Dense regression for object and grasp point detection and anomaly detection

Danijel Skočaj

University of Ljubljana

UNIVERSITY OF LJUBLJANA | FRI Faculty of Computer and Information Science

ViCOS Visual Cognitive Systems Lab

ROMANDIC Winter School 2026
Kranjska Gora, 12. 2. 2026

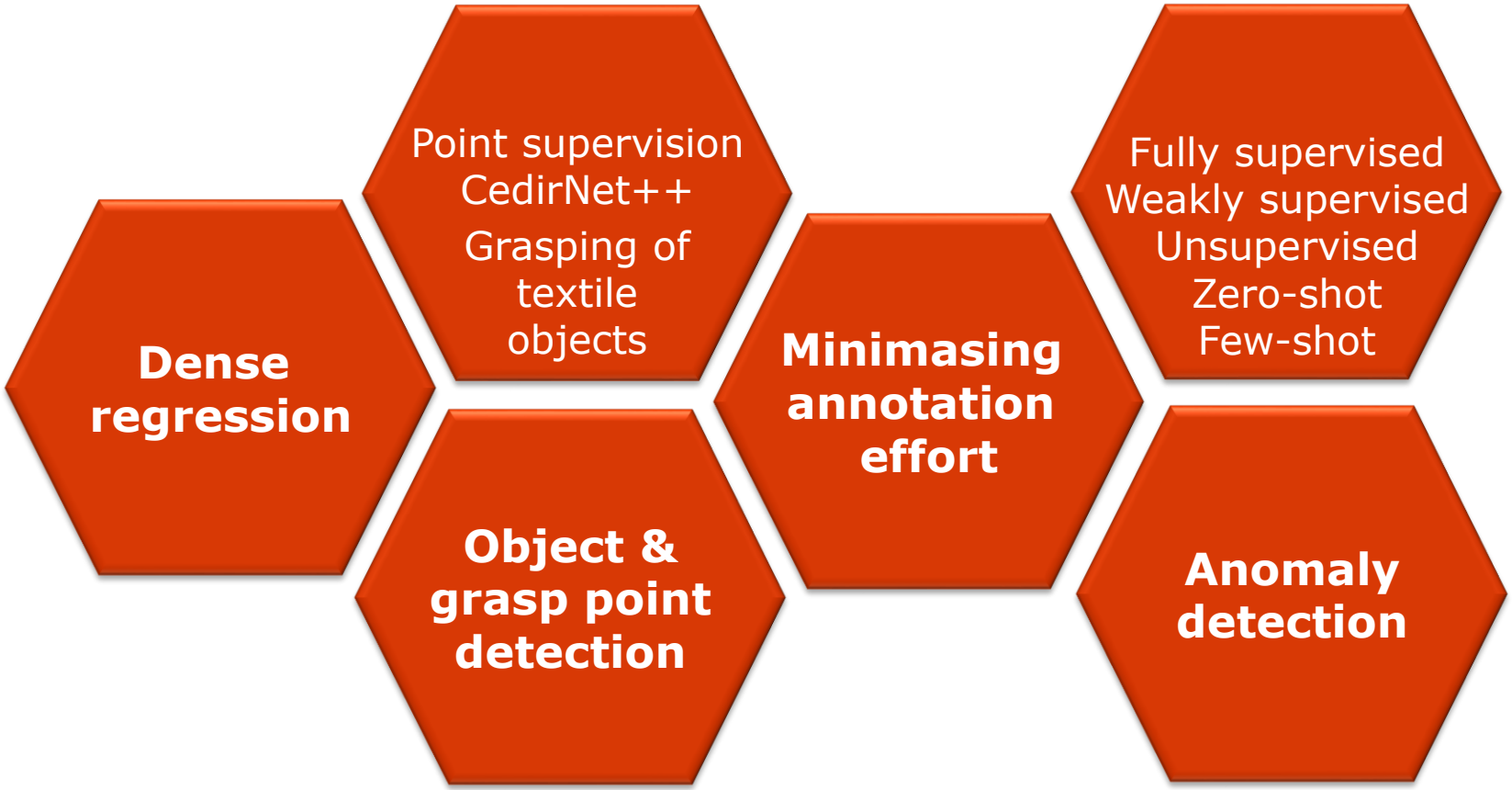
Danijel Skočaj

UNIVERSITY OF LJUBLJANA | FRI Faculty of Computer and Information Science

ViCOS visual cognitive systems lab

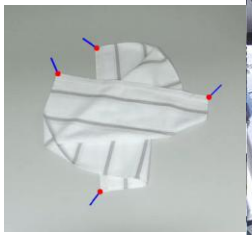
