Deep Learning Based Feature Descriptor and Matching for Endoscopy

Graduate Level

Xingjian Hao & Tengfei Jiang

Fall 2022 ENGN 2605

School *of* Engineering

Endoscopy:

An endoscopy procedure involves inserting a thin, flexible tube called an endoscope down your throat and into your esophagus. A tiny camera on the end of the endoscope lets doctor examine your esophagus, stomach and the beginning of your small intestine.

3D models[1] or panoramic images[2] reconstructed from endoscopy videos are widely used in clinical applications, helping investigating symptoms, diagnosing, and treating patients with related diseases.

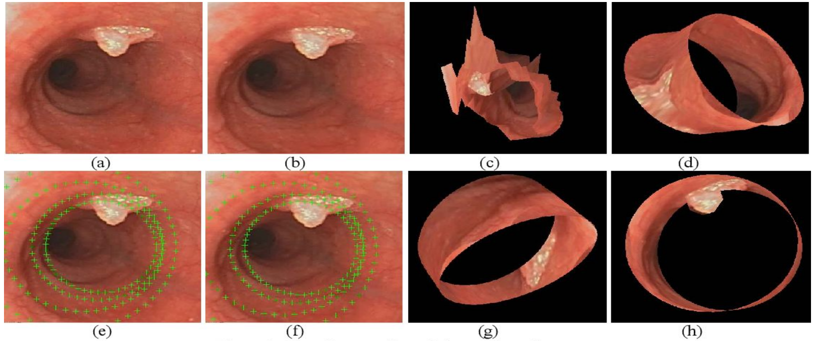
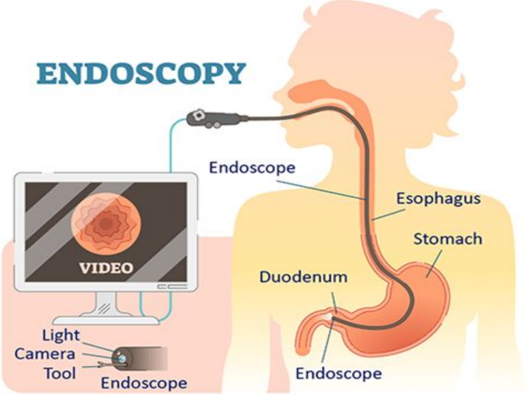


Figure 8. Experiment with real data with outlier parts

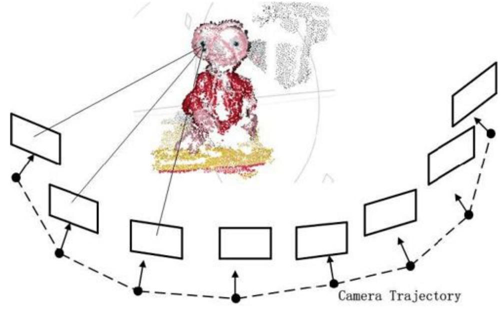


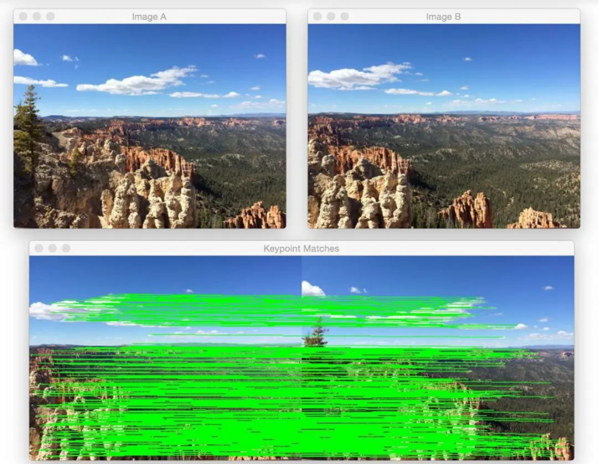
[1] Liu, Xingtong, Yiping Zheng, Benjamin Killeen, Masaru Ishii, Gregory D. Hager, Russell H. Taylor, and Mathias Unberath. "Extremely dense point correspondences using a learned feature descriptor." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4847-4856. 2020.
[2] Farhat, Manel, Houda Chaabouni-Chouayakh, and Achraf Ben-Hamadou. "Self-supervised endoscopic image key-points matching." *Expert Systems with Applications* 213 (2023): 118696.

Problem Statement:

In order to reconstruct based on videos taken by monocular camera, we need find the matching correspondences in images in sequence, thus we can reconstruct 3d model with dense features, or get panoramic images with homography matrix as we did in previous lab.

Therefore, the matching process is of great significance, while due to the **unsatisfactory quality of the video** and the **scarcity of texture** in the organ, the matching problem cannot be easily solved by traditional methods.





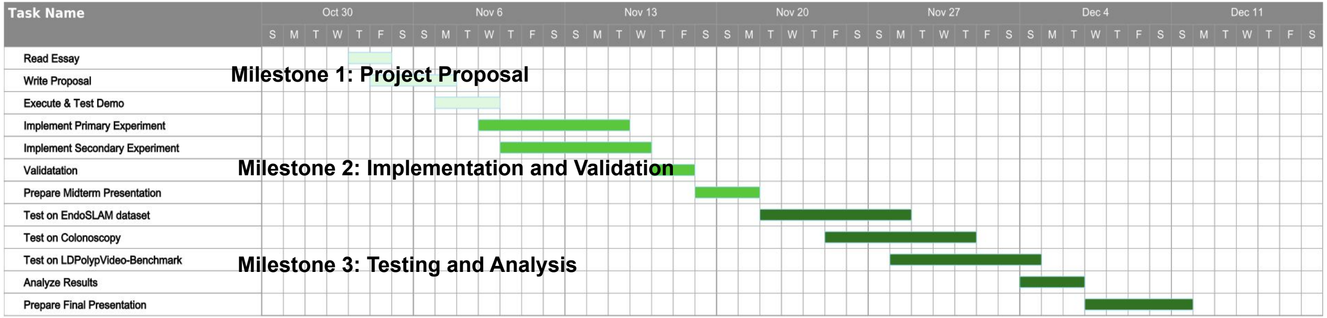
Project Goal:

To Solve:

- Self-labeled Descriptor for CNN training[1] (Primary Paper)
- Learnable Dense Feature Descriptor[2](Secondary Paper)

Goal:

- Recreate two experiments
- Compare and analysis the performance of two methods with multiple datasets[3][4][5]



