Investigate learning based end2end localisation methods in Colonoscopic Surgery

Supervisor: Jinjing Xu
Team: (TU-Dresden: CMS Team Project)
Prerana Chandratre
Anjali Sharma
Zhifan Yu
Jingjun Huang

NCT
NATIONAL CENTER
FOR TUMOR DISEASES
PARTNER SITE DRESDEN
UNIVERSITY CANCER CENTER UCC

Supported by:
German Cancer Research Center
University Hospital Carl Gustav Carus Dresden
Carl Gustav Carus Faculty of Medicine, TU Dresden
Helmholtz-Zentrum Dresden-Rossendorf

Introduction

End-to-End Localization in Colonoscopic Surgery

General Localization Pipeline:

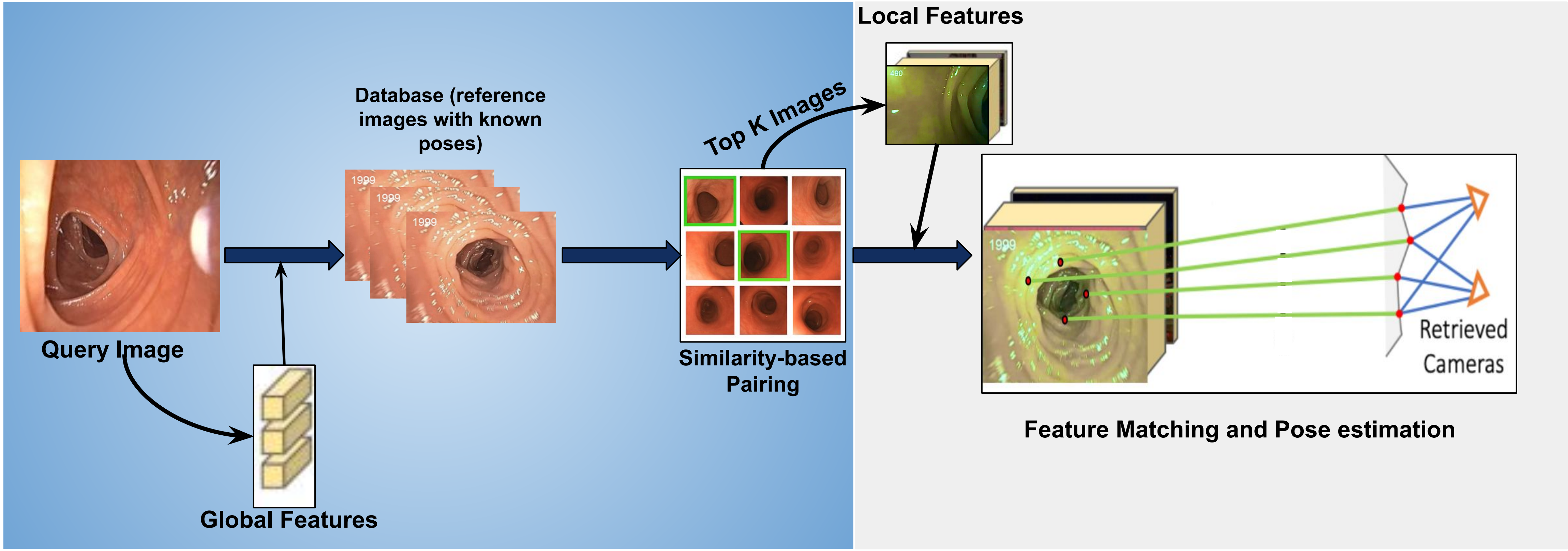
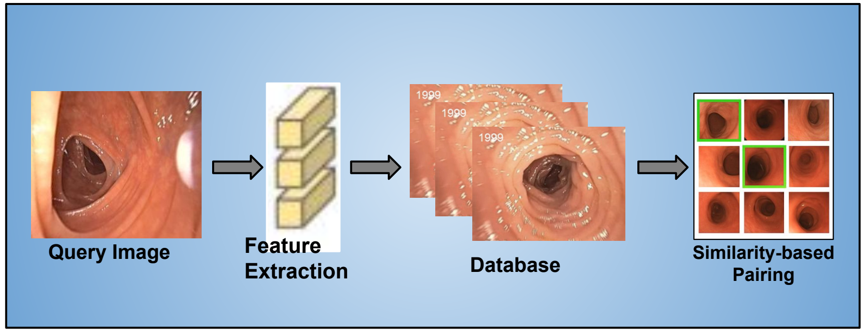


Image Retrieval

Pose Estimation

Existing **Approaches:**

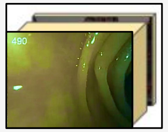
Existing approaches vary in how the general retrieval-localization pipeline is implemented.