danijel.skocaj@fri.uni-lj.si

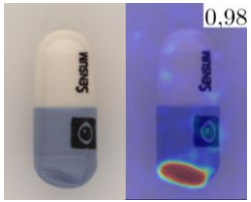
Talk outline



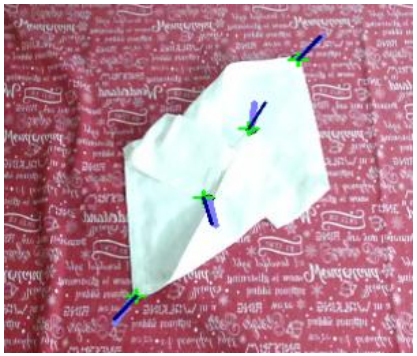
Outline



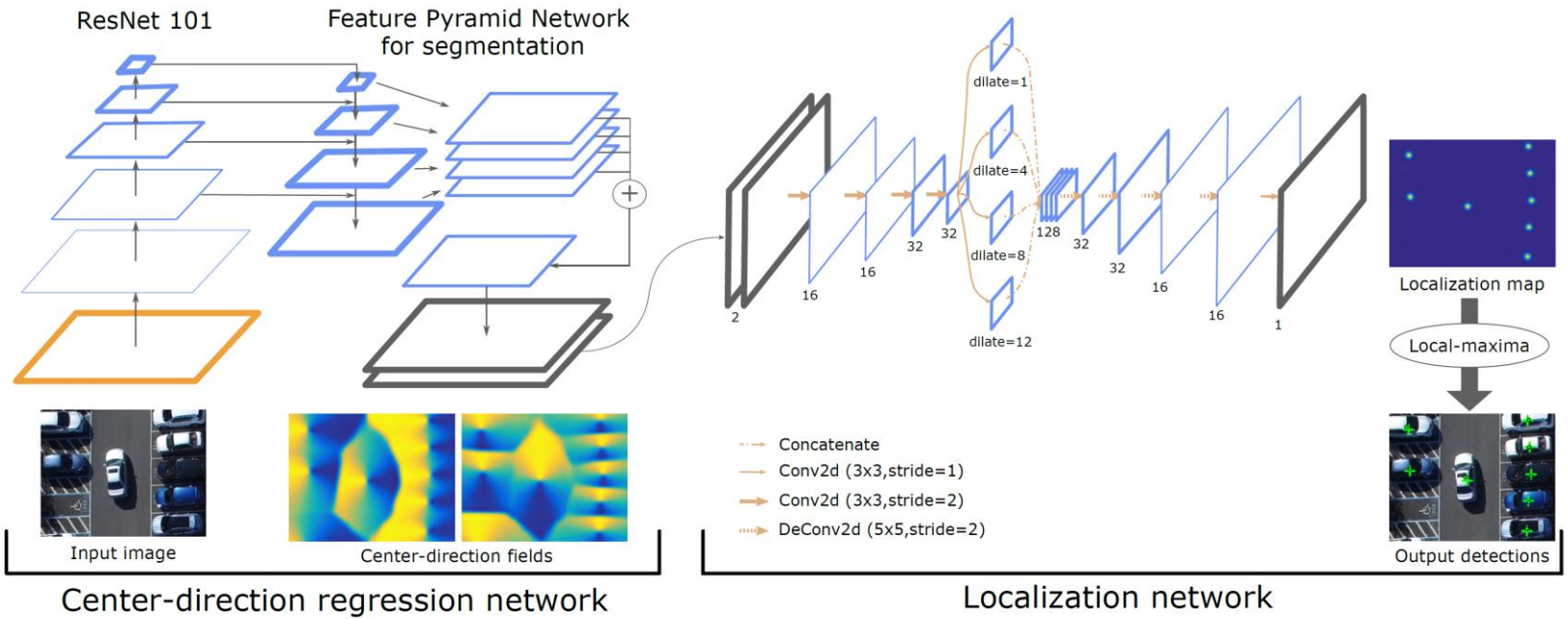



Detection of grasp points



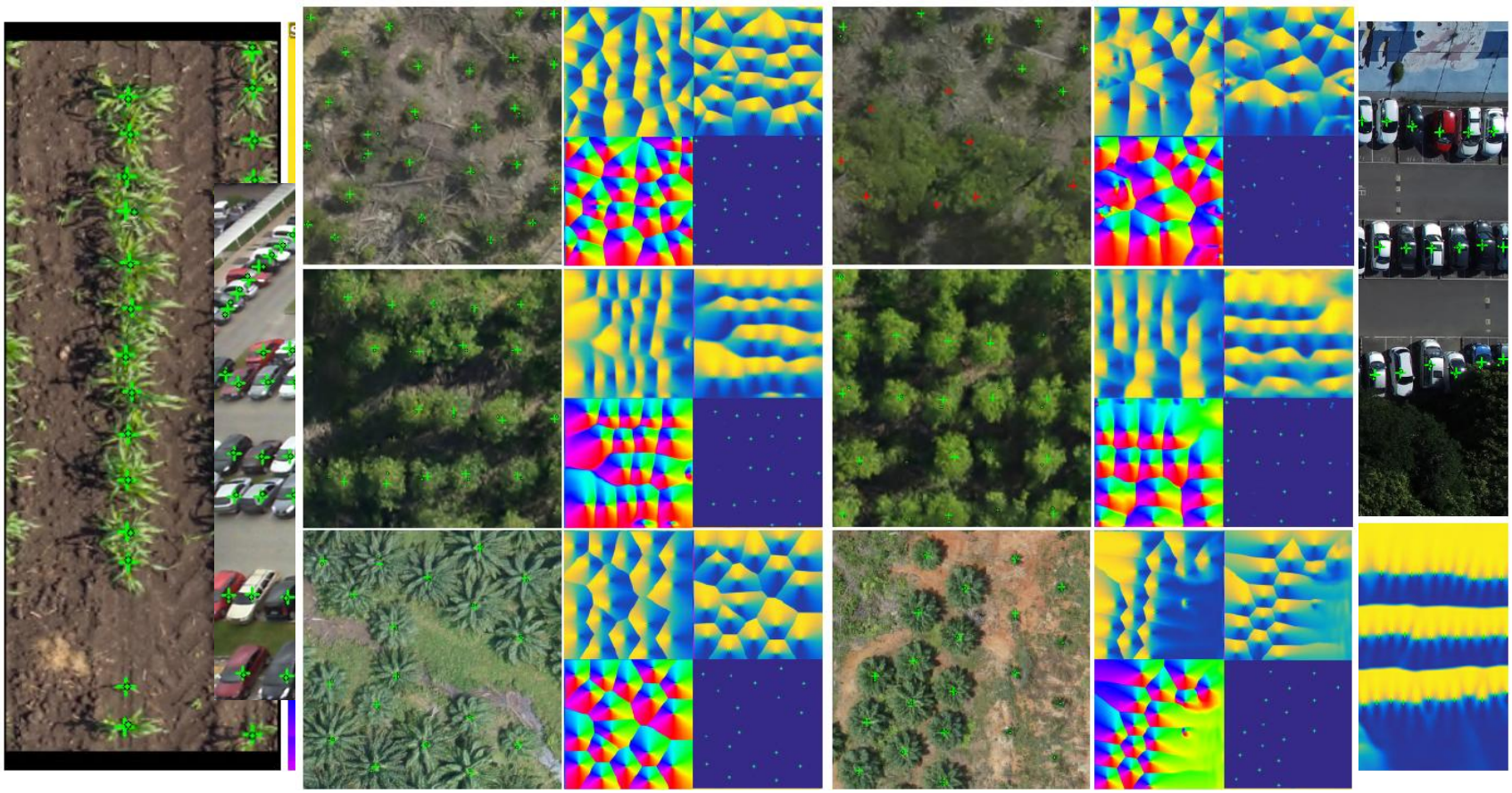
Object counting and localisation

- Point supervision

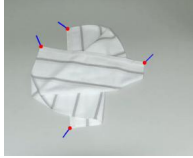


[1] D. Tabernik, J. Muhovič, and D. Skočaj, Dense center-direction regression for object counting and localization with point supervision, Pattern Recognition, 2024.

