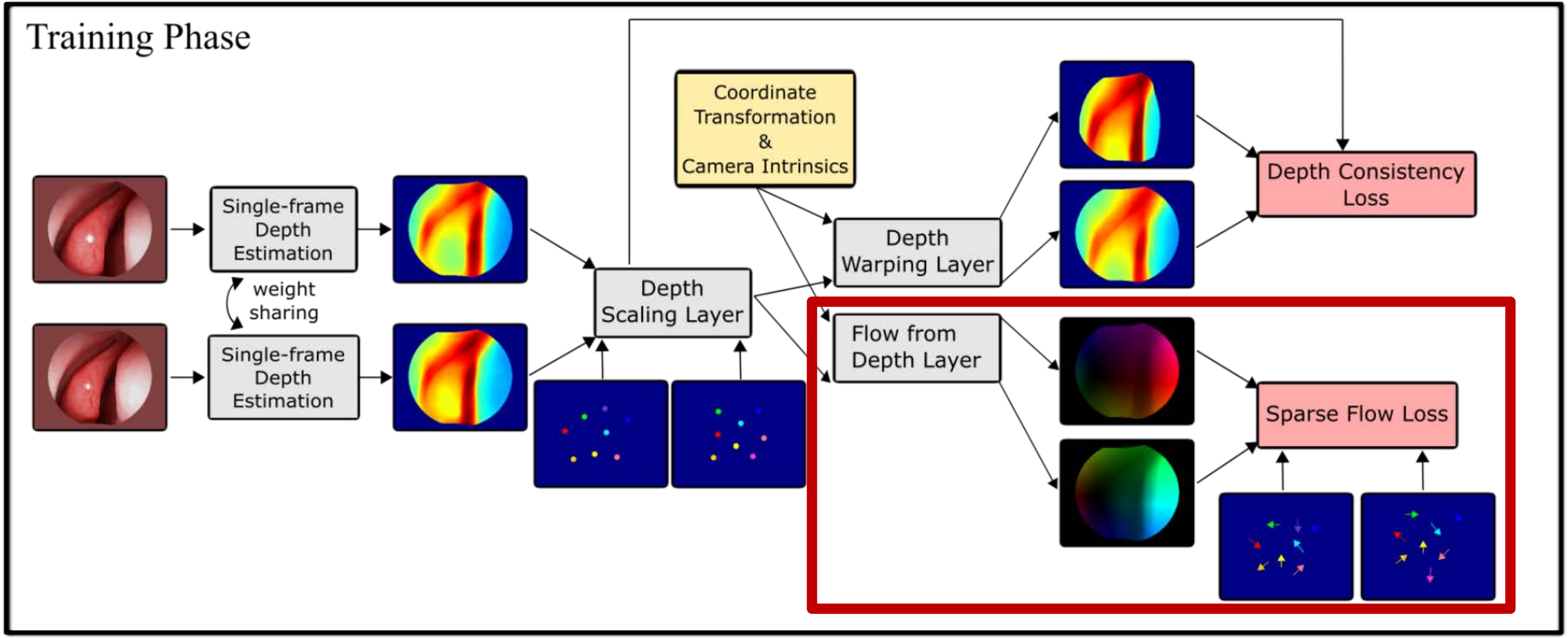
Training Phase





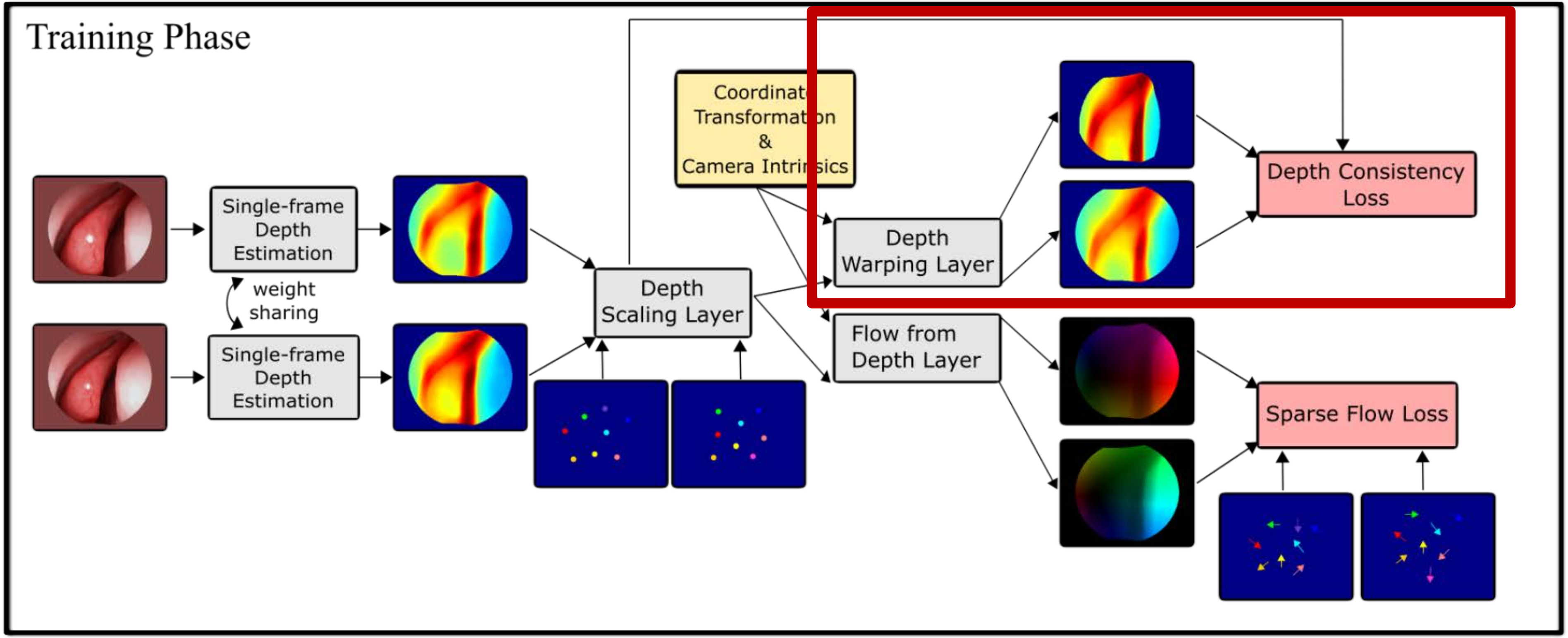
Structure from motion (SfM)-based self-supervision