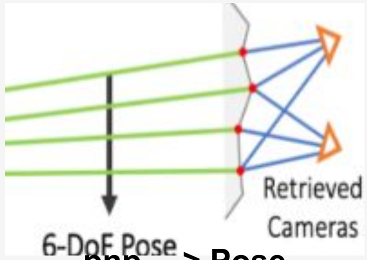
IR

1. Hierarchical Localization

Local Features



Feature Matching

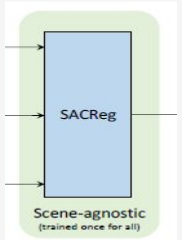


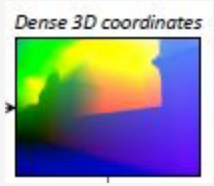
pnp —> Pose

✅ Interpretable

❌ Fails in low texture

2. Scene Coordinate Regression (SACReg)

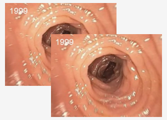




pnp —> Pose

✅ Robust to low-texture and specularities

3. Relative Pose Regression (Reloc3r)



Network

6 DoF Pose

✅ No PnP needed

Motivation

Our Focus

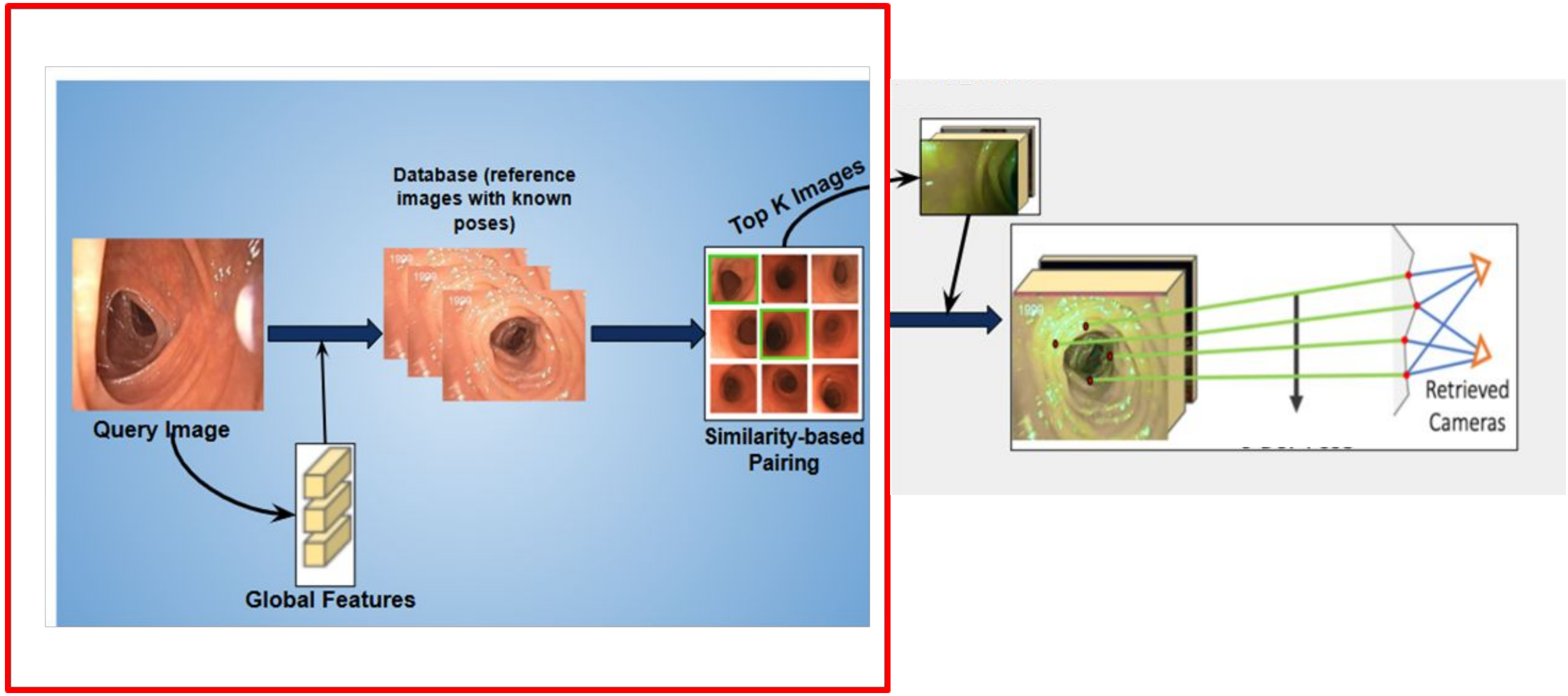
- Image Retrieval for localization with Pose Regression
- Exploring alternative Pose Estimation