Object counting and localisation



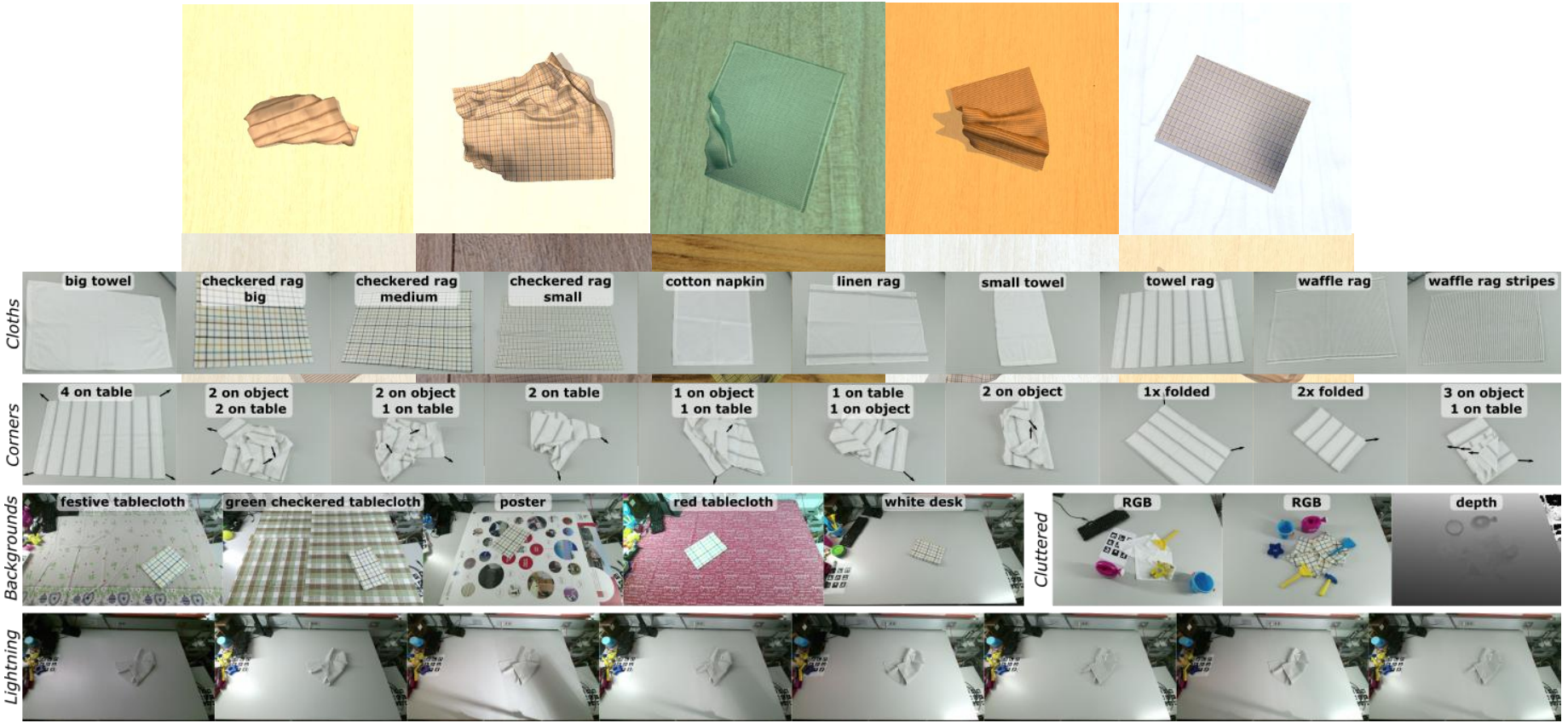
Grasping deformable objects



- Grasping deformable objects (cloth, textile) is challenging due to non-rigid transformations and occlusions.

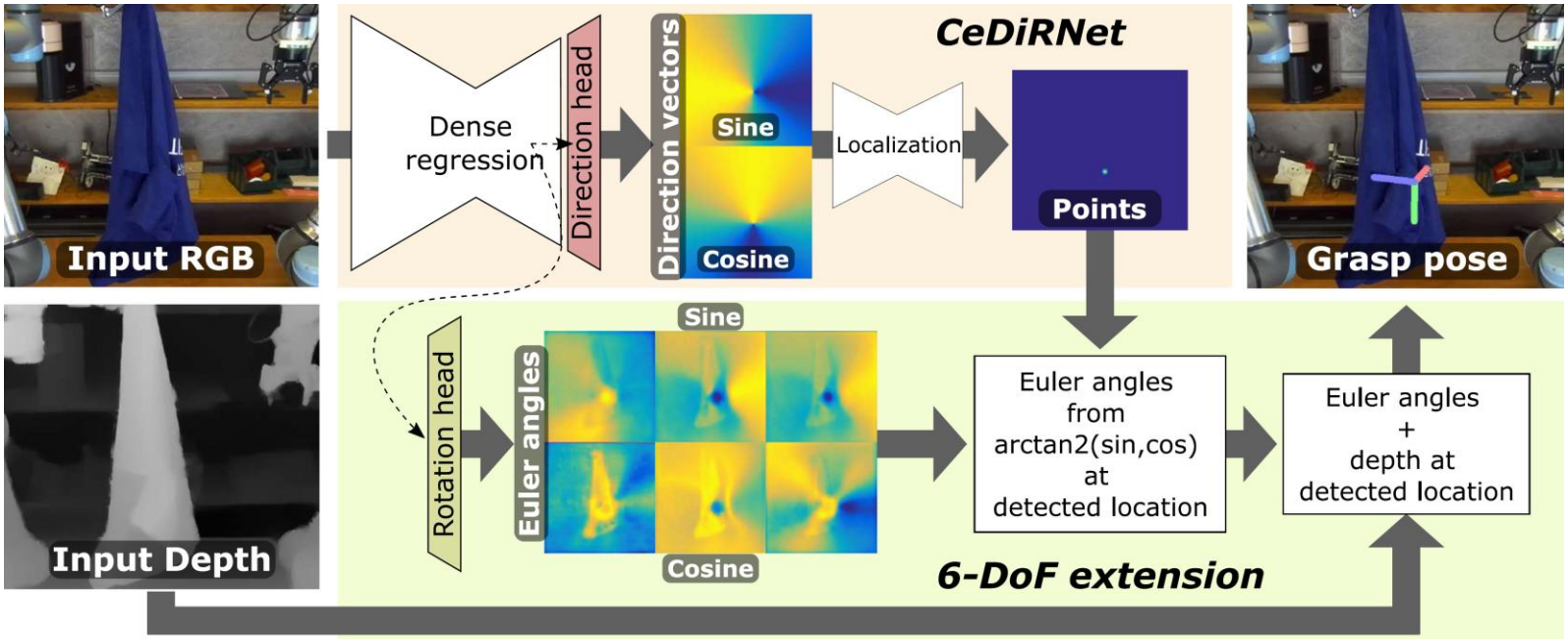
- Lack of standardized benchmarks, limited training and evaluation datasets.
- Goal:
  - Robust method for grasp point detection on cloths
  - Standardized, diverse dataset to benchmark deep-learning methods
  - 3DoF
  - 6DoF