- Run SfM on short video sequence (15 to 30 frames)
- Siamese network → Process multiple frames



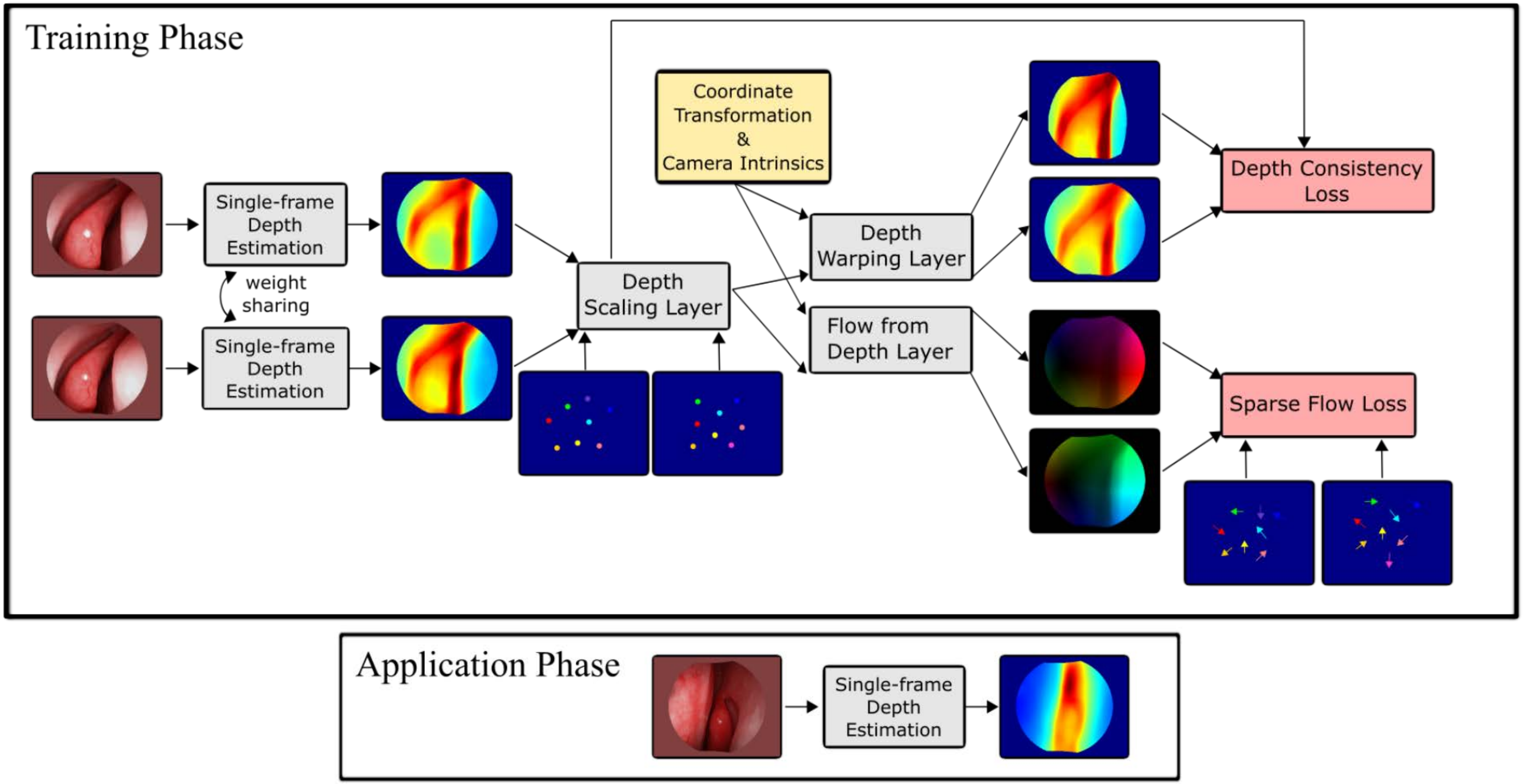
Sparse Flow Loss

- True 2D optical flow from 3D reconstructed points (SfM)
- Estimated optical flow from depth prediction