Addressing →

Challenges

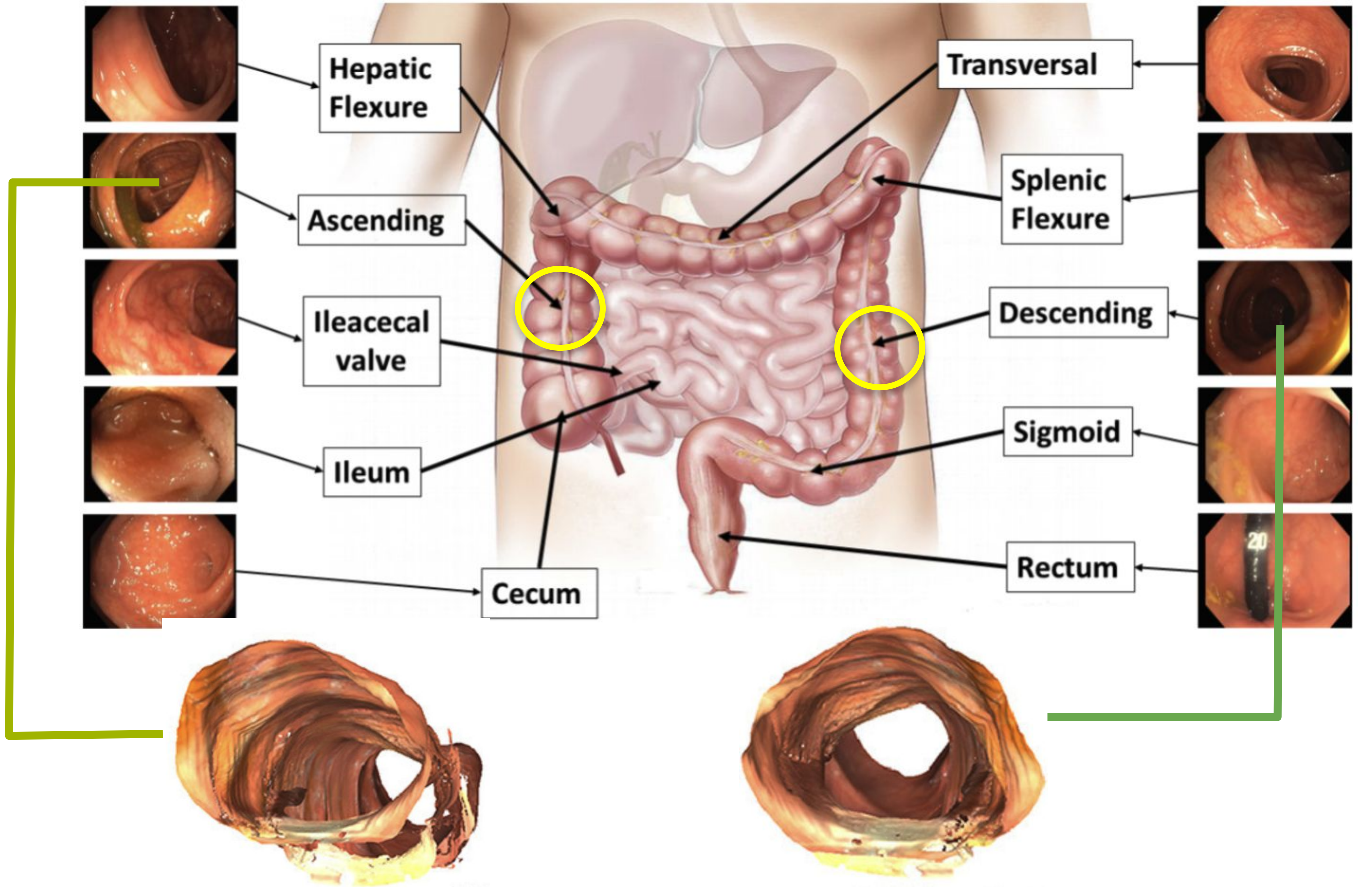
- Deformable envs
- Fluid
- Low Texture
- Repetitive
- Occlusion

Task 1:
Investigation on Image Retrieval



IR and Dataset Overview

IR overview

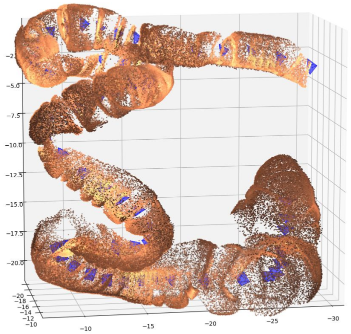


In short: IR helps to find a rough area for camera localization

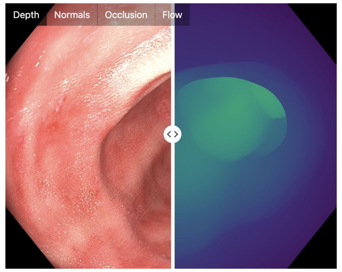
https://www.sciencedirect.com/science/article/pii/S1361841521001468
https://arxiv.org/pdf/2204.14240