Milestone 1: Project Proposal

Milestone 2: Implementation and Validation

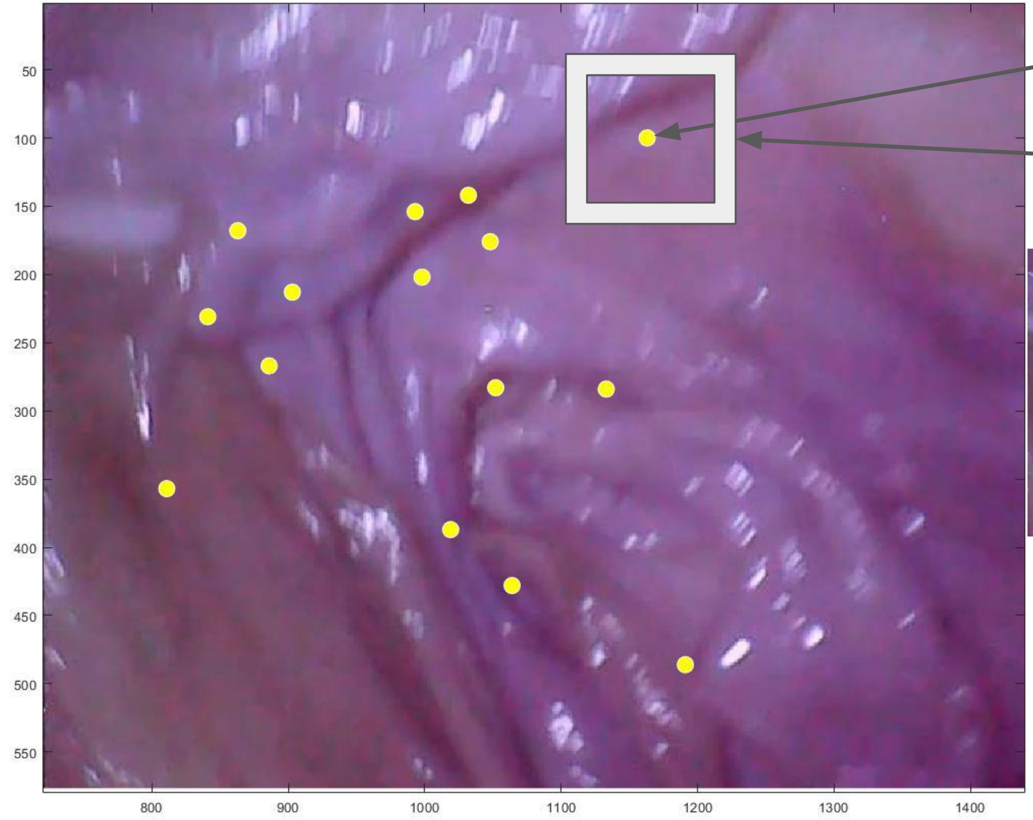
Milestone 3: Testing and Analysis

[1] Farhat, Manel, Houda Chaabouni-Chouayakh, and Achraf Ben-Hamadou. "Self-supervised endoscopic image key-points matching." *Expert Systems with Applications* 213 (2023): 118696.
[2] Liu, Xingtong, Yiping Zheng, Benjamin Killeen, Masaru Ishii, Gregory D. Hager, Russell H. Taylor, and Mathias Unberath. "Extremely dense point correspondences using a learned feature descriptor." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4847-4856. 2020.
[3] Kutsev Bengisu Ozyoruk, Guliz Irem Gokceler, Taylor L. Bobrow, Gulfize Coskun, Kagan Incetan, Yasin Almalioglu, Faisal Mahmood, Eva Curto, Luis Perdigoto, Marina Oliveira, Hasan Sahin, Helder Araujo, Henrique Alexandrino, Nicholas J. Durr, Hunter B. Gilbert, Mehmet Turan. "EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos." Medical Image Analysis, Volume 71, 2021, 102058,ISSN 1361-8415,
[4] Anita Rau, P. J. Eddie Edwards, Omer F. Ahmad, Paul Riordan, Mirek Janatka, Laurence B. Lovat, Danail Stoyanov. "Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy." International Journal of Computer Assisted Radiology and Surgery (2019) 14:1167–1176
[5] Yiting Ma, Xuejin Chen , Kai Cheng, Yang Li, and Bin Sun. "LDPolypVideo Benchmark: A Large-Scale Colonoscopy Video Dataset of Diverse Polyps." National Engineering Laboratory for Brain-inspired Intelligence Technology and Application

Primary Paper:

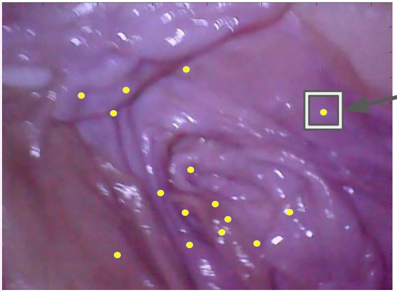
Self-Supervised Endoscopic Image Key-Points Matching

The Principle of training model:



Feature taken using SIFT

Crop around the key point, and make patch with 128*128 pixels P_j^t

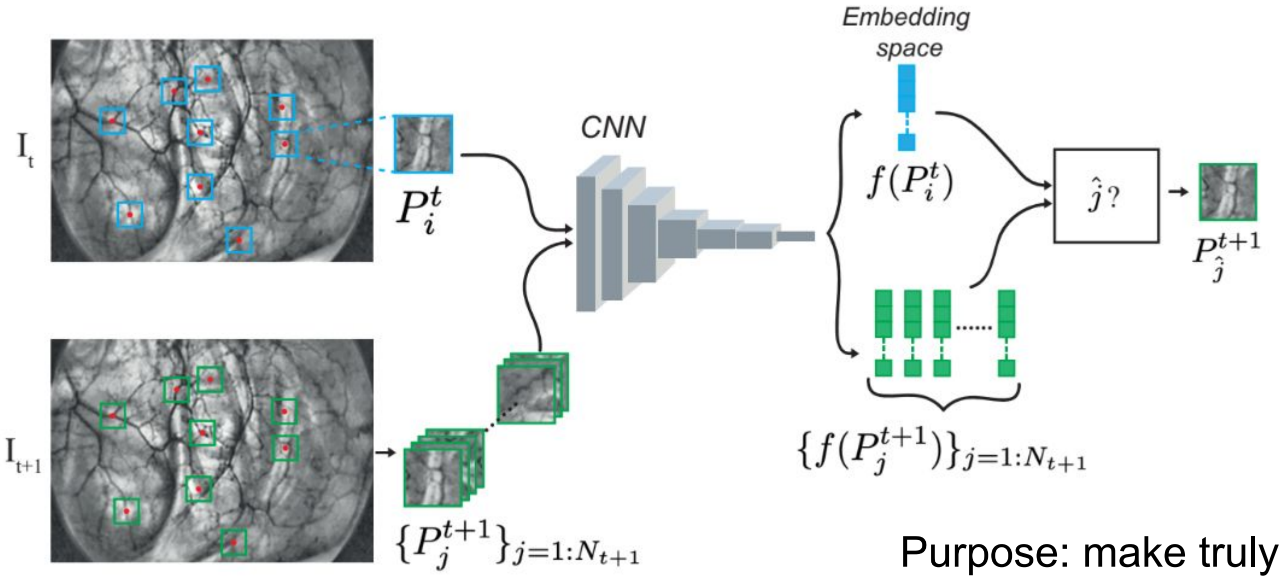


Patch up in the next frame

P_j^{t+1}

Extract features and make patches based on the features, and make them into two sets of patches. $\{P_j^t\}_{i=1:N_t}$ $\{P_j^{t+1}\}_{i=1:N_{t+1}}$

The Principle of training model:

