Endo-FASt3r: Endoscopic Foundation model Adaptation for Structure from motion

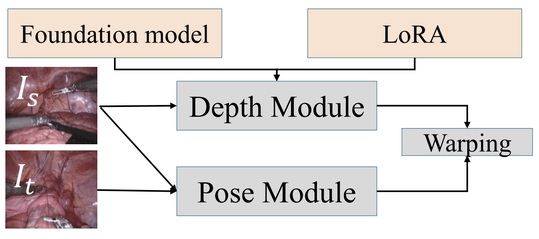
Mona Sheikh Zeinoddin[1,2], Mobarak I. Hoque[1,4], Zafer Tandogdu[3,6], Greg L. Shaw[3], Matthew J. Clarkson[1,4], Evangelos B. Mazomenos[1,4], Danail Stoyanov[1,5]

[1] Hawkes Institute, University College London, London, UK [2] Institute of Health Informatics, University College London, London, UK [3] Dept. of Urology, University College London Hospitals, London, UK [4] Dept. of Medical Physics & Biomedical Engineering, University College London, UK [5] Dept. of Computer Science, University College London, UK [6] Division of Surgery and Interventional Science, University College London, UK

MICCAI2025
28th International Conference on Medical Image Computing and Computer Assisted Intervention
23-27 September 2025
Daejeon Convention Center
Daejeon REPUBLIC OF KOREA
UCL

Motivation & Background

- **Camera Pose & Depth estimation** is essential to achieve 3D scene understanding in **robotic-assisted surgery.**
- The **self-supervised reprojection loss** [2] pipeline is widely used in the surgical domain due to the lack of ground truth data.

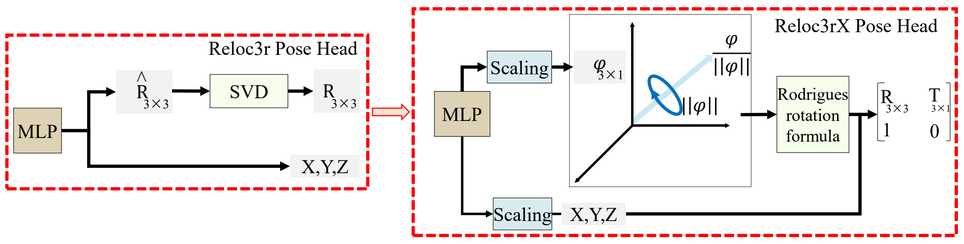


- Previous works have investigated the use of foundation models in the **depth module** via **Low-Rank Adaptation (LoRA)** [1] **–based** techniques.
- **Major limitations** of current literature:
  - No work has investigated the use of **foundation models** in the **pose module**.
  - The **limiting low rank update space** of **LoRA-based approaches.**

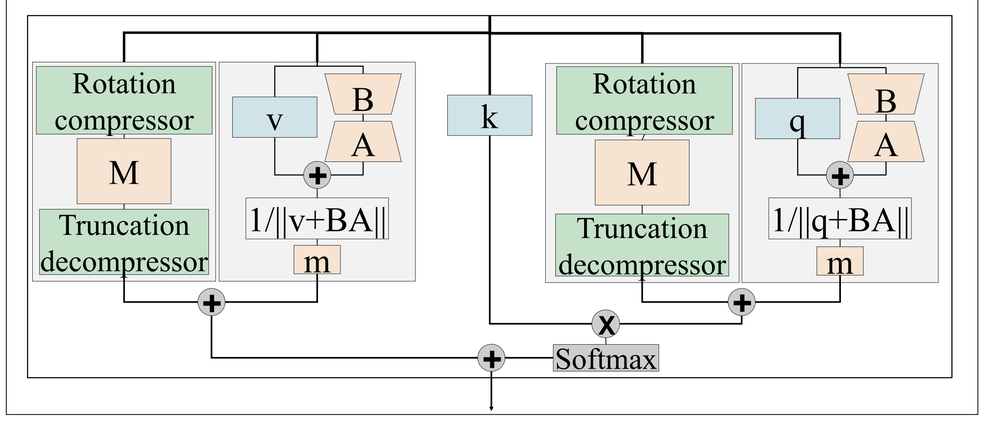
Endo-FASt3r Contributions

- In this work, **Endo-FASt3r: Endoscopic Foundation model Adaptation for Structure from motion** , we introduce:

- **Reloc3rX**: Extending the foundation model Reloc3r[3] by designing the **Axis Pose Head** to address **scale-mismatch.**



- **DoMoRA**: Enabling both **Low- and High-rank updates**.