Dataset Introduction

Train data

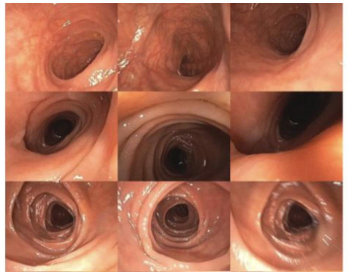


- Simcol3d
  - Simulated data
  - 37,000 frames



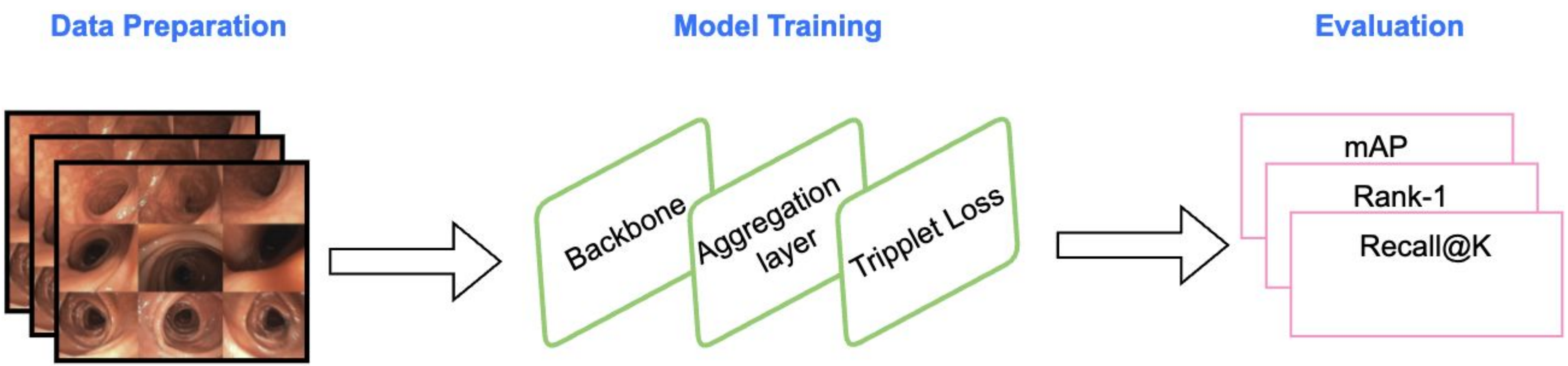
- C3VD
  - Based on real models
  - 10,015 frames

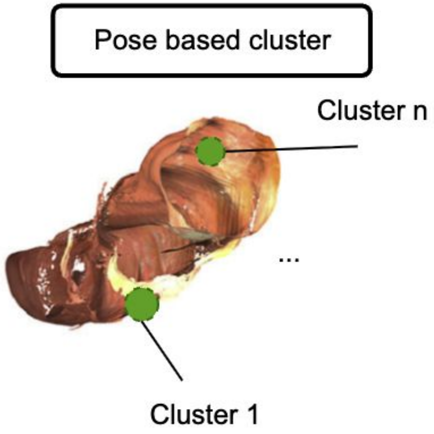
Test data



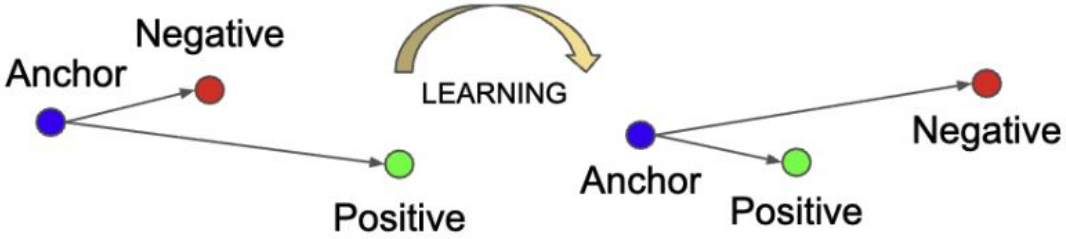
- Colon10k
  - Real colon data with expert labels
  - 10126 frames

IR Architecture and implementation details





$$D(A,P) \ll D(A,N)$$



$$L(A,P,N) = \max(0, D(A,P) - D(A,N) + \text{margin})$$

ViT V.S ResNet
- Transformer & CNN
- Global context & local details

GeM V.S Netvlad
- Global pooling method & aggregate local features
- Simple and efficeint & Complex and rich information

https://arxiv.org/abs/1503.03832

IR results and analysis

	Test on: Conlon10k	
	mAP	RANK-5
ViT + GeM	54.02%	99.36%
ViT + NetV	12.01%	33.10%
ResNet + Netv	7.34%	44.05%
ResNet + GeM	29.35%	72.20%

$$mAP = \frac{1}{|Q|} \sum_{Q \in Q} AP_Q$$

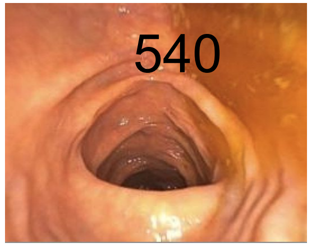
$$\text{Rank-5} = \frac{\sum_{q \in Q} \text{is\_correct}(q, 5)}{|Q|}$$

$|Q|$: The total number of queries in the evaluation query set.