Depth Consistency Loss

- Differentiable warping operation to warp estimated depth into neighbor frame
- Enforces consistency among predictions


Training Phase
Single-frame Depth Estimation
weight sharing
Single-frame Depth Estimation
Depth Scaling Layer
Coordinate Transformation & Camera Intrinsics
Depth Warping Layer
Flow from Depth Layer
Depth Consistency Loss
Sparse Flow Loss
Application Phase
Single-frame Depth Estimation

Dataset and Architecture

- Endoscopic video (no tools) of 6 consenting patients
  - 8 minutes of video total; rectified, and downsampled to 256 x 320 pixels
  - Different endoscopes for every patient
  - 4 patients with corresponding CT data (ground truth, disregarding erectile tissue)

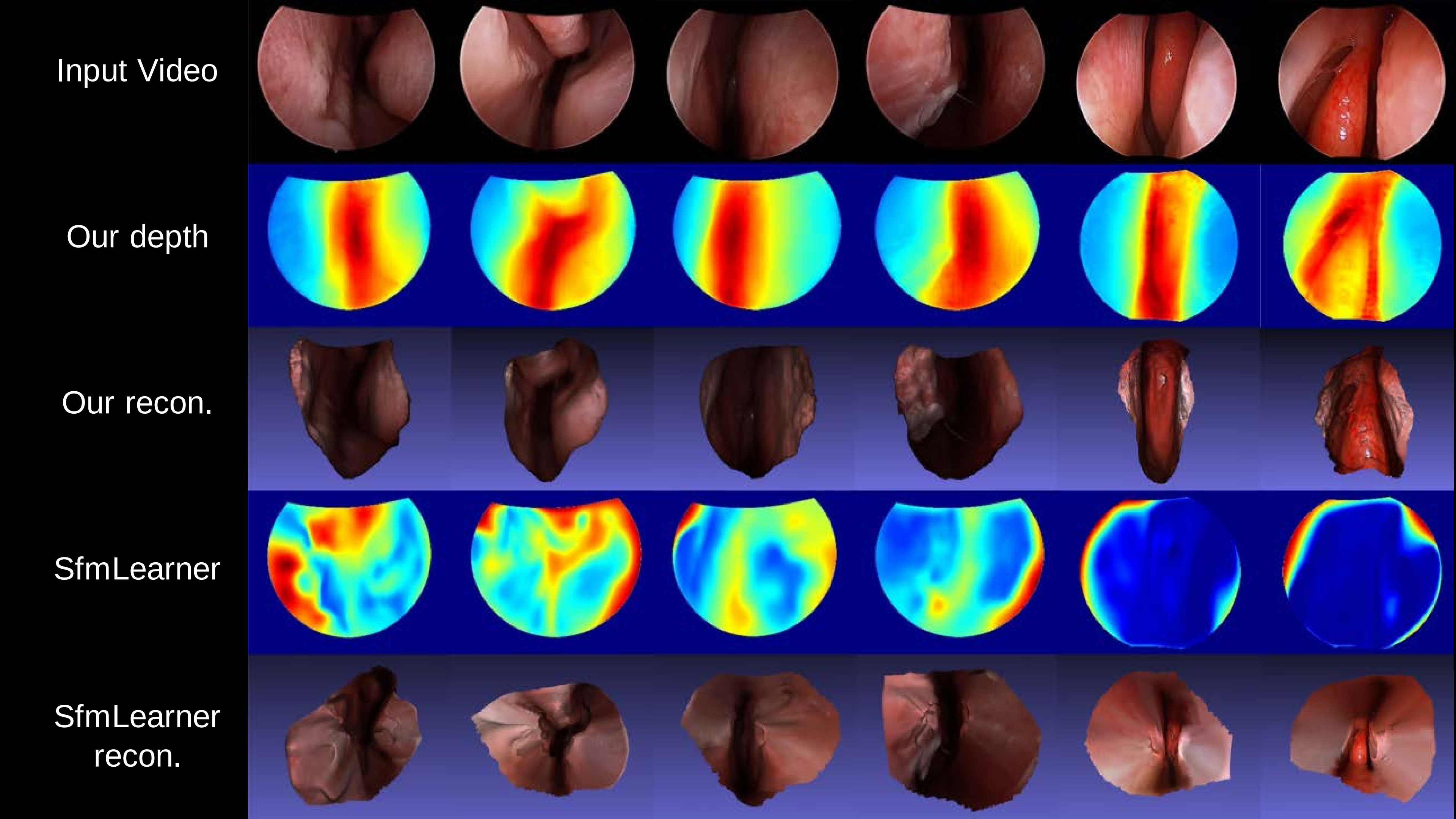
Liu, X., Sinha, A., Ishii, M., Hager, G. D., Reiter, A., Taylor, R. H., & Unberath, M. (2019). Self-supervised Learning for Dense Depth Estimation in Monocular Endoscopy. arXiv:1902.07766 and under review at IEEE TMI.