References

[1] Hu, Edward S, et al. "LoRA: Low-Rank Adaptation of Large Language Models." In *ICLR* (2022)
[2] Shao, Shuwei, et al. "Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue." In *ICRA* (2021)
[3] dong, Siyan, et al. "Reloc3r." In *CVPR* (2025)

Let's discuss

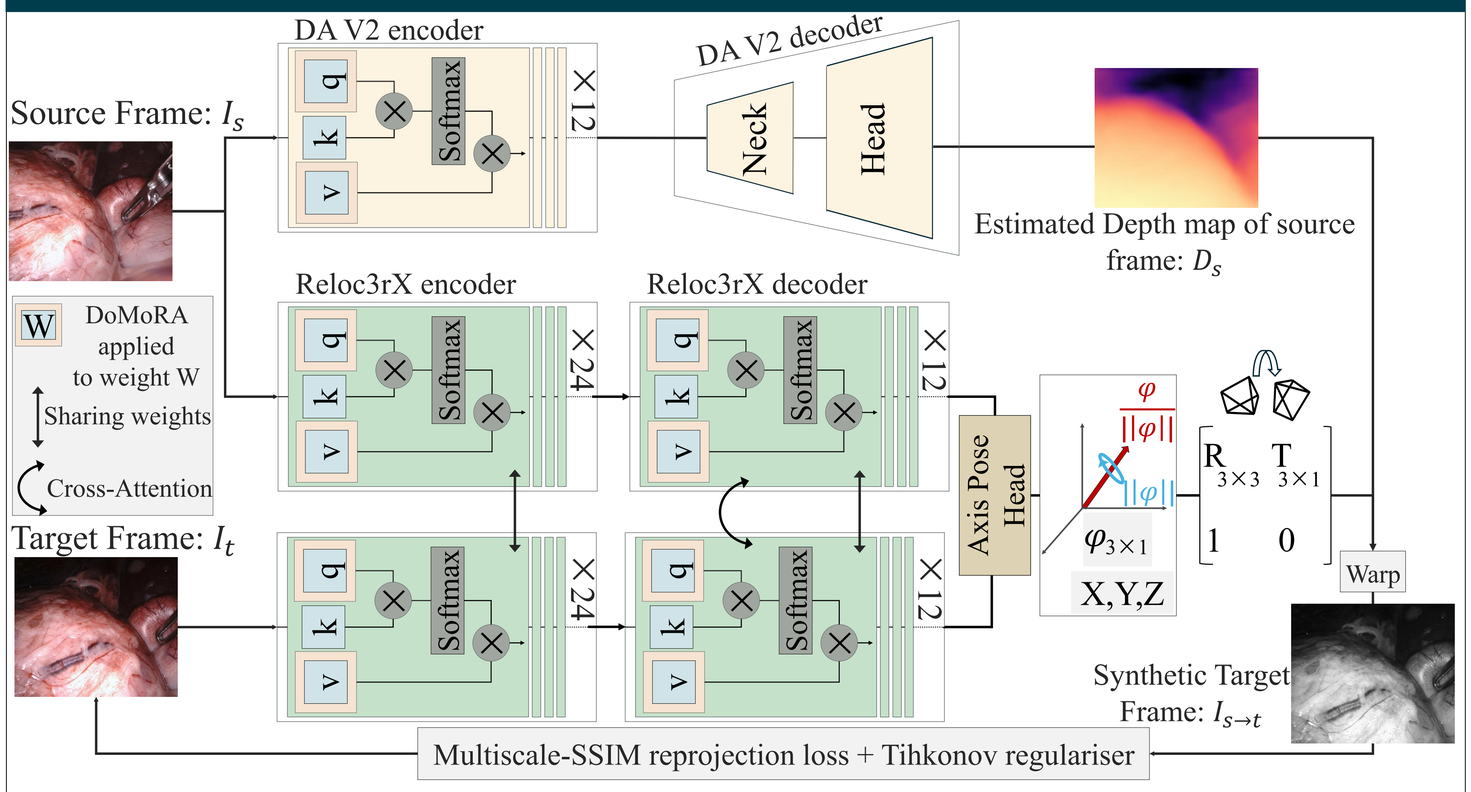
Can you answer these questions:
- Why was the **change to Reloc3r's pose head** necessary?
- We used **DUSt3r, MASt3r and MICKEY** instead of **Reloc3r** and none worked, what do you think was the reason?
- What is the main difference of **LoRA and DoMoRA**?

Interested in our work?

Github :

Paper:

Endo-FASt3r Architecture



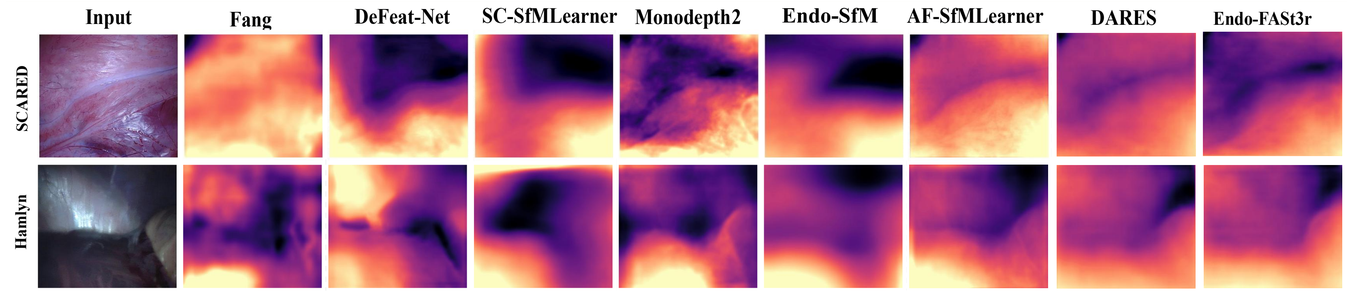
Results & Conclusion

- **Trained** on **the SCARED dataset and Evaluation** performed on **the rigid SCARED, Hamlyn and non-rigid StereoMIS datasets.**