Dataset


Cloths
big towel
checkered rag big
checkered rag medium
checkered rag small
cotton napkin
linen rag
small towel
towel rag
waffle rag
waffle rag stripes
Corners
4 on table
2 on object 2 on table
2 on object 1 on table
2 on table
1 on object 1 on table
1 on table 1 on object
2 on object
1x folded
2x folded
3 on object 1 on table
Backgrounds
festive tablecloth
green checkered tablecloth
poster
red tablecloth
white desk
Cluttered
RGB
RGB
depth
Lightning

CeDiRNet-3DoF/6DoF

- Extension of CeDiRNet to 3DoF/6DoF



Added orientation for 3DoF problem:

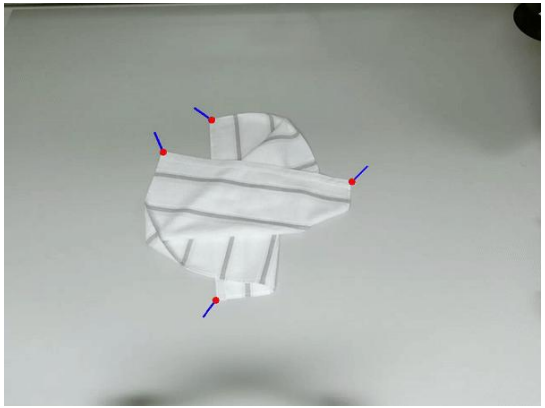
- Encoded with sine/cosine
- During training: only around center (30x30 px)
- During inference: read at center of detection

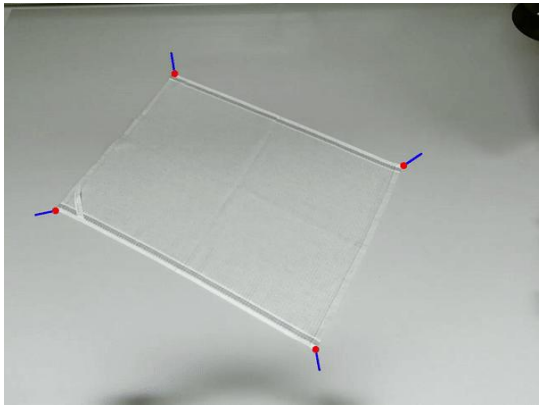
Extension to 6DoF:

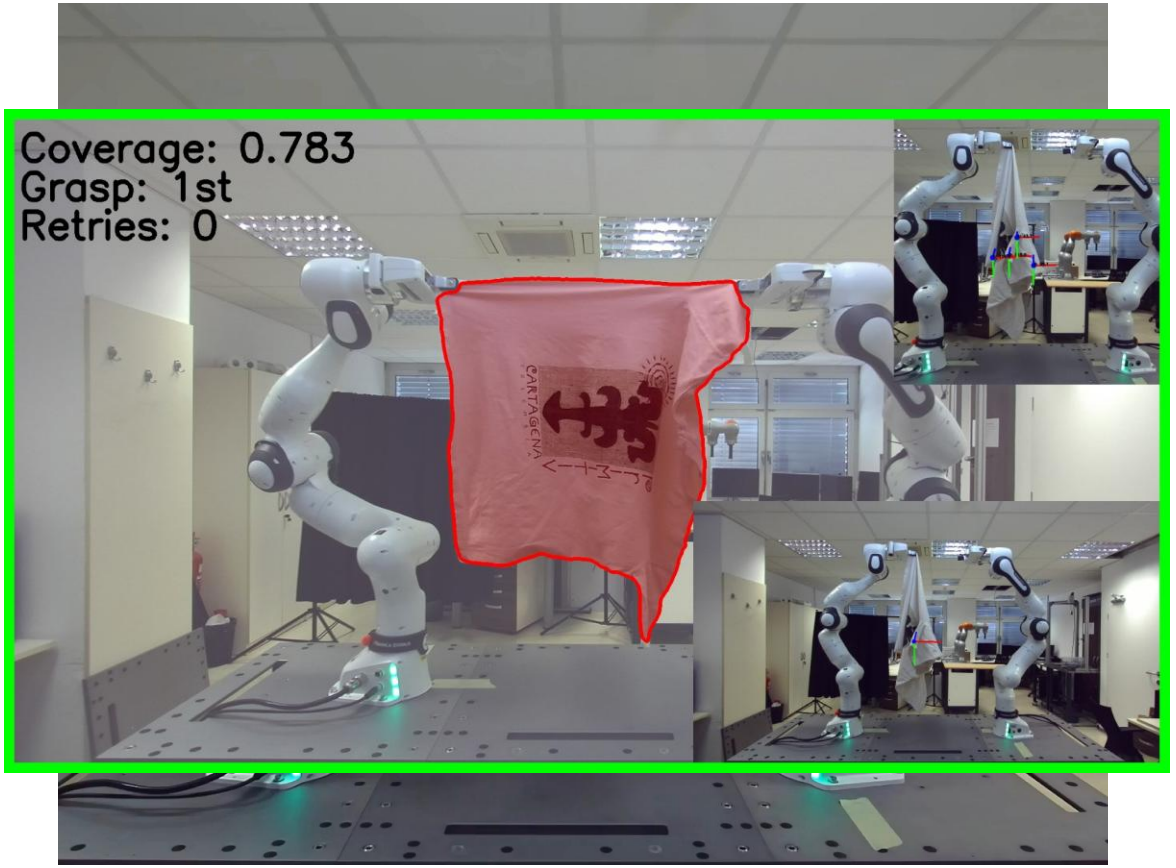
- Prediction of remaining two orientations
- Calculation of thord translation

[2] D. Tabernik, J. Muhovič, M. Urbas, and D. Skočaj, Center direction network for grasping point localization on cloths. IEEE Robotics and automation letters. Oct. 2024.

