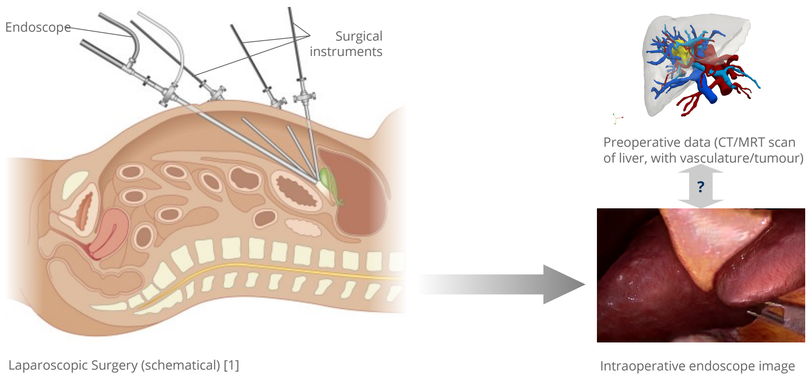
Fast High-Resolution Disparity Estimation for Laparoscopic Surgery

Jan Müller, Reuben Docea, Matthias Hardner, Katja Krug, Paul Riedel, Ronald Tetzlaff
Technische Universität Dresden / National Center for Tumor Diseases (NCT), Partner Site Dresden, Germany

Laparoscopic Liver Surgery



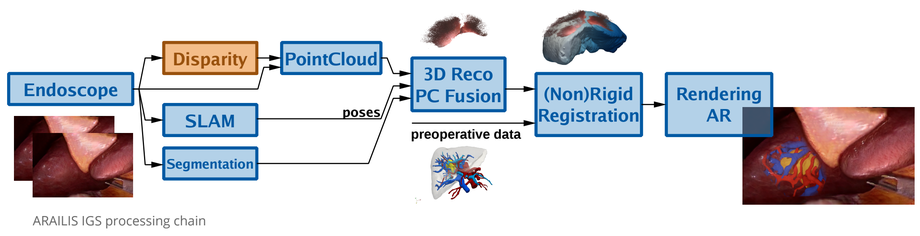
Advantages
- minimally invasive
- lesser morbidity
- faster patient recovery

Problems
- correlation image MRT/CT data
- location of vasculature/tumours
- motion/alterations of organs

→ Image Guidance System

Intraoperative Image Guidance System

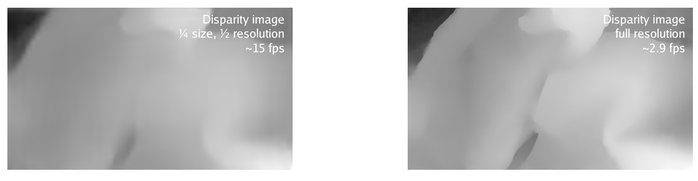
ARAILIS project: Augmented Reality and Artificial Intelligence supported Laparoscopic Imagery in Surgery



ARAILIS IGS processing chain

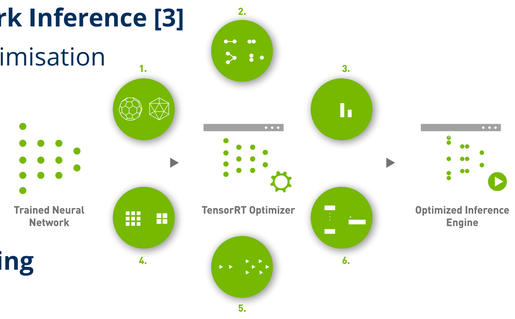
- Point Cloud generation/fusion requires *dense accurate* disparity estimation
- endoscope's *full framerate* at *FullHD resolution* necessary for smooth rendering

Disparity Estimation: Hierarchical Stereo Matching (HSM) neural network [2]



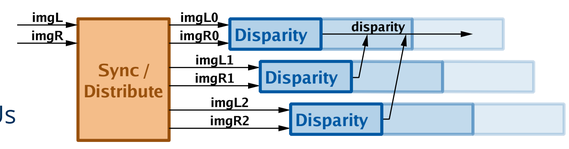
Acceleration Methods

NVIDIA TensorRT: Optimise Network Inference [3]
- GPU/CUDA-centric acceleration/optimisation
- layer/tensor fusion
- memory footprint minimisation
- mixed-precision data (FP32, FP16, INT8)



NVIDIA DALI: Network Pre-Processing
- pipelined input processing on GPU: image transposition, conversion, normalisation, padding/cropping
- streaming data transfer

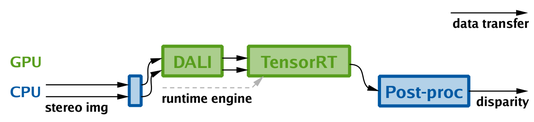
Multithreading/Multi-GPU
- time *synchronisation* of stereo images → pairs
- tagging of stereo pairs → *distribution* to threads / GPUs