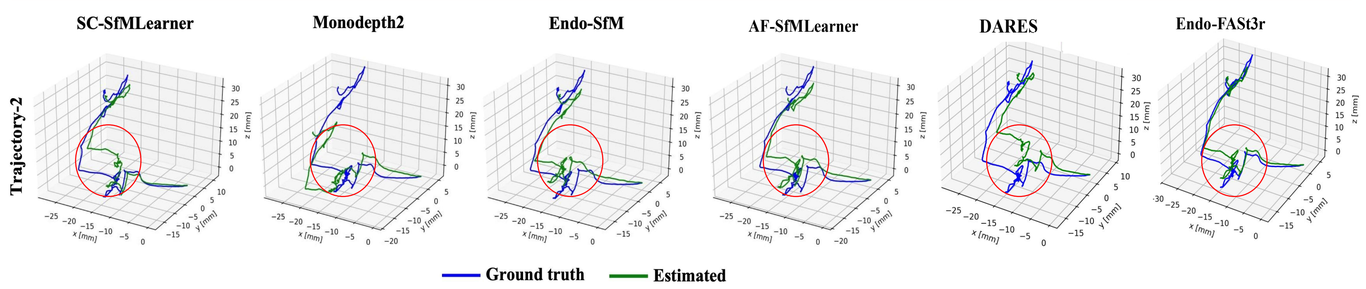
Comparison on **rigid scenes** with benchmark methods in depth estimation and pose estimation matrices

	Method	AbsRel↓	SqRel↓	RMSE↓	$\delta\uparrow$	ATE-T1 ↓	ATE-T2 ↓	Total	Train	Speed
SCARED	DeFeat-Net	0.077	0.792	6.688	0.941	0.1765	0.0995	14.8	14.8	-
	SC-SfMLearner	0.068	0.645	5.988	0.957	0.0767	0.0509	14.8	14.8	-
	Monodepth2	0.069	0.577	5.546	0.948	0.0769	0.0554	14.8	14.8	-
	Endo-SfM	0.062	0.606	5.726	0.957	0.0759	0.0500	14.8	14.8	-
	AF-SfMLearner	0.059	0.435	4.925	0.974	0.0757	0.0501	14.8	14.8	8.0
	Yang et al.	0.062	0.558	5.585	0.962	0.0723	0.0474	2.0	2.0	-
	Zero-Shot DA V2	0.091	1.056	7.601	0.916	-	-	-	-	-
	Zero-Shot Reloc3r	-	-	-	-	0.0938	0.0735	-	-	-
	DARES	0.052	0.356	4.483	0.980	0.0752	0.0498	24.9	2.88	15.6
	EndoFASt3r (Ours)	**0.051**	**0.354**	**4.480**	**0.998**	**0.0702**	**0.0438**	24.9	2.93	19.1
Hamlyn	Endo Depth & Motion	0.185	5.424	16.100	0.732	-	-	-	-	-
	AF-SfMLearner	0.168	**4.440**	13.870	0.770	-	-	14.8	14.8	7.7
	EndoFASt3r (Ours)	**0.166**	4.529	**13.718**	**0.778**	-	-	24.9	2.93	19.1

Qualitative comparison of our depth estimation on **rigid scenes** against benchmarks

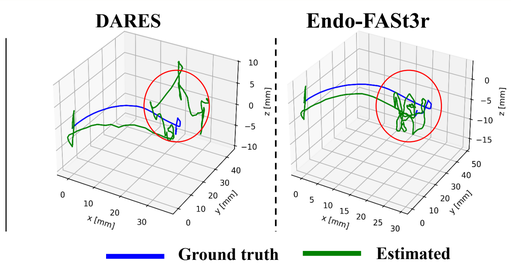


Qualitative comparison of our ego-motion estimation on **rigid scenes** against benchmarks



Comparison of our ego-motion estimation on **non-rigid scenes** against the second-best approach

StereoMIS	Method	ATE↓
	DARES	0.0715
	EndoFASt3r (Ours)	**0.0664**



Interested in seeing a video of our results?

- **Endo-FASt3r** marks the **first framework** to use foundation models for **pose estimation** in **surgical environments**, and it does so with **NO ground truth data.**
- **Endo-FASt3r** surpasses all SOTA methods, reaching an improvement of **9.34%** in **camera pose estimation** and **2%** in **depth estimation** over the nearest competitor.

Mona Sheikh Zeinoddin
mona.zeinoddin.22@ucl.ac.uk

Hawkes Institute
Engineering and Physical Sciences Research Council
EPSRC
UCL Health Informatics
SURGICAL ROBOT VISION research group