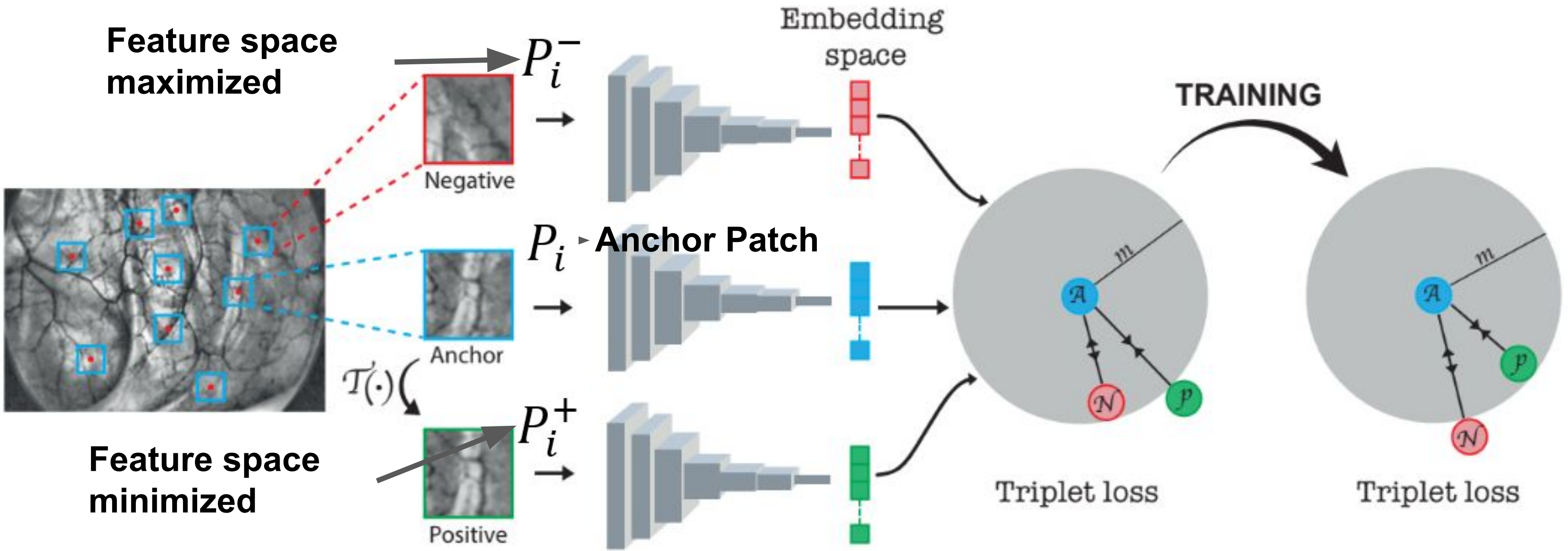


$$\hat{j} = \underset{j}{\operatorname{argmin}} \left\| f(P_j^{t+1}) - f(P_i^t) \right\|_2$$

Purpose: make truly matching patches in two sets **as similar as possible**, the similarity between patches is evaluated by Euclidean Distance.

The network is power by L2-Net

The Principle of training model:



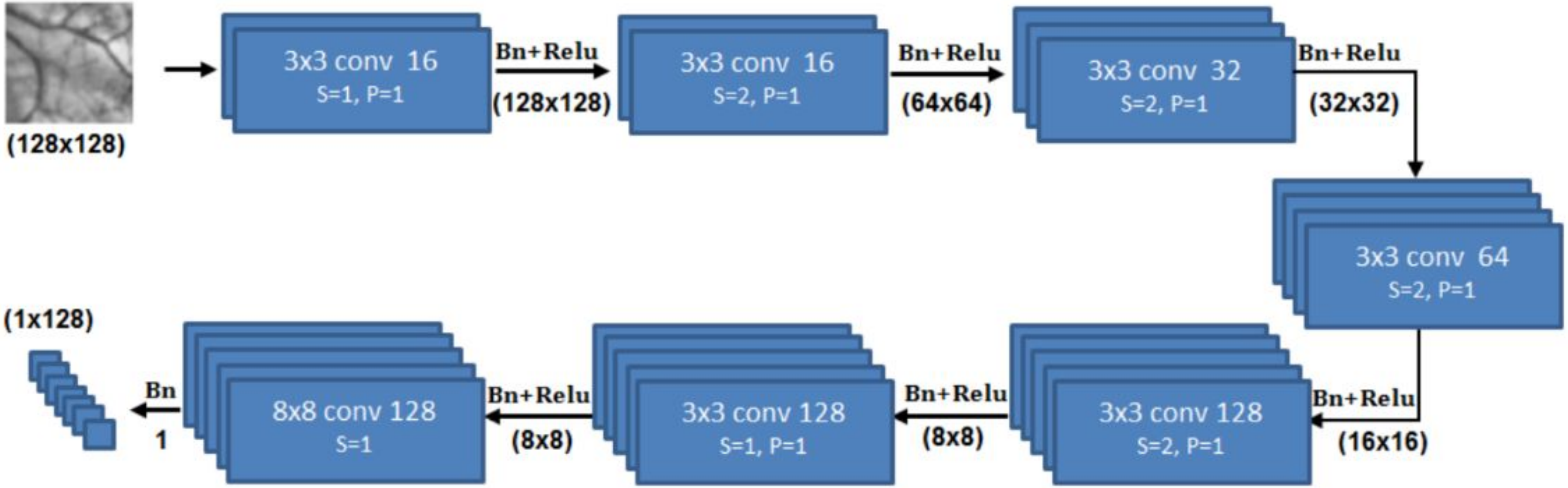
Feature space maximized

Anchor Patch

Feature space minimized

The Pi is seen as **Anchor Patch**, and our goal is to make distance between **Pi & P-** maximum, and **Pi & P+** minimum. The selection of P+ is based on **simulated homography transformation**, and the selection of P- is inspired by strategy proposed in **HardNet** method(could be chosen randomly).

The Function of training model:



Basically, the model convert **128*128 patch** into **1*128 descriptor**, and the Euclidean distance between descriptors can be taken as the similarity between features, then with simple sort function, we can list the corresponding matches with the distances.

Training and testing details:

We took **EndoSLAM dataset** as the training and testing source, which consists of real endoscopy video frames taken by different cameras, corresponding calibration data and the pose information. Thus the essential matrix between frames can be calculated .

Besides the **Pretrained model** given by the paper, I trained **Specific Organ Model** with corresponding organ dataset, and compared their results with matches generate by **SIFT**.

Eventually, I trained 5 models for 2 Cameras (LowCam, HighCam), and tested 46 trajectories in total. Each trajectory contains around 600-1300 frames.

The process of Evaluation:

Take the coordinates extracted with SIFT into the model and **match the correspondence** and save the feature coordinates after matching.

Load the **Pose** file of dataset which contains [tx,ty,tz,rx,ry,rz,rw], and calculate the **Essential Matrix** between two consecutive frame. Thus, we have epipolar lines plotted in the second frame.

$$P_{c1} = R_1 R_w + T_1; \;\; P_{c1} = R_1 R_w + T_1$$

$$P_{c2} = R_2 R_1^T P_{c1} + T_2 - R_2 R_1^T T_1;$$

$$R_{12} = R_2 R_1^T; \; T_{12} = T_2 - R_2 R_1^T T_1$$

$$E = [T]_x R$$

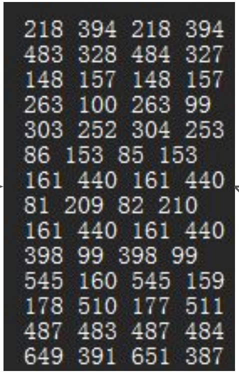
$$\begin{bmatrix} \bar{\xi} & \bar{\eta} & 1 \end{bmatrix} \begin{bmatrix} e_{11} & e_{12} & e_{13} \\ e_{21} & e_{22} & e_{23} \\ e_{31} & e_{32} & e_{33} \end{bmatrix} \begin{bmatrix} \xi \\ \eta \\ 1 \end{bmatrix} = 0.$$

$$A\bar{\xi} + B\bar{\eta} + C = 0$$

Flow Chart:

School of Engineering

Dataset



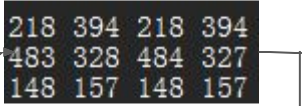
```
218 394 218 394
483 328 484 327
148 157 148 157
263 100 263 99
303 252 304 253
86 153 85 153
161 440 161 440
81 209 82 210
161 440 161 440
398 99 398 99
545 160 545 159
178 510 177 511
487 483 487 484
649 391 651 387
```

SIFT features Coordinates

Train Model in advance

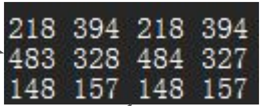
Matching Process

Specific Organ Model



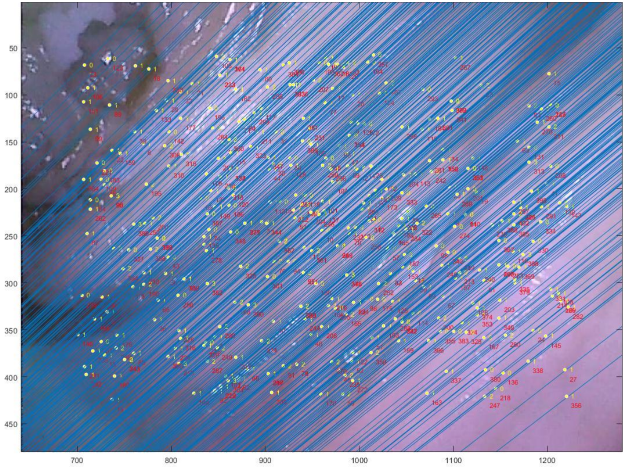
```
218 394 218 394
483 328 484 327
148 157 148 157
```

Pretrained Model



```
218 394 218 394
483 328 484 327
148 157 148 157
```

Calculate **Essential Matrix** with **Pose**.
Plot **Epipolar lines** and get distance between them and features in the next frame.
Judge whether it is valid match based on **threshold**
Loop over frames and threshold
Record the **Accuracy** & **Total distance**



Evaluation with Epipolar Lines:

The paper uses dataset with ground truth which indicates the matches and their coordinates, and it evaluates the performance with **absolute accuracy**.

Due to the **scarcity of labelled dataset** in endoscopy, I evaluated the accuracy with substitution method with **Epipolar Line**.

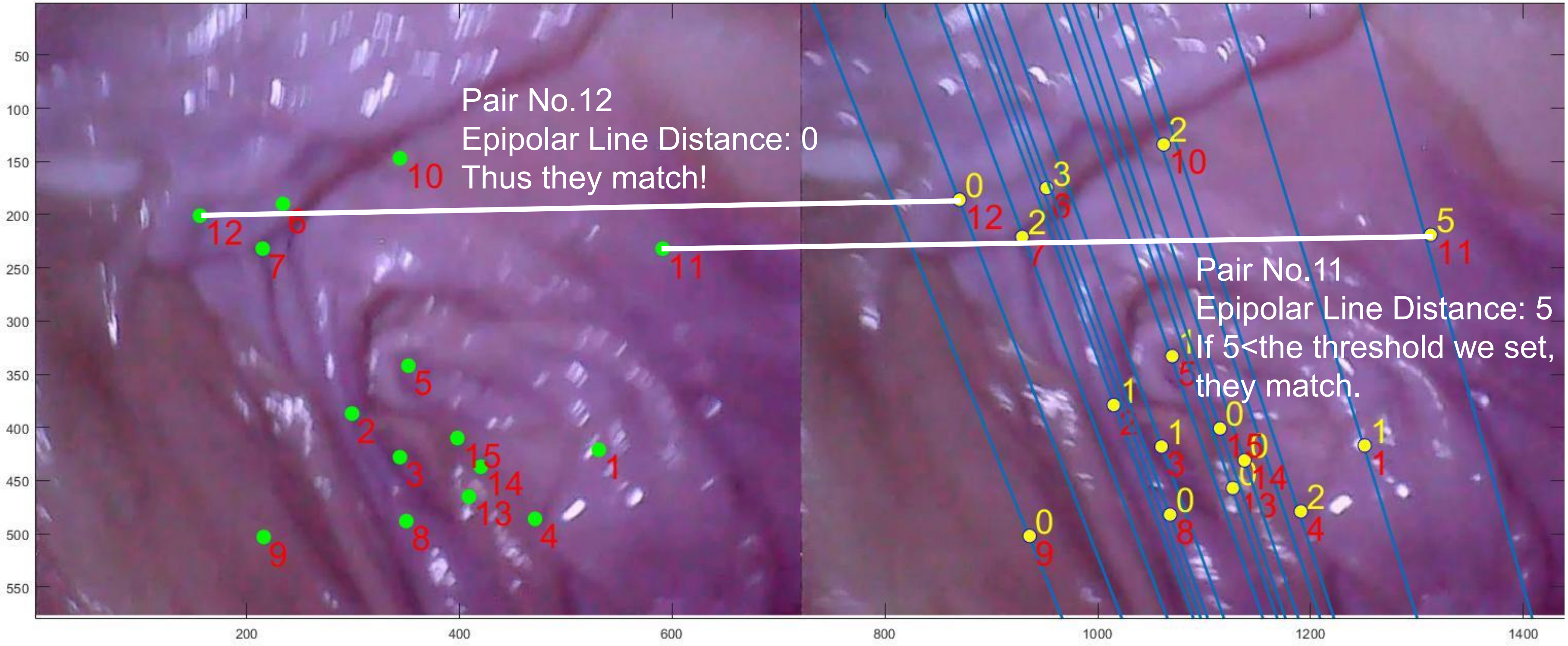
Since not only camera moves in consecutive frames, **the organ moves as well**, thus the feature may deviate from the place where it should be assumed based simply on pose changes. Also, it makes the accuracy nearly impossible to be 100%, since there must be **points moving out of frame**, and organ is moving. If the framerate of video is relatively low or camera is moving fast, the accuracy could be even **lower**.

When mistaken feature point l**ocate along the line**, the distance between point and epipolar line could be little while it is still away from the ground truth, and it may make accuracy **higher.**

Therefore, we set **multiple thresholds** of distance as the criteria of judging correspondences matching or not.

Epipolar Line:

Green points: Feature Points in Frame n
Yellow points: Feature Points in Frame n+1
Blue lines: Epipolar Lines based on Pose