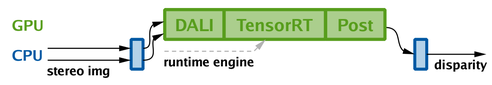
Implementations

ROS (Robot Operating System) Integration
- ROS/python callback function
- interface, inference, output

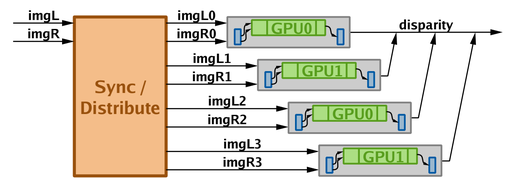
TensorRT & DALI
- build runtime engine once
- integration in ROS infrastructure
- DALI pipeline: DMA interface



GPU-centric implementation
- post-processing on GPU
- minimum data transfer: direct operation on results



Multithreading
- syncing of images by time stamps
- thread tagging/renaming
- ROS scripts: assign threads to GPUs



Experimental Results

Quality: Disparity Error

Models

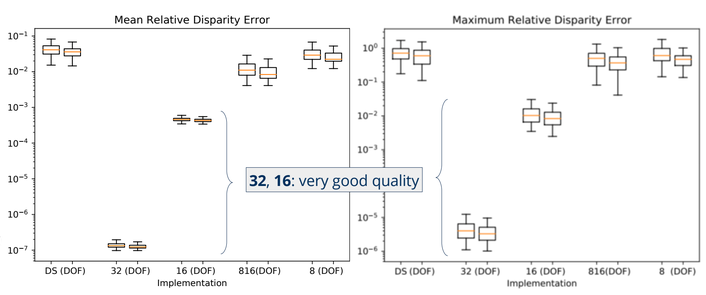
REF	original HSM algorithm, PyTorch, 32-bit float (FP32) data	FullHD resolution (1920×1080 px)
DS	original HSM algorithm (see REF), downscaled	1/16 of FullHD (480×270 px)
32	TensorRT engine & DALI pipeline, FP32 data only	FullHD
16	TensorRT & DALI, mixed-precision FP16 & FP32 data	FullHD
816	TensorRT & DALI, mixed-precision INT8 & FP16 & FP32	FullHD
8	TensorRT & DALI, mixed-precision INT8 & FP32	FullHD

Data Recording
- endoscope: EinsteinVision, frame rate 30 frames / s (fps), depth-of-field **DOF** [20 mm, 200 mm] → DOF disparity range [48.5, 485] px
- data set: 926 stereo images (31 s), with disparity range [3.24, 560] px

Error Measures
- disparity error to REF

$$e_d = \frac{d - d_{ref}}{d_{ref}}$$

- mean over image $\bar{e}_d$
- maximum in image $\hat{e}_d$ in whole image / DOF



Speed: Throughput

Setup
- workstation **XB:** CPU 2 × Xeon 4216, GPU 4 × RTX A5000
- measurement: average data volume (frames) per time (s)
- models, data set → see Quality, more workstations → see paper

Implementation model	Throughput / fps	
Original serial execution **REF** (single CPU/GPU)	2.9	
Downscaled serial model **DS** (single CPU/GPU)	14.9	downscaled for "acceptable" throughput
TensorRT & DALI (single GPU) Models **32** / **16** / **816**	8.3 / 21.5 / 27.3	significant acceleration
Multi-threaded / TRT & DALI (2 threads / 1 GPU)	- / 20.2 / 27.0	more threads no advantage
Multi-threaded / parallelised (2 threads / 2 GPU)	- / 42.9 / 51.0	full frame rate (> 30 fps)
Multi-threaded / parallelised (3 threads / 3 GPU)	- / 64.9 / 60.2	
Multi-threaded / parallelised (4 threads / 4 GPU)	- / 74.1 / 90.1	

throughput of model **8** very similar to model **816**

Distributed Processing
- frame grabber, stereo rectification on "front-end" workstation
- syncing/disparity estimation on "back-end" workstation **GRAPH?**

Recent & Current Work

- integration in processing chain of ARAILIS Image Guidance System
- latest measurement: model **16** on *single* GPU RTX 3090 Ti → 35 fps
- integration of Point Cloud calculation: on *single* RTX 3090 Ti → 32 fps
- reference measurements (on body phantom)

Summary & Conclusions

- improvement/acceleration of dense FullHD disparity algorithm
- model **16**: very good quality, full frame rate on 2 GPUs
  → more accurate 3D reconstruction & registrations
  → improved usability

Acknowledgement

The authors gratefully acknowledge funding for this research by the State of Saxony via Sächsische Aufbaubank (SAB) in the scope of the ARAILIS project (100400076). This measure is co-financed with tax funds on the basis of the budget passed by the Saxon state parliament.

References

[1] Image: Hörstmann & Todoroff www.chirurgie-mallorca.com

[2] YANG, Gengshan, et al. Hierarchical deep stereo matching on high-resolution images. Proc. CVPR, 2019. pp. 5515-5524

[3] NVIDIA, "NVIDIA TensorRT documentation," https://docs.nvidia.com/deeplearning/tensorrt/, online, accessed 2022-06-01