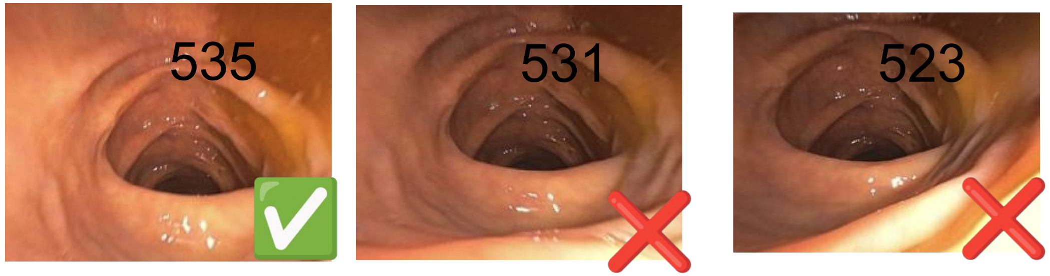
ResNet + NetV Evaluation

- Strictness of Evaluation
  - Define expert's manual annotation as same physical area
- Limitations of ResNet backbone
  - Focus on local features

Query image



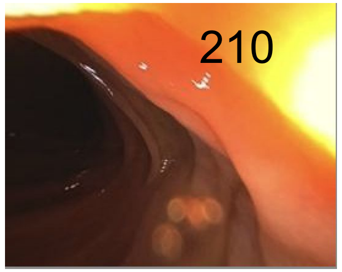
Top-10 Result Examples for ResNet+ NetV



Ground Truth(533-549)

ViT + GeM: A Deep Dive into Success and Challenges

Query



TOP-10 retrieval and similarity results

Ground truth
(206-215)

Conclusion & Success: Top-5 are all above 84% similarity, and it is capable to deal with minor ambient light and position variance.

Unsolved issues: mAP score is not perfect; real dataset with ground truth is precious.

Validate IR in Localization Pipeline
IR (NetVLAD) + Reloc3r

IR: NetVlad (baseline) + Reloc3r

Evaluation Data:

- SimCol3D Data: Synthetic Colon_III

Metrics:

- Absolute Translation Error(ATE)
- Relative Pose Error (RPE)

IR Method	ATE	RPE
NetVLAD (baseline)	8.015 m	15.34°
ViT+GeM (Ours)	6.0092 m	11.76°

Translation Accuracy (ATE)

- **ATE dropped by ~25%** when using ViT+GeM instead of NetVLAD.
- Model **predicts the camera location more precisely** with ViT+GeM.

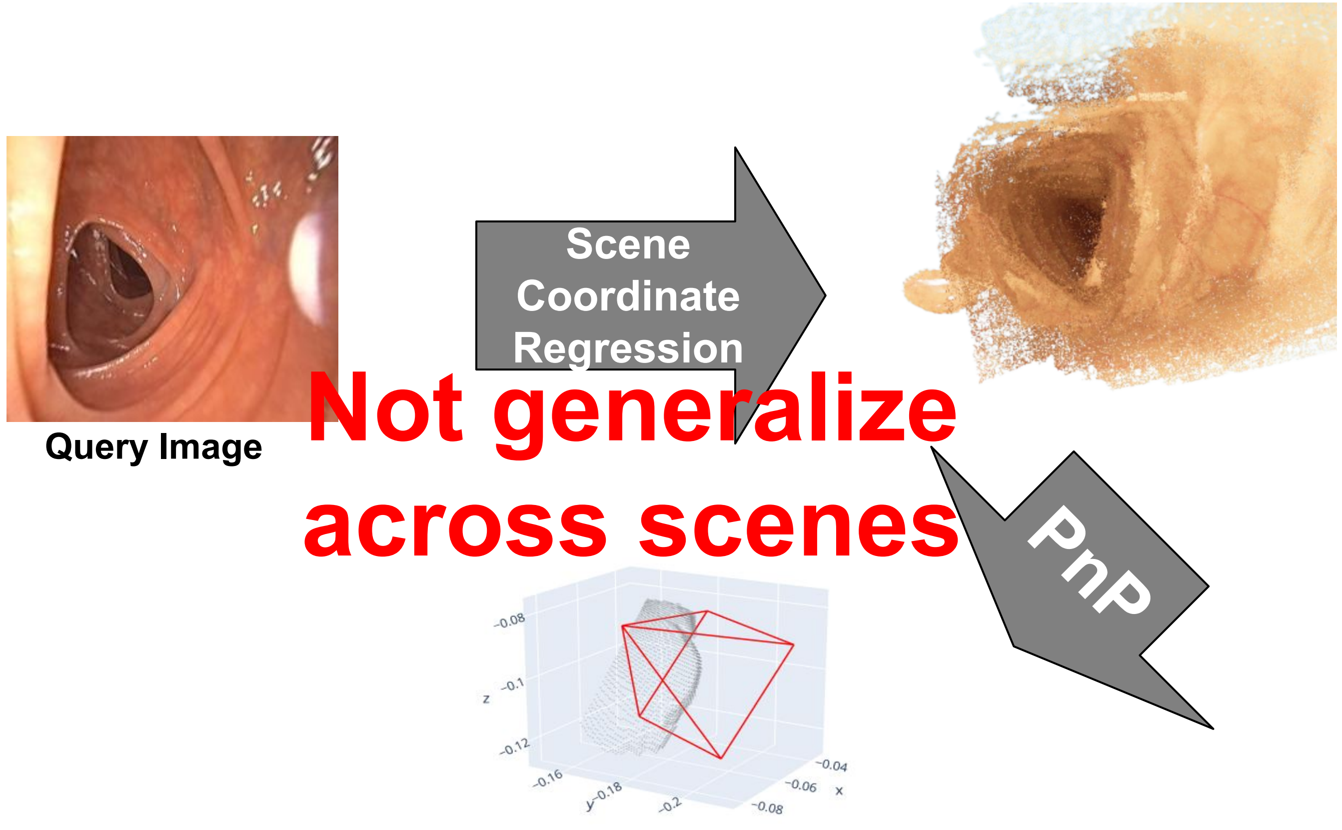
Rotation Accuracy (RPE)