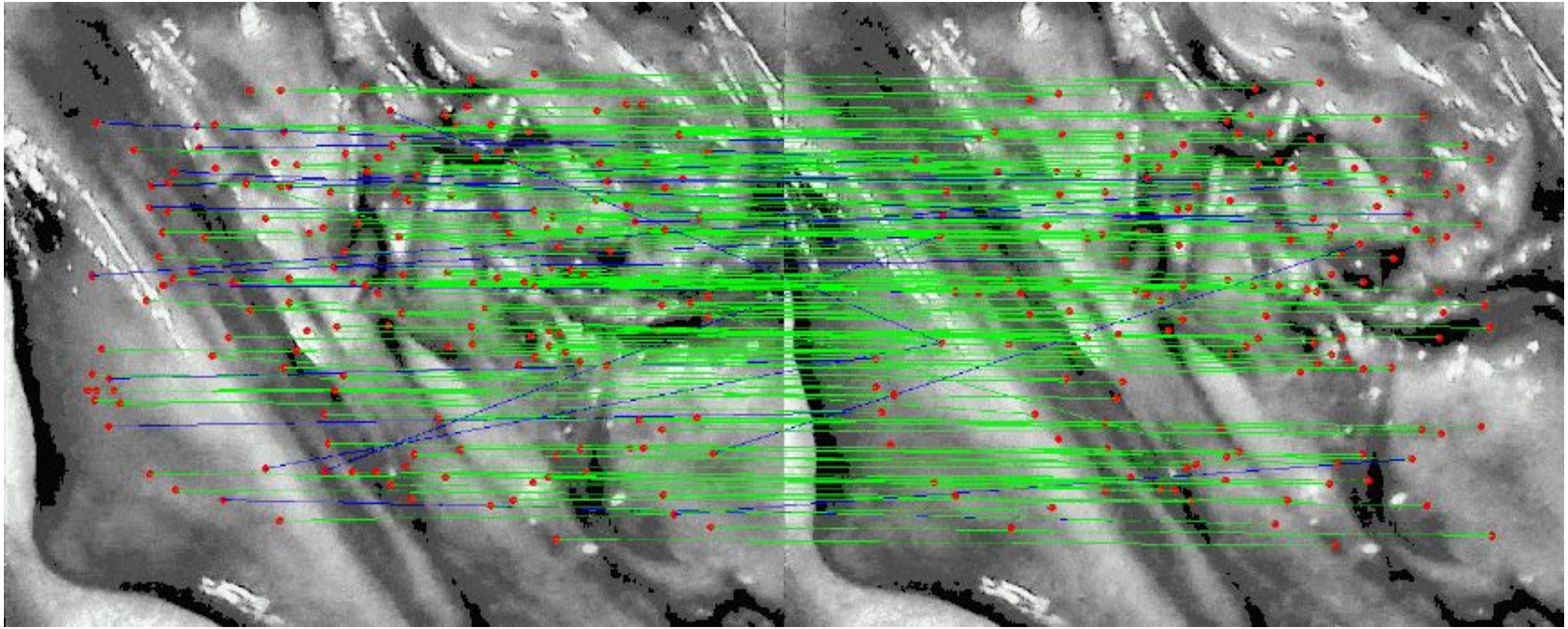


Pair No.12
Epipolar Line Distance: 0
Thus they match!

Pair No.11
Epipolar Line Distance: 5
If 5<the threshold we set,
they match.

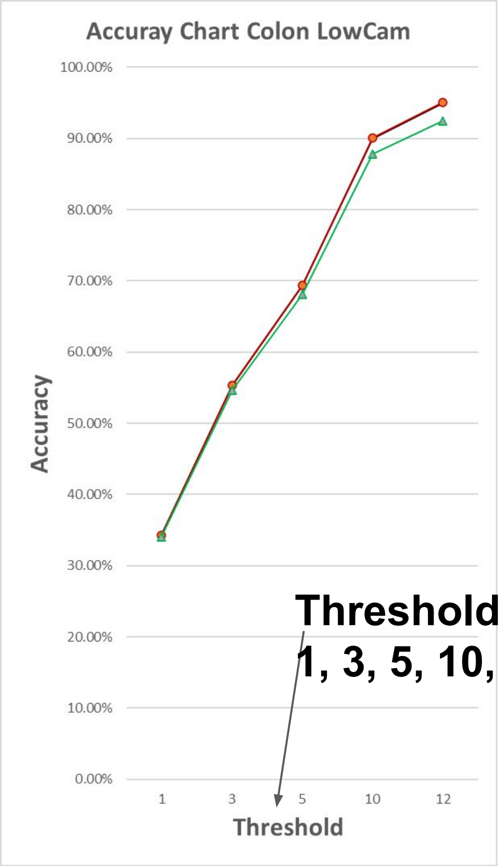
Matching Demo:

Matching Results with SIFT extraction and sorting

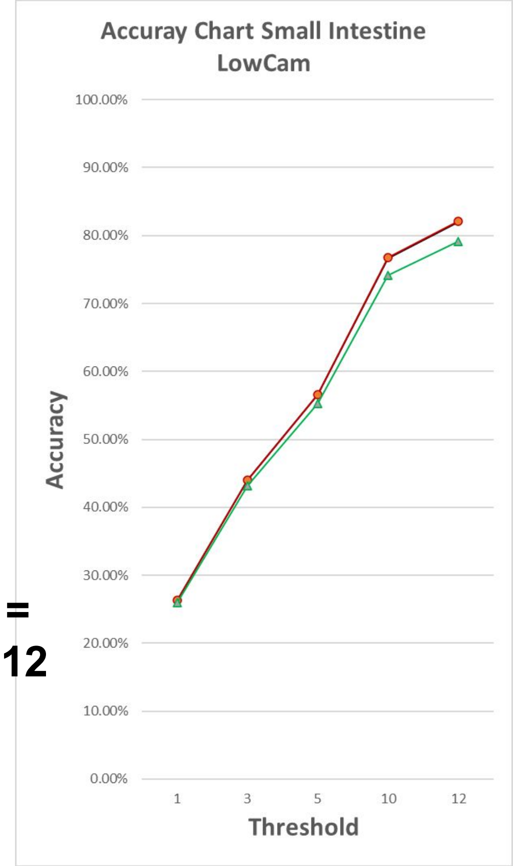


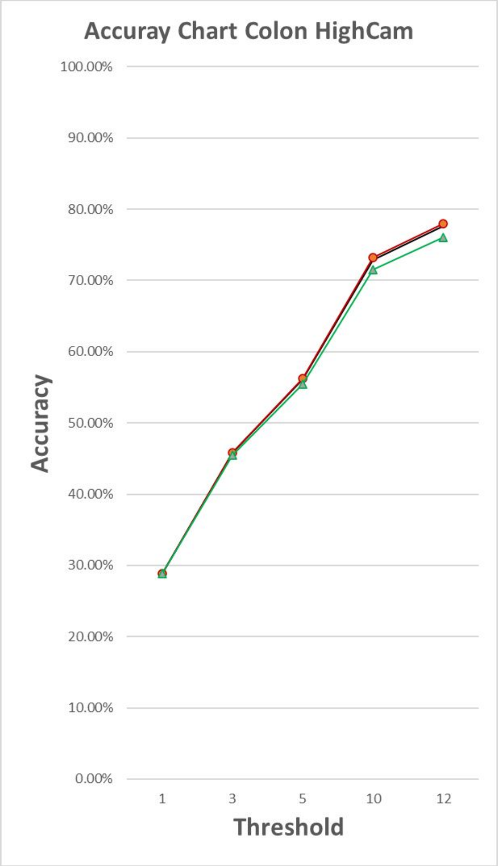
There are some **'X' shape** matching lines cross the others, which indicate that they are **wrong matches**.