Dataset and Architecture

- Endoscopic video (no tools) of 6 consenting patients
  - 8 minutes of video total; rectified, and downsampled to 256 x 320 pixels
  - Different endoscopes for every patient
  - 4 patients with corresponding CT data (ground truth, disregarding erectile tissue)

- Depth estimation architecture
  - U-Net (8 M params): East to train on sparse signals but overfits heavily
  - FC-DenseNet-57 (1.5 M params): Generalizes well but hard to train from scratch
  - Teacher-Student approach
    - Teacher self-supervised learning
    - Teacher supervises student
    - Student self-supervised learning
  - Code available on GitHub: lppllppl920/EndoscopyDepthEstimation-Pytorch

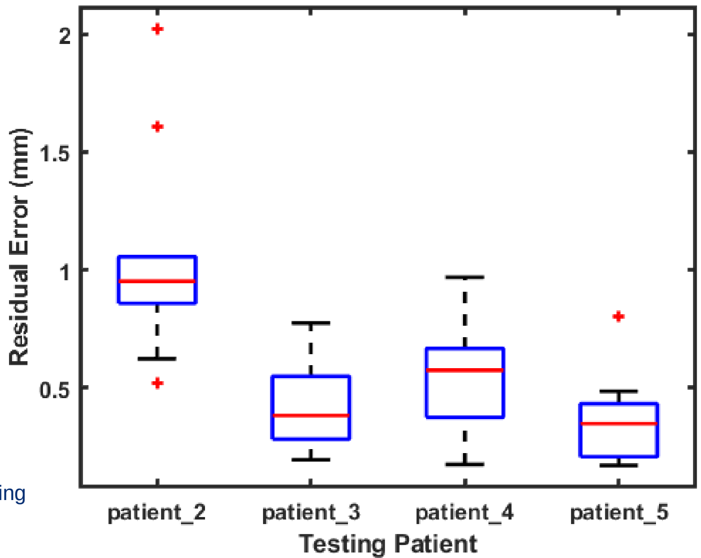
Liu, X., Sinha, A., Ishii, M., Hager, G. D., Reiter, A., Taylor, R. H., & Unberath, M. (2019). Self-supervised Learning for Dense Depth Estimation in Monocular Endoscopy. arXiv:1902.07766 and under review at IEEE TMI.


Input Video
Our depth
Our recon.
SfmLearner
SfmLearner recon.

Quantitative Results

- Leave-one-out training
- Randomly sample 20 frames per left-out patient
  - Estimate depth
  - Register to patient CT surface via GD-IMLOP (no shape deformation)
  - Compute residual error

- Sub-millimeter accuracy in most cases!
  - SfmLearner: > 10 mm
  - Deep (dark) regions exhibit high variation → Outliers
  - CT is imperfect ground truth (erectile tissue)

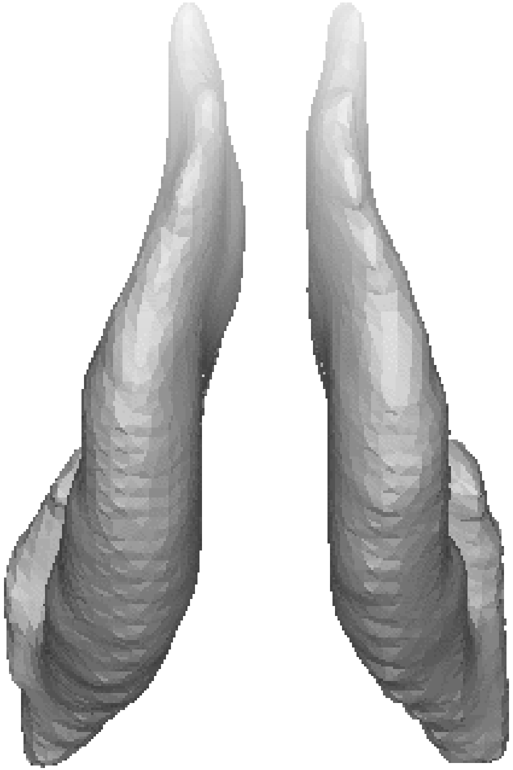


Liu, X., Sinha, A., Ishii, M., Hager, G. D., Reiter, A., Taylor, R. H., & Unberath, M. (2019). Self-supervised Learning for Dense Depth Estimation in Monocular Endoscopy. arXiv:1902.07766 and under review at IEEE TMI.