- **RPE dropped by ~23%**, from 15.34° to 11.76°.
- Indicates **better angular alignment** between predicted and ground-truth poses using ViT+GeM.

NCT

Localization Method — Investigating SACReg

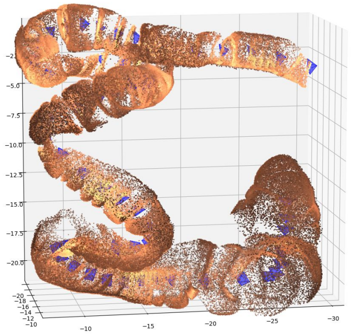
Scene Coordinate Regression overview



Not generalize across scenes

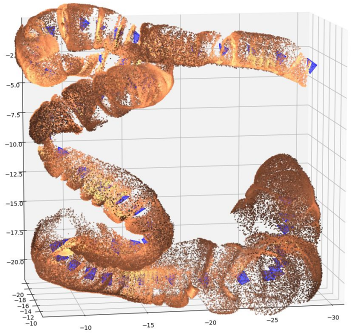
Dataset Used

Train data



- Simcol3d
  - Simulated data
  - 37,000 frames

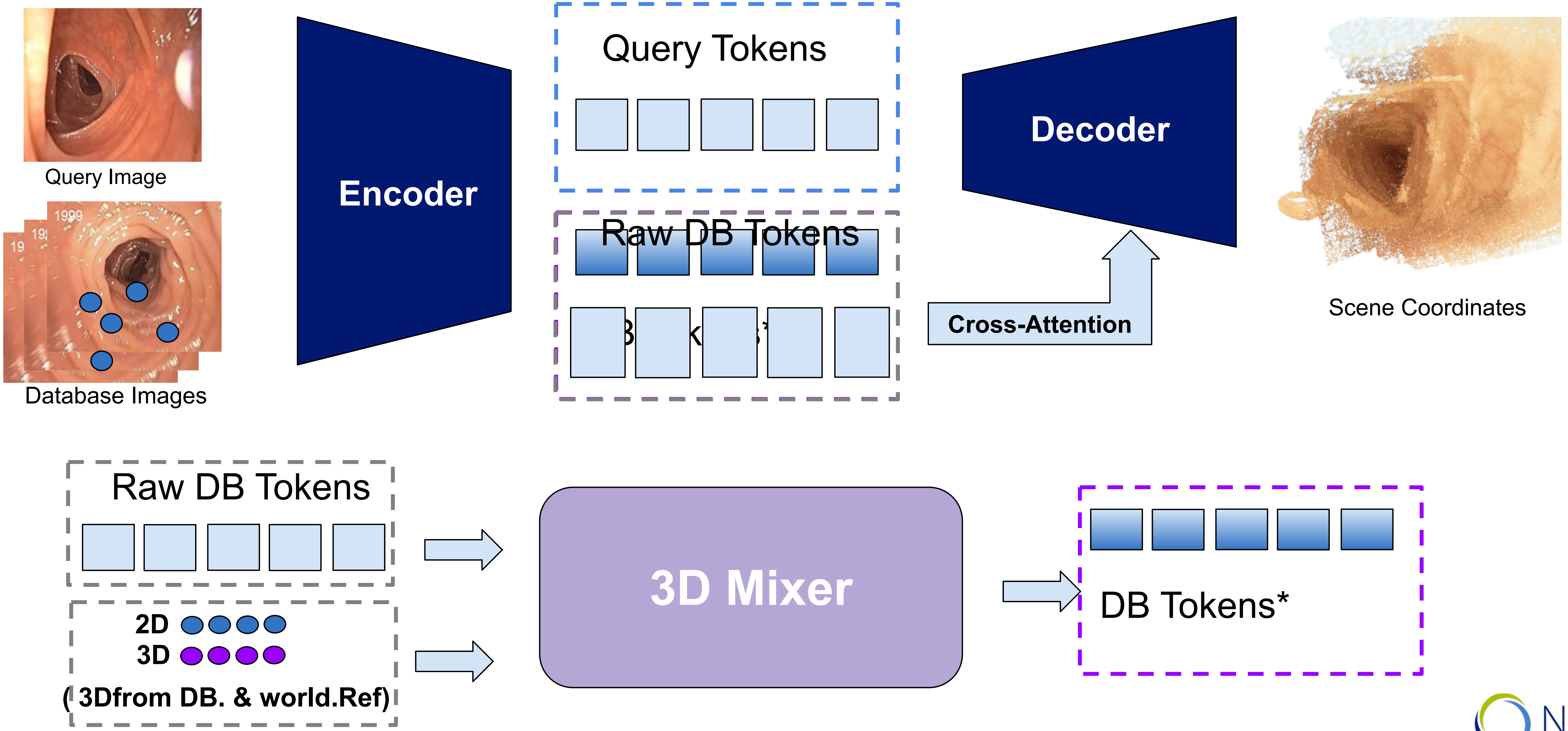
Test data



- Simcol3d
  - Official test split
  - 600 DB frames
  - 1200 queries

In short: SCR have the issue, while SACReg enhance the cross scene generalization.

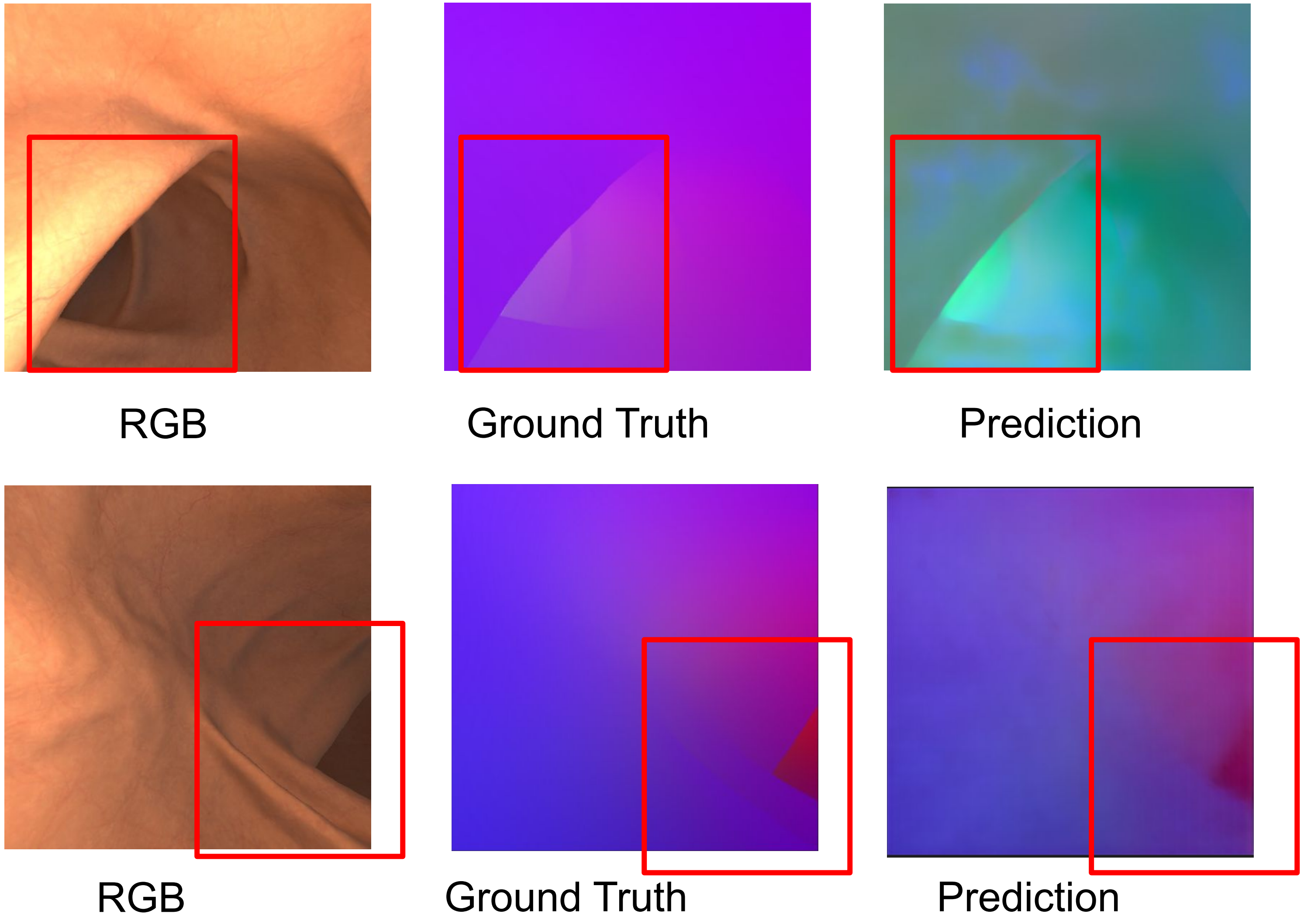
SACReg Architecture and 3D Points Embedding