Accuracy Evaluation Results: (Part of the evaluation results)


Accuray Chart Colon LowCam

Threshold =
1, 3, 5, 10, 12


Accuray Chart Small Intestine LowCam

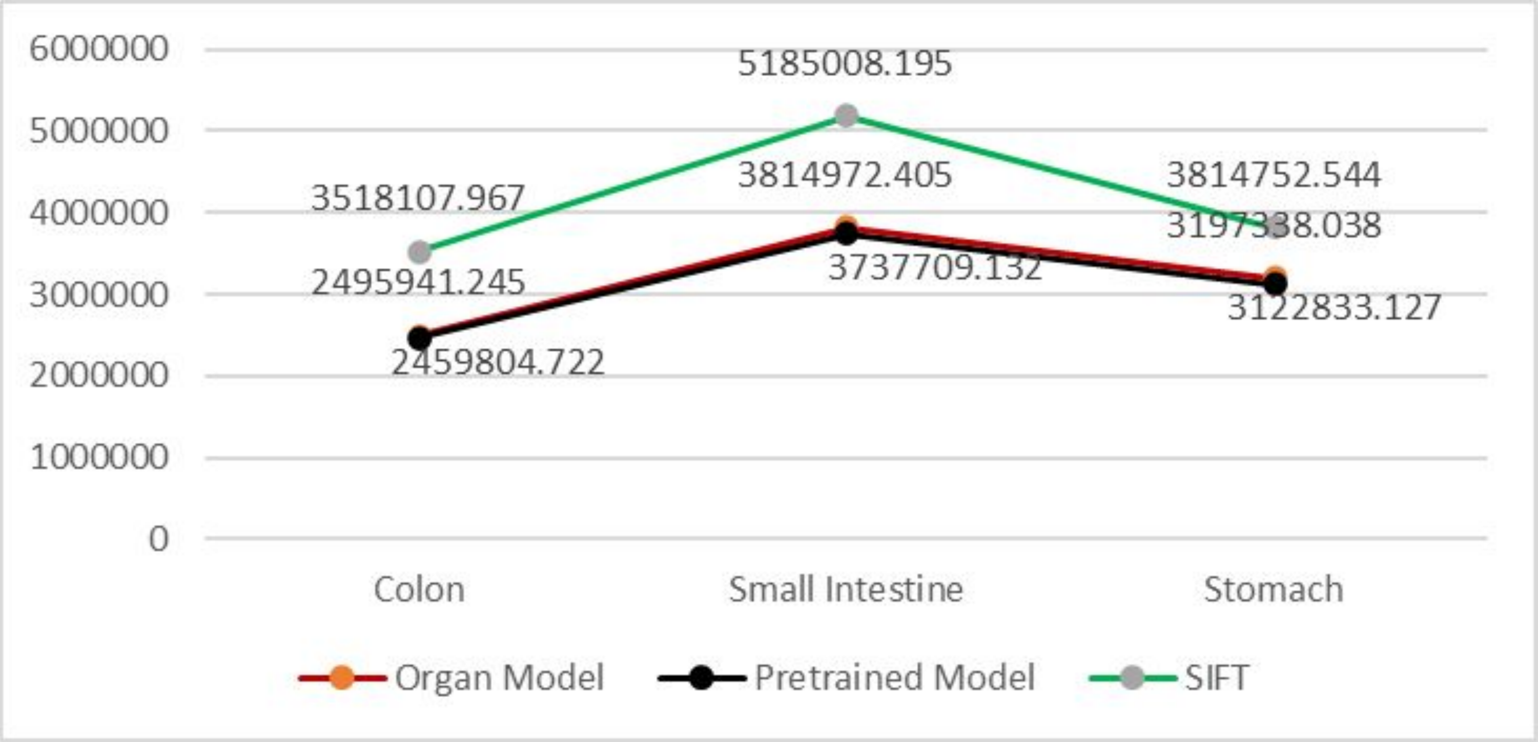

Accuray Chart Colon HighCam

Organ Model

Pretrained Model

SIFT

Distance Evaluation Results: (Part of the evaluation results)



Organ Model

Pretrained Model

SIFT

Result Analysis:

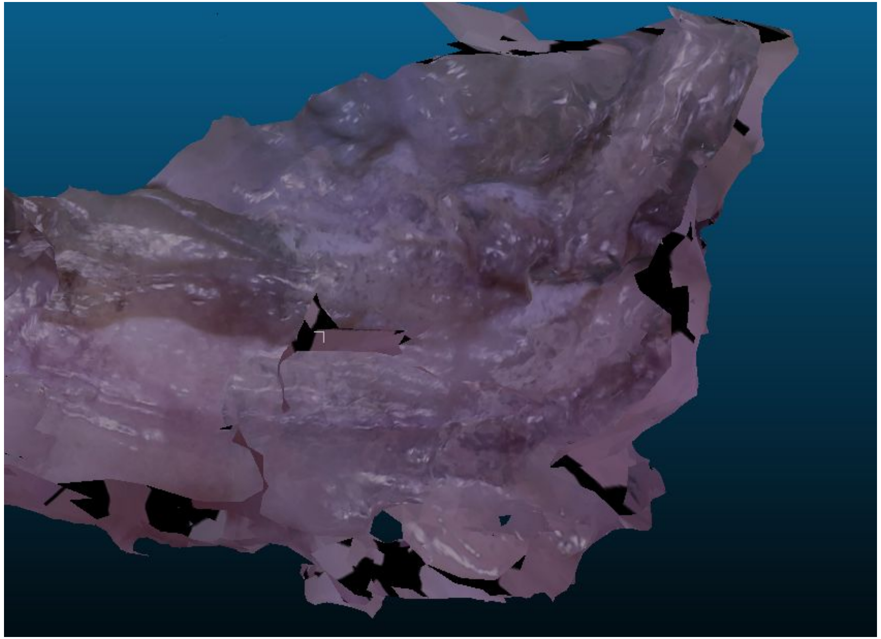
School *of* Engineering

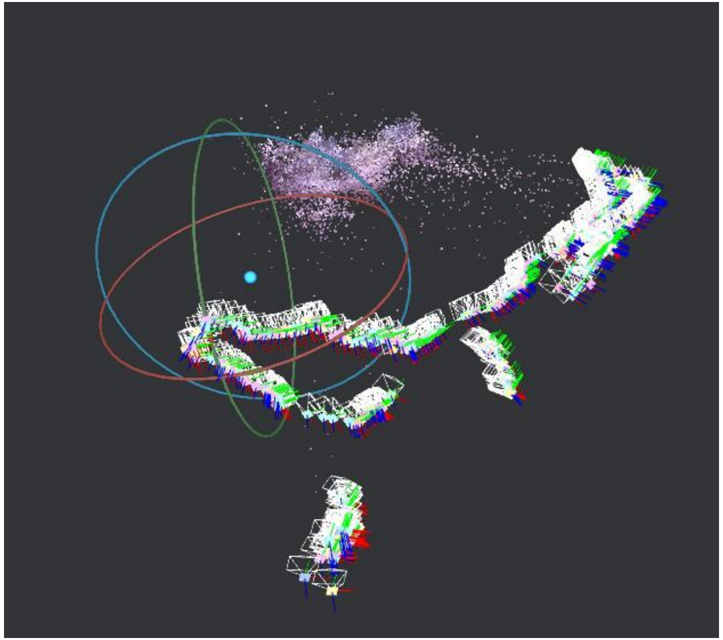
- As seen in the chart, the improvement in accuracy is minor. Basically, the improvement is around **1~2%** in accuracy when threshold is set to **5 pixel**.
- Since we extract around 800 features in this dataset, this method outputs around **10 more pairs per frame**.
- However, the improvement towards the total distance of feature points and corresponding epipolar lines is obvious, the decrease rate is around **25-30%**.
- While in practice, we wouldn't take every point into matching, even with SIFT matching. We normally sort SIFT features and take part of matches with higher score, and it is effective to filter out the feature untrustworthy. Thus, the decrease in total distance is of little practical value to matching.

The purpose of matching:

School *of* Engineering

The matching points in consecutive frames is needed for 3d reconstruction, as shown below.
The details of the organ can be clearly seen through the model and help diagnosis of possible disease.





Conclusion:

The details is important since the disease spot could be small, thus we should still cherish the little increase in accuracy.

To reconstruct the organ which requires hundreds of features points, and obviously with higher accuracy comes with better reconstruction performance.