CeDiRNet-3DoF/6DoF



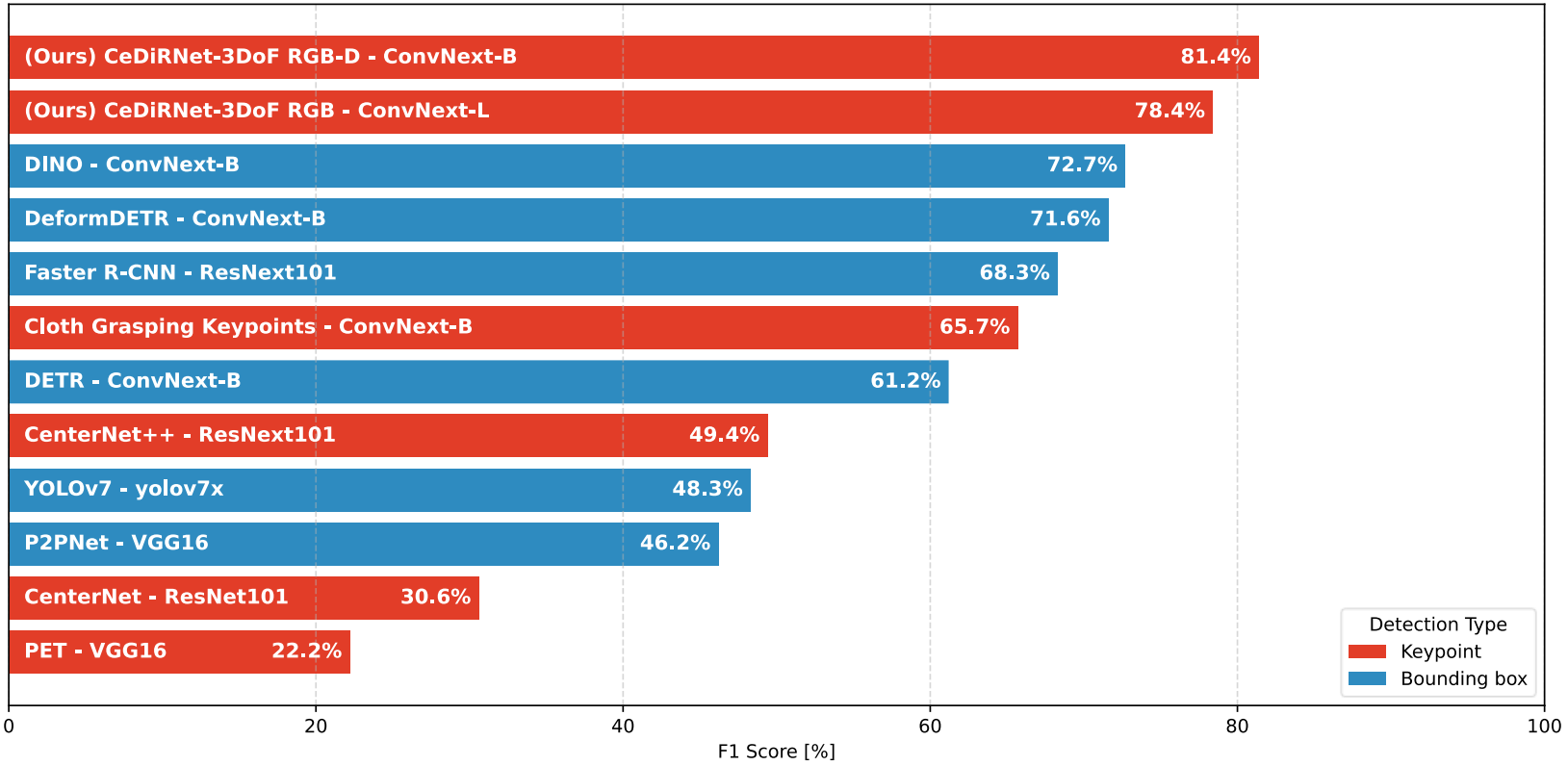




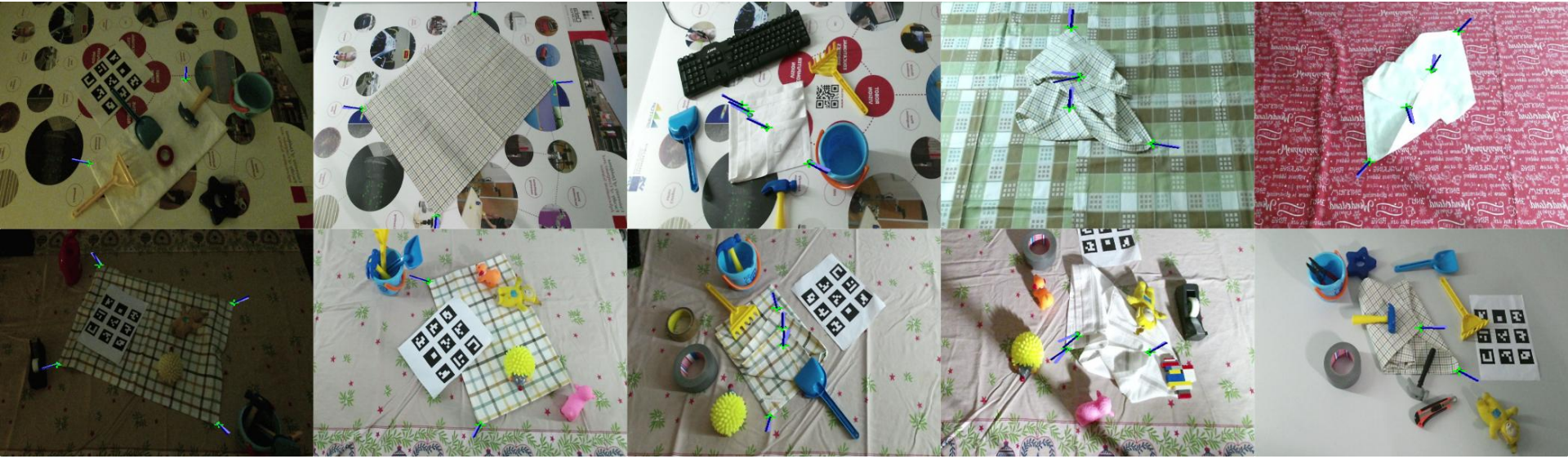
Experimental results– 3DoF

- Training set: 12k synthetic + 5k real images
- Test set: 2880 real images

- *Localisation error*: 1.6 px
- *Orientation error*: 6.4°



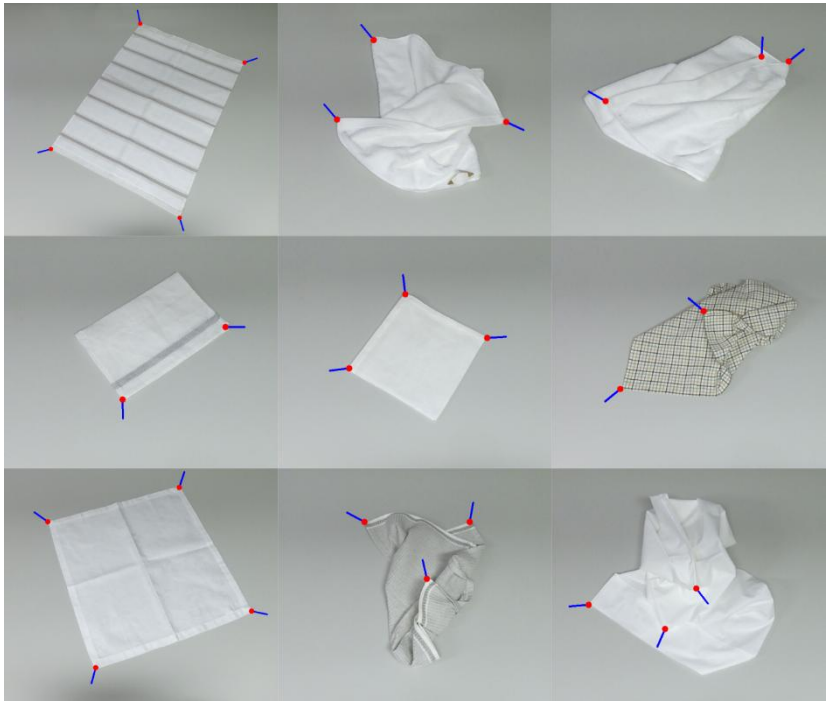
Detection examples



Failure cases:

Cloth Challenges @ ICRA