SACReg results and analysis

- Finetuning:
  - n = 28,824(image pairs)
  - coordinate shifted uniformly
  - ~38cm/134°

- Overfitting:
  - n=16
  - Good prediction
  - ~2cm/30°

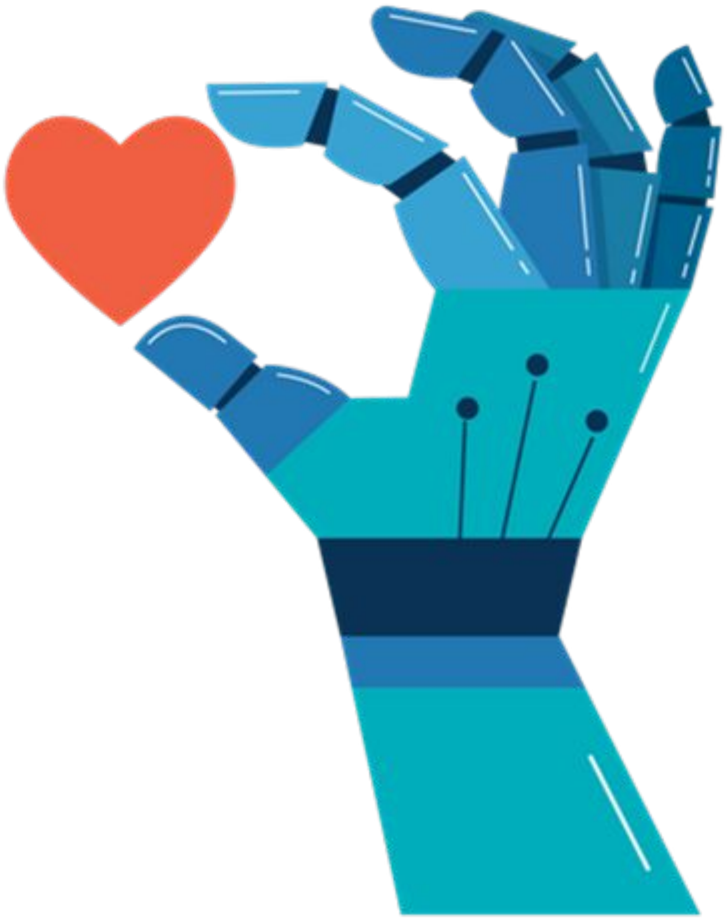


Visualization with normalized scene coordinate

Conclusion

- We contributed an Image Retrieval model enhances localization accuracy, when combined with a **pose regression-based** foundation model.
- We investigated the feasibility of using **Scene Coordinate Regression**—another category of pose estimation which enhanced explainability—for localization.

THANK YOU



References

Datasets:

1. https://durrlab.github.io/C3VD/
2. https://rdr.ucl.ac.uk/articles/dataset/Simcol3D_-_3D_Reconstruction_during_Colonoscopy_Challenge_Dataset/24077763
3. https://www.synapse.org/Synapse:syn26707219

Models:

1. REVAUD, Jerome, et al. **Sacreg: Scene-agnostic coordinate regression** for visual localization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024. S. 688-698.
2. DONG, Siyan, et al. **Reloc3r: Large-scale training of relative camera pose regression** for generalizable, fast, and accurate visual localization. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025. S. 16739-16752.
3. SARLIN, Paul-Edouard, et al. From coarse to fine: Robust hierarchical localization at large scale. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019. S. 12716-12725.
4. RUIZ, Lina, et al. COLON: The largest COlonoscopy LONg sequence public database. *arXiv preprint arXiv:2403.00663*, 2024.

NCT