However, considered of the computing consumption is relatively high in this method, I won't say it is a promising method to deal with endoscopy matching problem under this dataset.

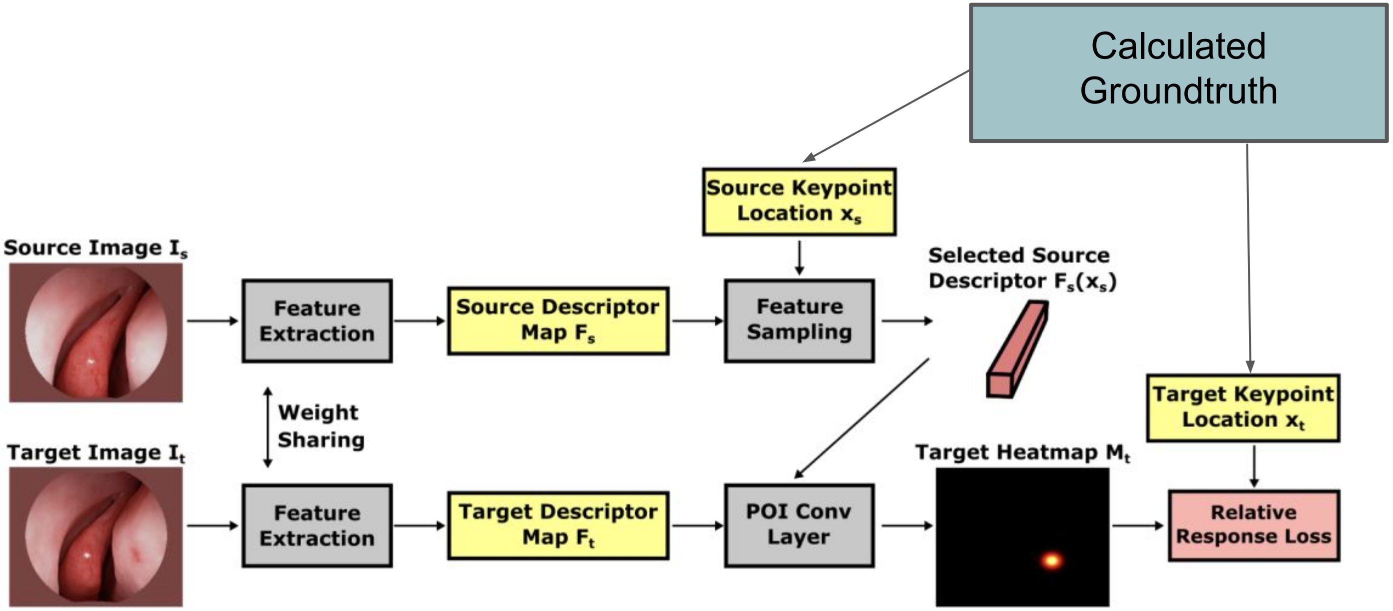
The relatively unsatisfied accuracy improvement could be caused by overfitting of the model, the methodology of the training principle is basically effective.

With further investigation and improvements, the results could be better.

Secondary Paper:

Extremely Dense Point Correspondences using a Learned Feature Descriptor

Principles - Secondary:



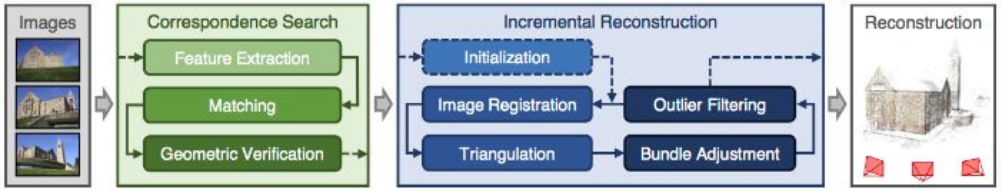
Groundtruth Matching

Groundtruth matching used in feature sampling and loss evaluation are estimated:

- COLMAP apply SfM method on video to generate 3D reconstruction
- Project the sparse 3D reconstructions onto the image planes
- Generate pairwise feature correspondence

COLMAP workflow

- Generate sparse and dense reconstruction with images and camera pose
- Build Database for faster I/O



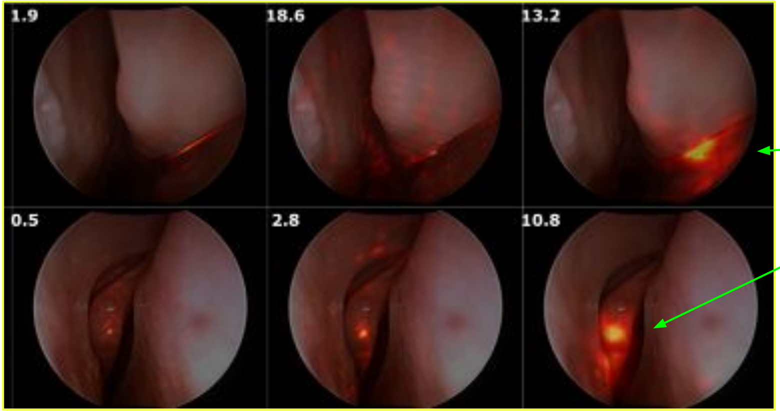
SfM(Structure-from-Motion)

- Feature detection and extraction
- Feature matching and geometric verification
- Structure and motion reconstruction

Point of Interest Layer and Heatmap

Conversion of Descriptor Learning to Key point Matching

- Extract dense feature descriptors
  - Source descriptor size = C x H x W, target descriptor size = C x 1 x 1
- 1 x 1 Convolutional kernel on target descriptor to generate Heatmap
- Heatmap stores similarity between the source descriptor and every target descriptor



Generated heatmap representing estimated matching on descriptors with various pixel error

Evaluation Metrics

Relative Response Loss

- The loss is proposed with the intuition that a target heatmap should present a high response at the groundtruth target keypoint location and the responses at other locations should be suppressed as much as possible.
-

$$\mathcal{L}_{\mathrm{rr}} = -\log\left(\frac{\mathrm{e}^{\sigma \mathbf{M}_{\mathrm{t}}(\mathbf{x}_{\mathrm{t}})}}{\sum_{\mathbf{x}} \mathrm{e}^{\sigma \mathbf{M}_{\mathrm{t}}(\mathbf{x})}}\right)$$

Spatial Softmax to speed up convergence

Matching Score

- (number of Inliner Matches) / (number of Features)
- Matching margin set to 10px from ground truth keypoint to target descriptor