Towards Next-generation Image Guidance

Navigating Without Prior Information

Estimating Patient-specific Anatomy



Potential sources of patient-specific models

- CT scans
- Statistical shape model
- …

Can we build a patient-specific, dense 3D model

- intra-operatively and
- on-the-fly?

Estimating Patient-specific Anatomy

l sources of patient-specific models

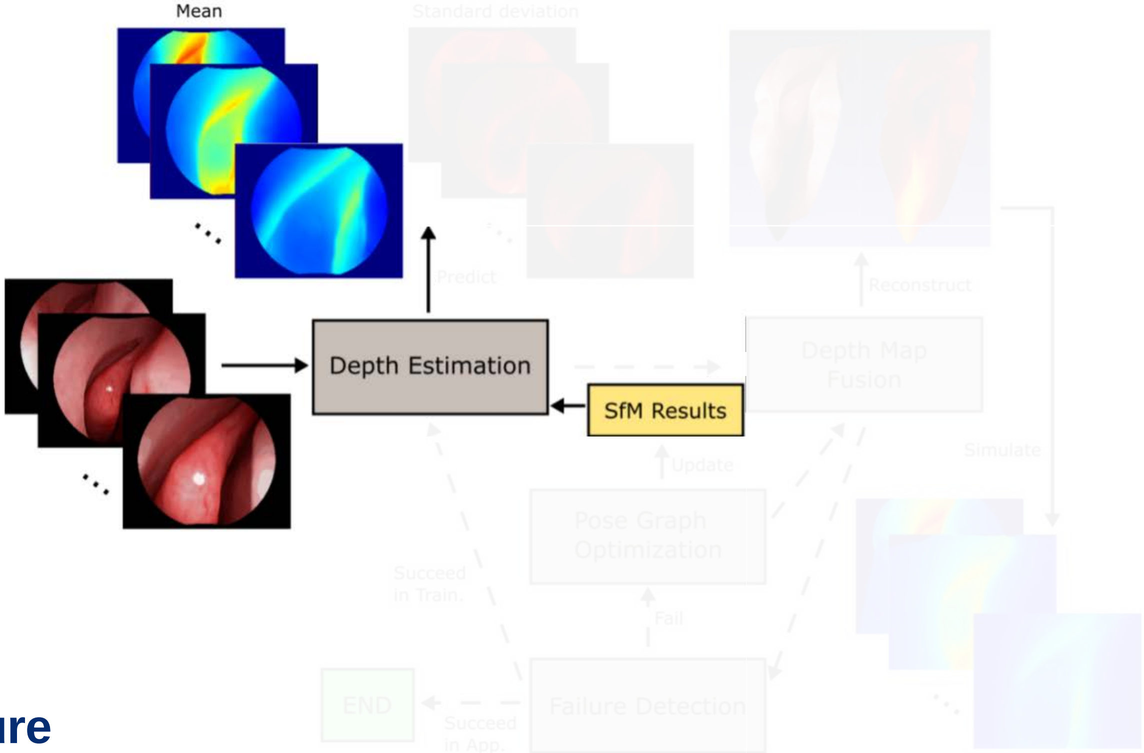
- CT scans
- Statistical shape model
- …

build a patient-specific, dense 3D model

- ntra-operatively and
- n-the-fly?

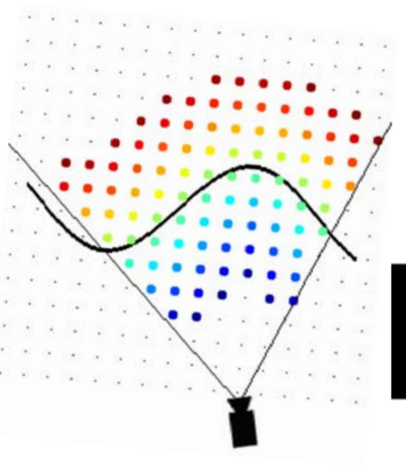
d we benefit two ways

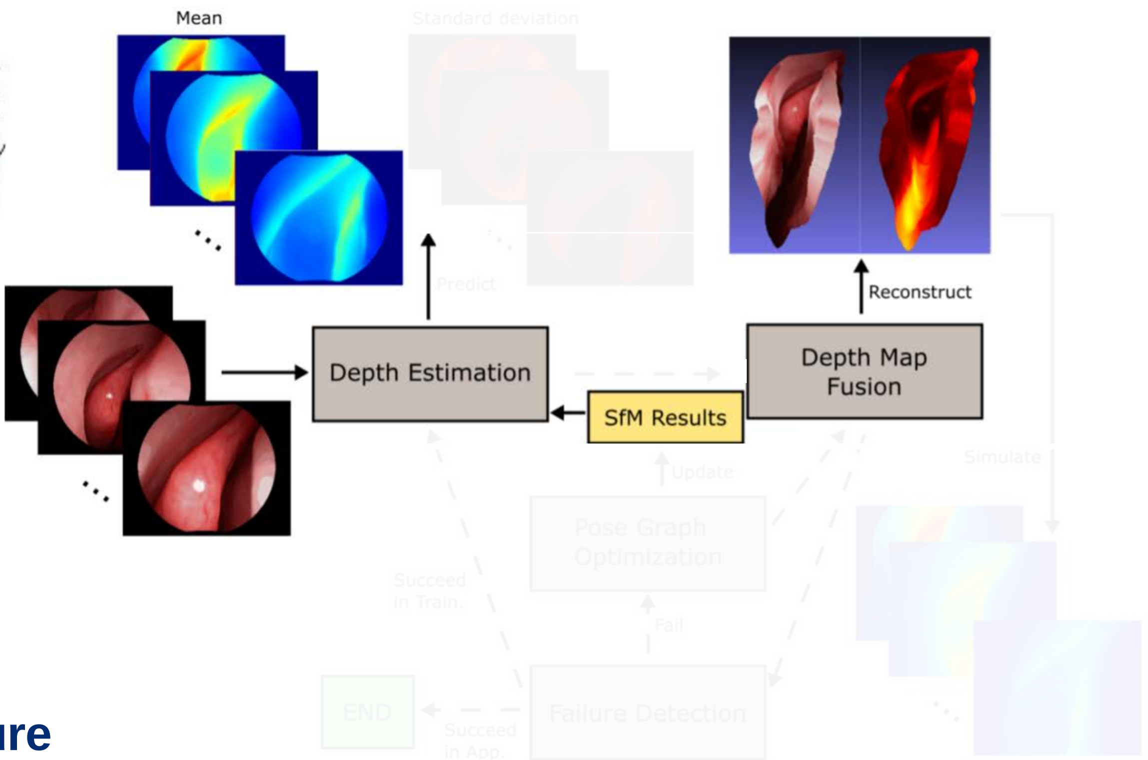
- Bootstrapping for **dense depth supervision**
- **Uncertainty of depth estimates**