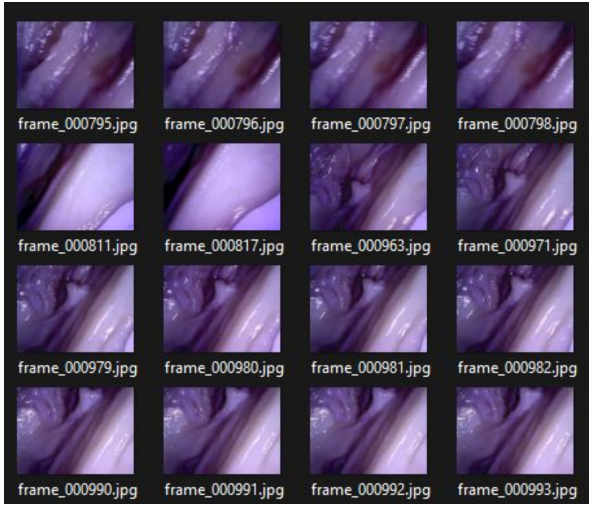
Testing Details

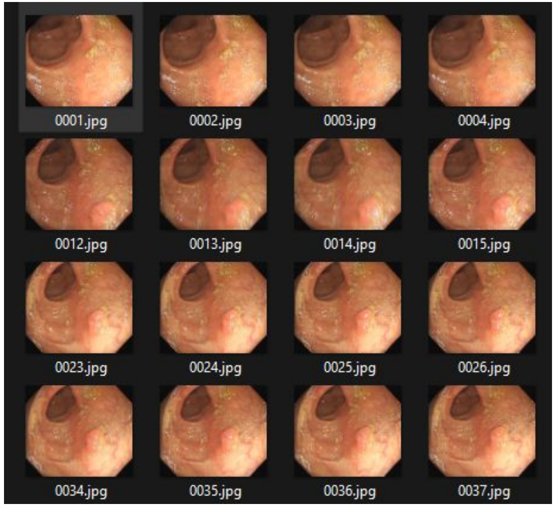
Endo-Slam dataset and **LDPolypVideo-Benchmark dataset** is our major testing datasets. Colonoscopy dataset doesn’t come with camera poses, thus cannot be pre-processed with COLMAP script.

Testing on Endo-Slam dataset consist of images taken by high camera and low camera. Colon-IV, small intestine and Stomach-I are the chosen organs, each contains 4 to 5 tumor free trajectories. 28 image sequences were tested in total.

Testing on LDPolypVideo-Benchmark dataset mostly focused on T1 and T2 dataset. Each dataset consist of multiple low-texture endoscopic videos with over 700 frames. Total number of image sequence tester is 10.

Dataset Demo



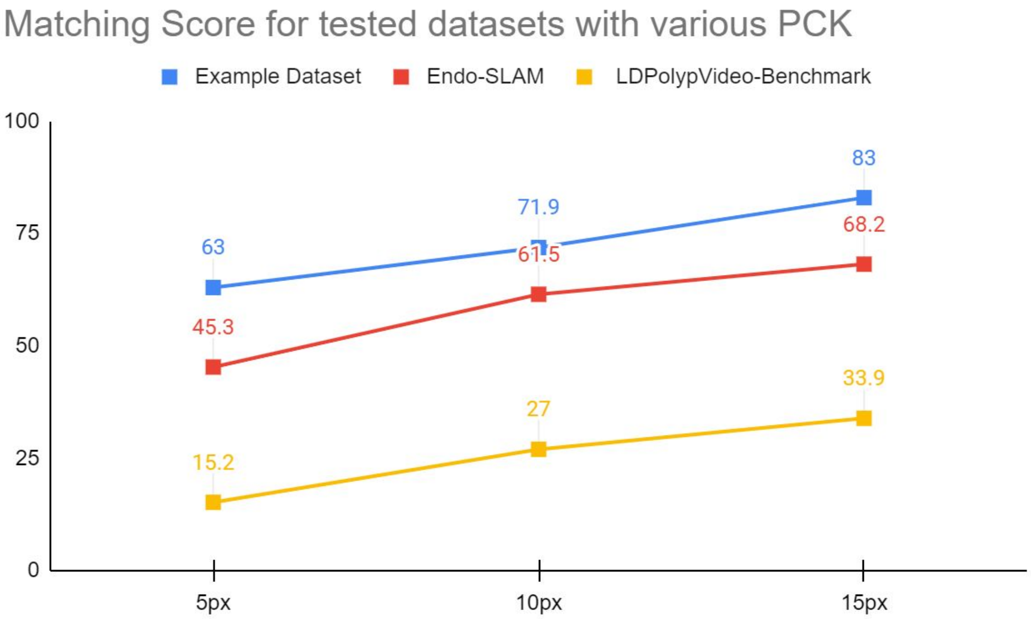


Endo-Slam dataset

LDPolypVideo-Benchmark dataset

Result - Secondary

Testing with pre-trained model constraints the parameters of predictions.
Scores on two tested datasets are averaged



```
Testing parameter
–adjacent_range 1 50
--image_downsampling 4.0
--network_downsampling 64
--input_size 256 320 --id_range 1
--batch_size 4 --num_workers 4
--num_pre_workers 4
--lr_range 1.0e-4 1.0e-3
--validation_interval 1
--rr_weight 1.0
--inlier_percentage 0.99
--training_patient_id 1
--testing_patient_id 1
--validation_patient_id
--feature_length 256
--filter_growth_rate 10
--matching_scale 20.0
--cross_check_distance 5.0
--heatmap_sigma 5.0
--visibility_overlap 20
```

Statistic Analysis :

Performance on two testing datasets are not desirable as example datasets. Reasons can be conclude as:

- Model falls to overfitting issue. Result have a high bias and relatively low variance toward training dataset.
- Pre-set parameters for testing is constrained, thus the best-performance-setup is not achieved. The author also mentioned this variance on preset in the essay, and the data collected for comparison was the best-performance setup. However, the testing process in original paper also contains training for each datasets, which makes the pre-trained model meaningless.
- Low quality of datasets. Network doesn’t solve texture scarcity issue completely.

Conclusion

School *of* Engineering

Performance Variance among Datasets

- The model solved the issue of texture to some extends and performances extraordinary with stable illumination and texture condition.
- Project is deployed as a system but not a model. For the best performance, optimal parameters need to be extracted from multiple examinations,
- Training process is required for individual datasets is required for high-quality 3D reconstruction.

Future Improvement

- To continue the experiment, the next step can be automatically fitting the dataset with best-performance parameter setup. Using adaptive hyper-parameters for future training would improve the accuracy in various condition.

Labor Division Table

	Tengfei Jiang	Xingjian Hao
Recreation of primary experiment	□	
Recreation of secondary experiment		□
Testing on new datasets	□	
Analyzing Results	□	

Thanks for watching!

Reference:

School *of* Engineering

[1] Farhat, Manel, Houda Chaabouni-Chouayakh, and Achraf Ben-Hamadou. "Self-supervised endoscopic image key-points matching." *Expert Systems with Applications* 213 (2023): 118696.

[2] Liu, Xingtong, Yiping Zheng, Benjamin Killeen, Masaru Ishii, Gregory D. Hager, Russell H. Taylor, and Mathias Unberath. "Extremely dense point correspondences using a learned feature descriptor." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4847-4856. 2020.

[3] Kutsev Bengisu Ozyoruk, Guliz Irem Gokceler, Taylor L. Bobrow, Gulfize Coskun, Kagan Incetan, Yasin Almalioglu, Faisal Mahmood, Eva Curto, Luis Perdigoto, Marina Oliveira, Hasan Sahin, Helder Araujo, Henrique Alexandrino, Nicholas J. Durr, Hunter B. Gilbert, Mehmet Turan. “EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos.” Medical Image Analysis, Volume 71, 2021, 102058,ISSN 1361-8415,

[4] Anita Rau, P. J. Eddie Edwards, Omer F. Ahmad, Paul Riordan, Mirek Janatka, Laurence B. Lovat, Danail Stoyanov. “Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy.” International Journal of Computer Assisted Radiology and Surgery (2019) 14:1167–1176

[5] Yiting Ma, Xuejin Chen , Kai Cheng, Yang Li, and Bin Sun. “LDPolypVideo Benchmark: A Large-Scale Colonoscopy Video Dataset of Diverse Polyps.” National Engineering Laboratory for Brain-inspired Intelligence Technology and Application