The big picture

1. Self-supervised training of depth estimation (now on long video sequences)

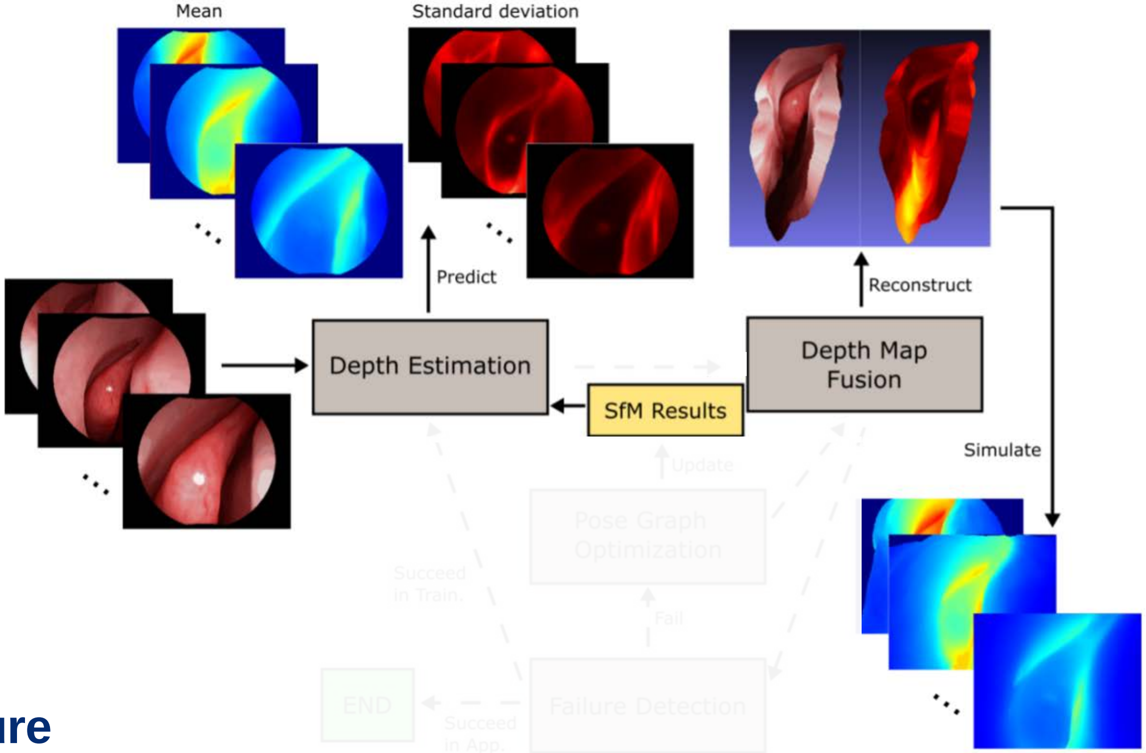




The big picture

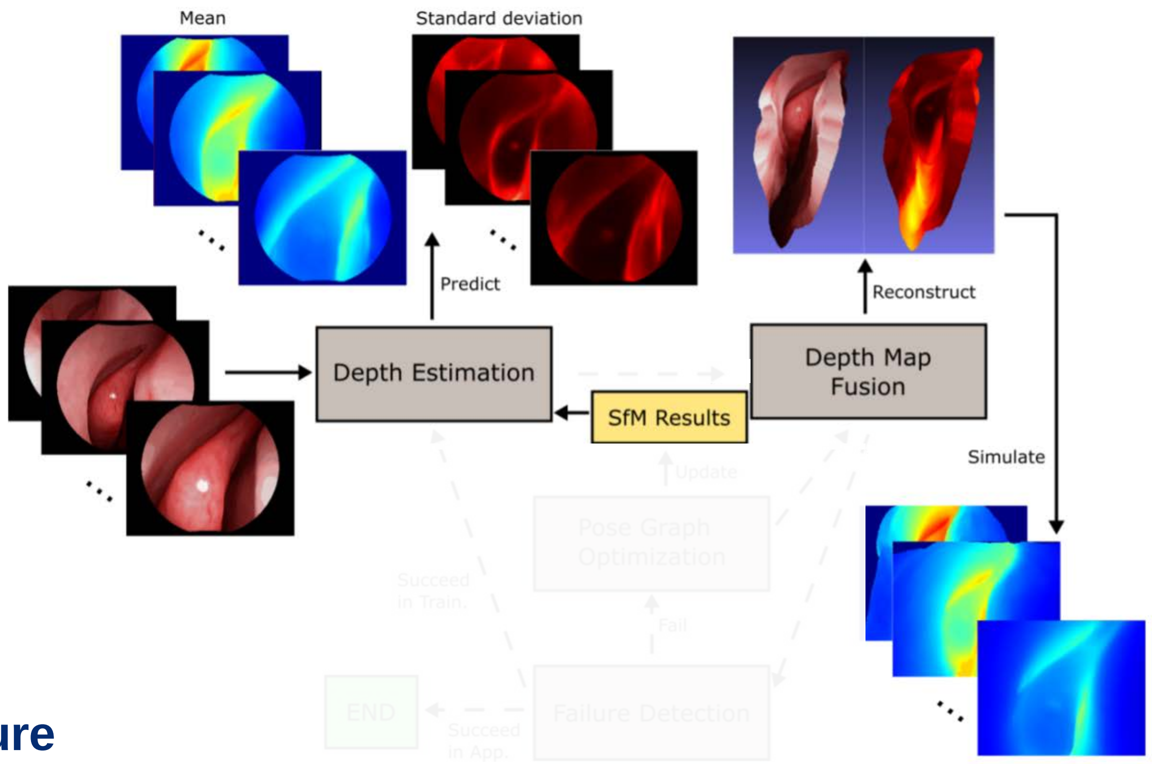
1. Self-supervised training of depth estimation (now on long video sequences)
2. Volumetric fusion (truncated signed distance function) → Mean, STD

Fusion modified from: Curless, B., & Levoy, M. (1996). A volumetric method for building complex models from range images.



The big picture

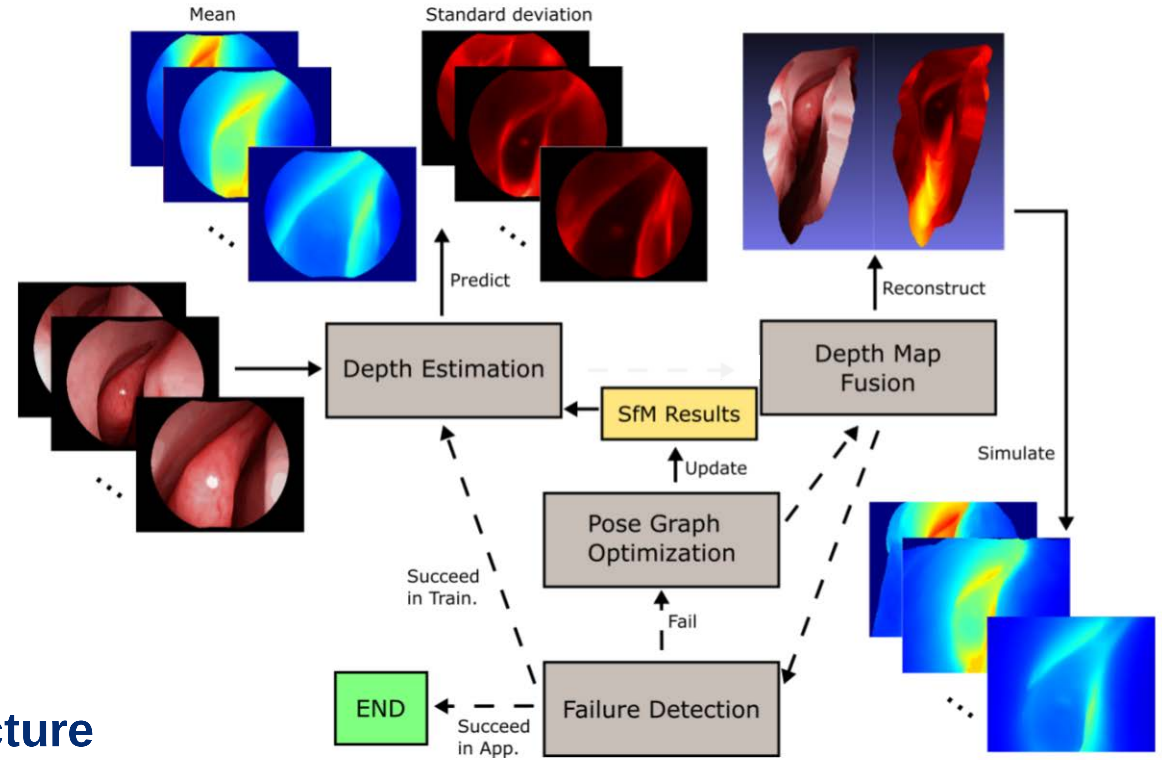
1. Self-supervised training of depth estimation (now on long video sequences)
2. Volumetric fusion (truncated signed distance function) → Mean, STD
3. Bootstrapping → Dense supervision of mean depth and uncertainty



The big picture

1. Self-supervised training of depth estimation (now on long video sequences)
2. Volumetric fusion (trunca[illegible]tion) → Mean, STD
3. Bootstrapping → Dense supervision of mean depth and uncertainty

But wait, there's more!

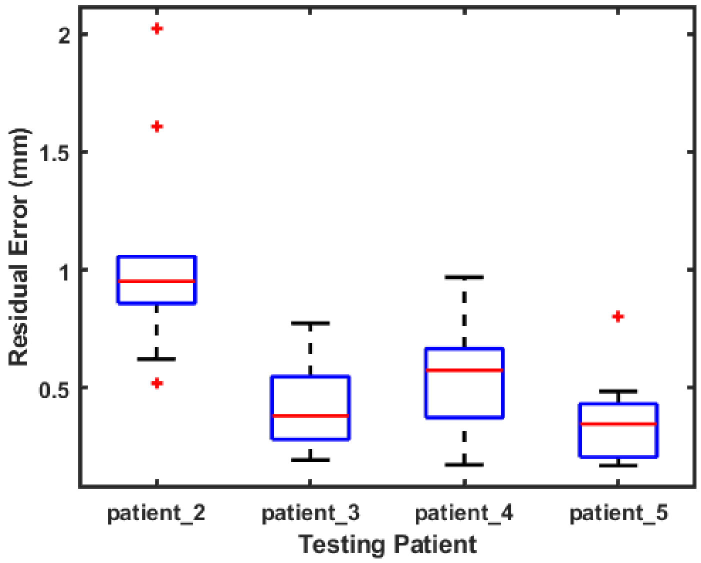


More big picture

- SfM results can be incorrect (few points etc.) → **Fusion will be wrong**
- Consistency between simulated and estimated depth → **Failure detection**
- If close → Pose graph refinement; If far off → Re-run SfM

Results and Observations

- Again, leave-one-out and GD-IMPLOP to patient CT
- Sub-millimeter errors
- Error seems higher → **Misleading**
  - Reconstruction is of ~ 1 minute video not just a single frame
  - Registration has larger residual, but average is over much larger region



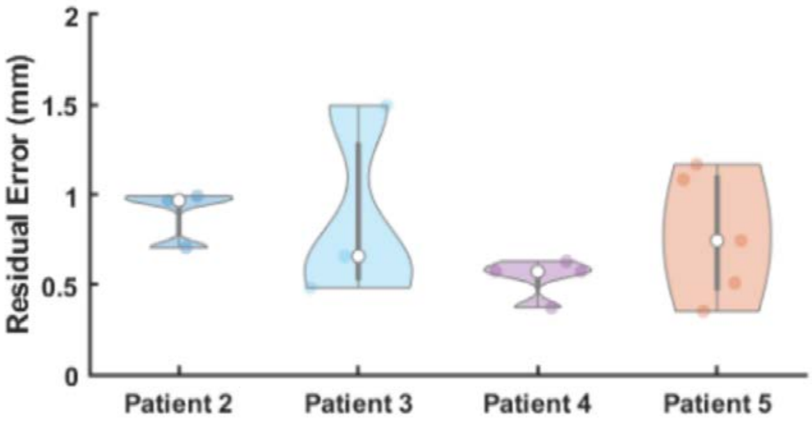


Image Guidance for Endoscopic Procedures

Concluding Remarks – Accounting for Anatomical Change

Where do we go from here?

Quantitative endoscopy

- Longitudinal monitoring of anatomical change
- E.g. for monitoring polyp behavior after steroid injection

The fairly untapped supreme discipline…

Monitoring anatomical change during surgery

- How to deal with tools?
- Blood, gore, and all other sorts of unseen variation?

Thank you.
Questions?