Cloth Competition - ICRA 2023



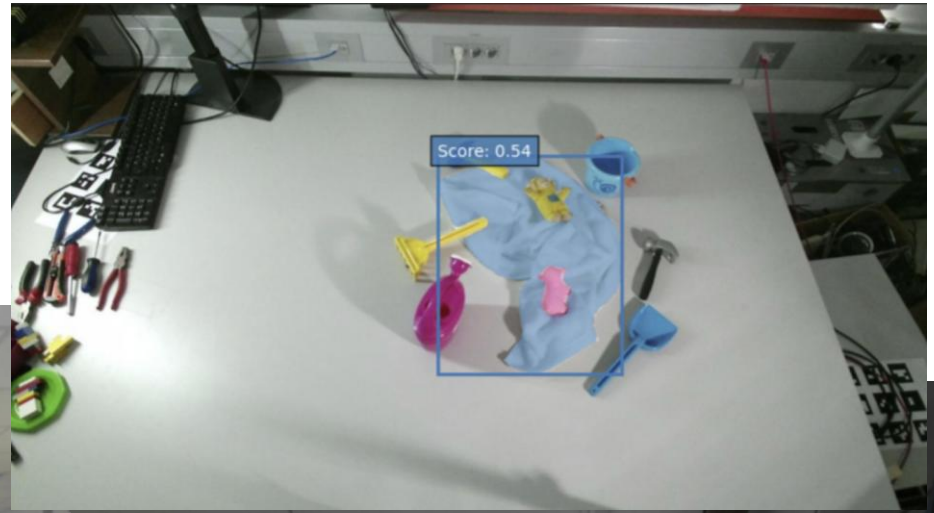
Cloth Unfold Competition - ICRA 2024

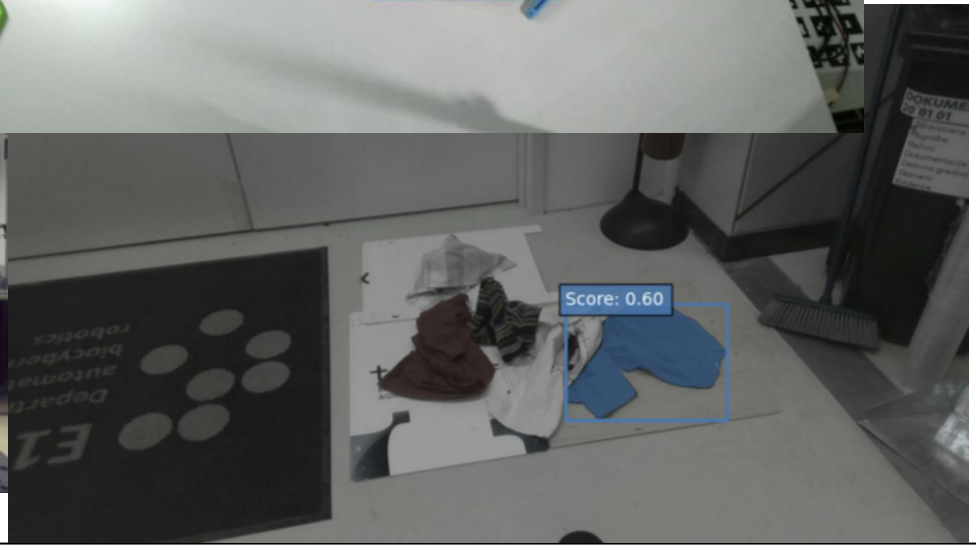
1st place – *grasp point detection*

2nd place – *grasping and stretching*

Towards general visual models

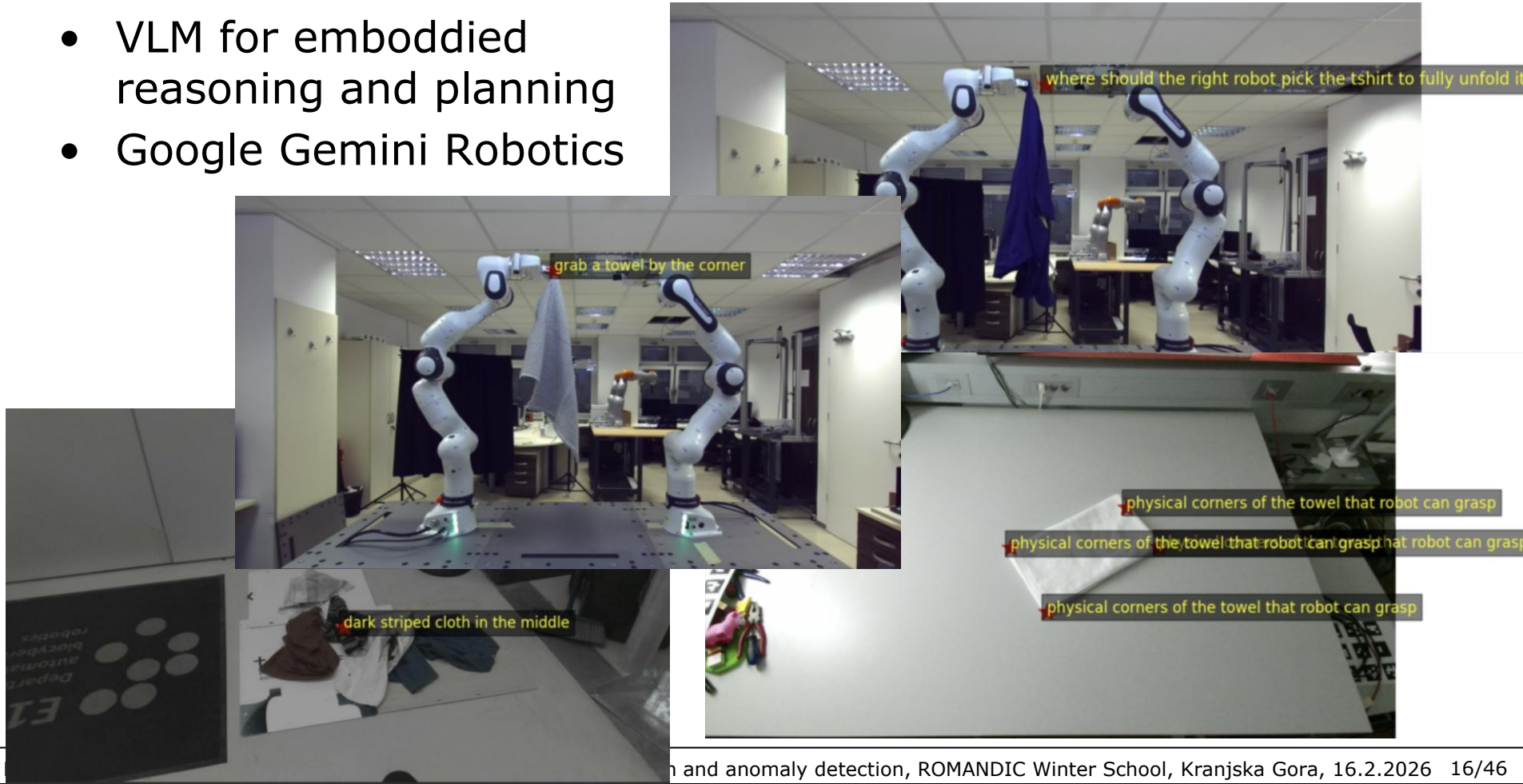
- FVM for general visual understanding
- SAM3
  - Semantic segmentation:





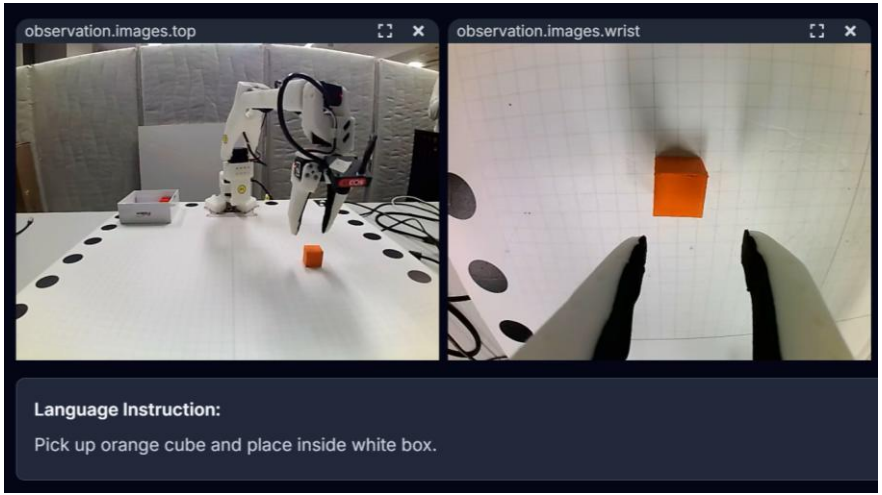
Towards general vision-language models

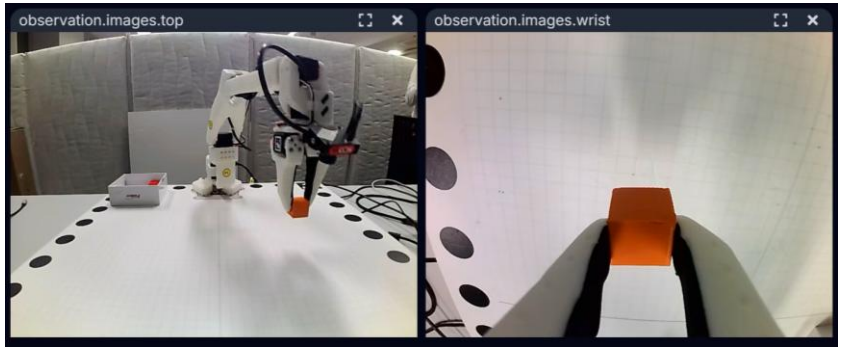
- VLM for emboddied reasoning and planning
- Google Gemini Robotics

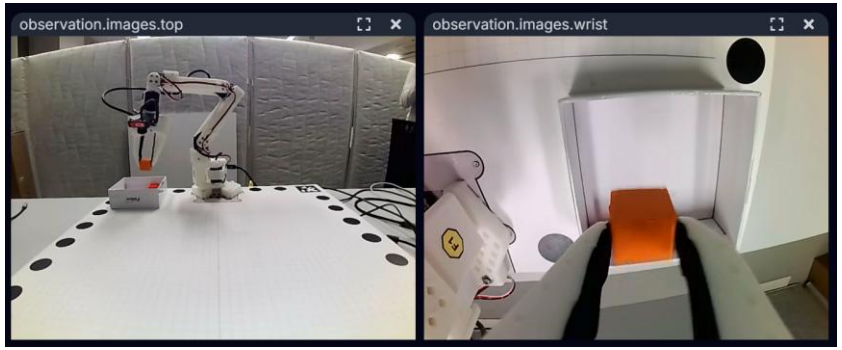


Towards general vision-language-action models

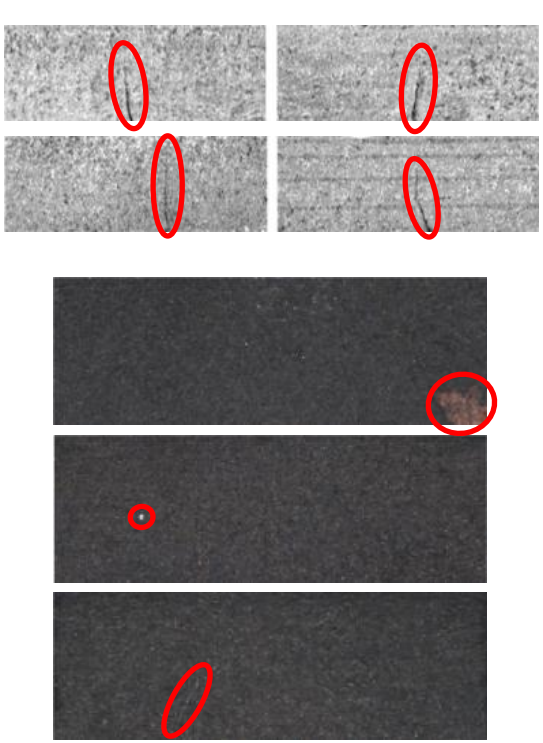
- VLA for direct vision-robot manipuation
- SmolVLA:



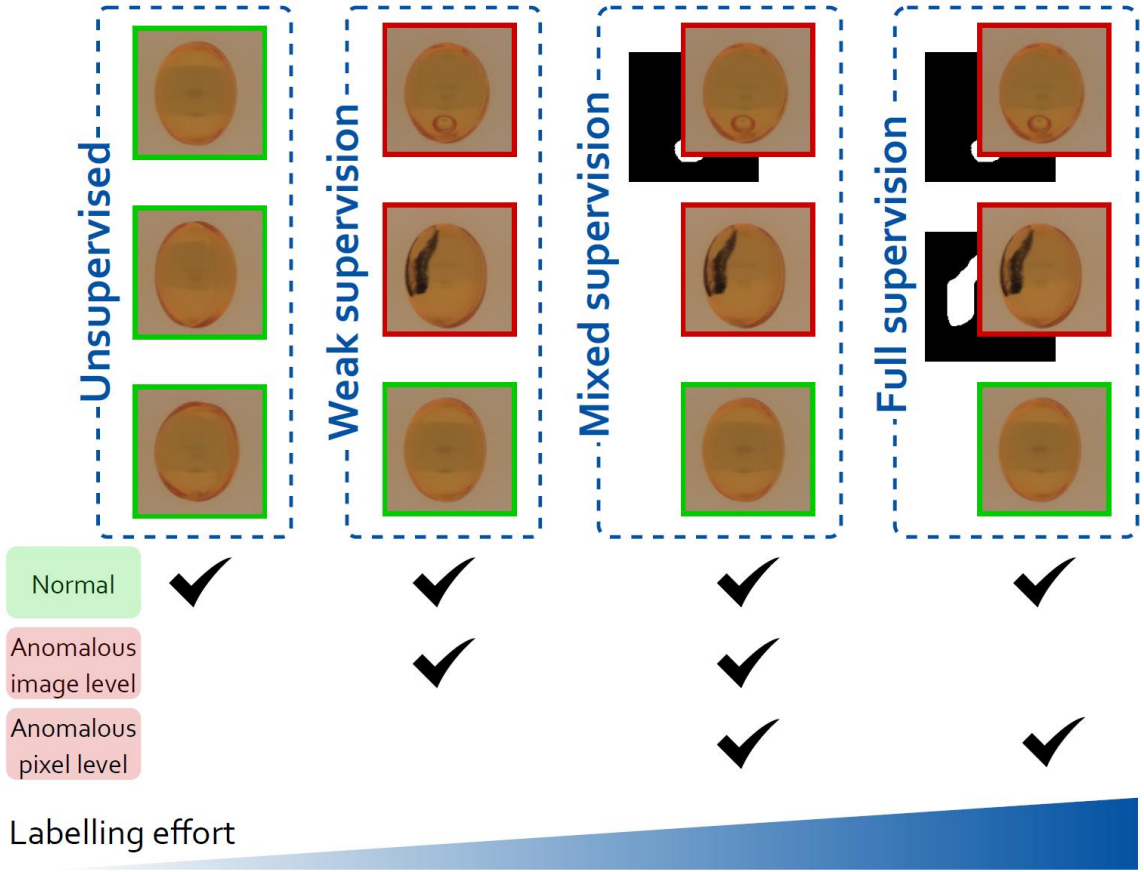




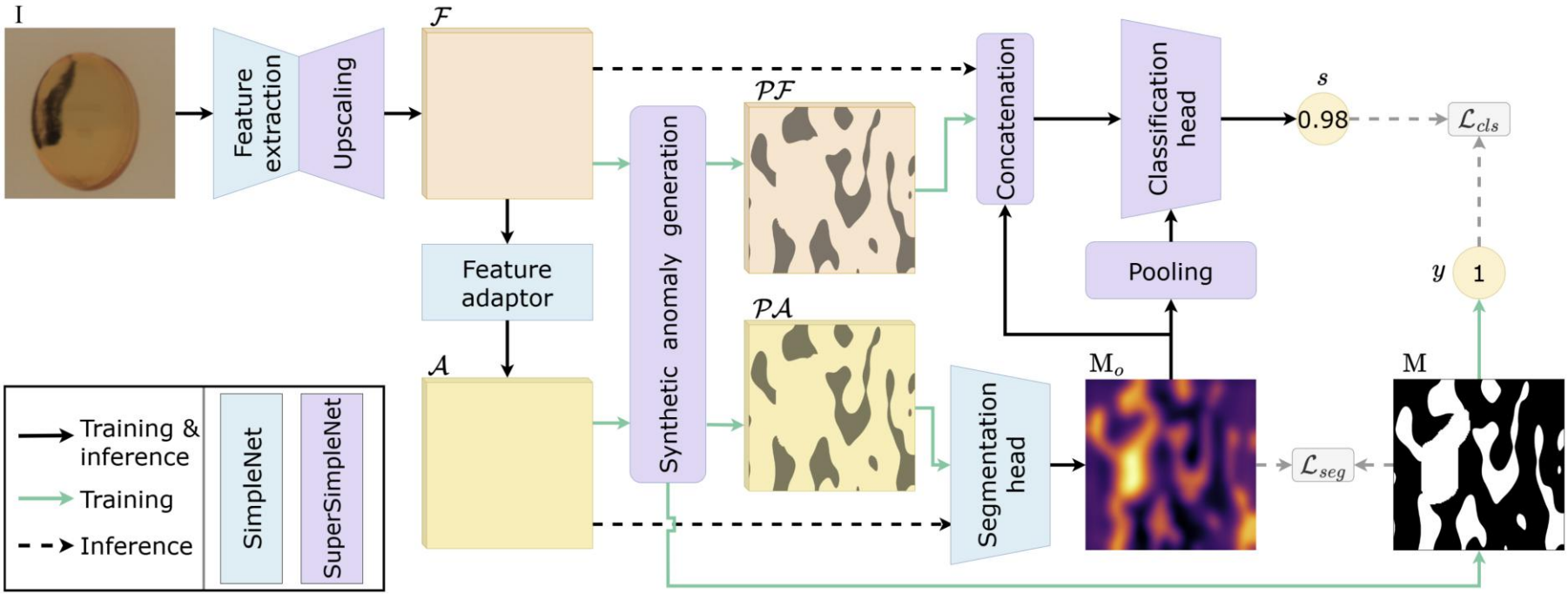
Surface anomaly detection problem



Learning regimes



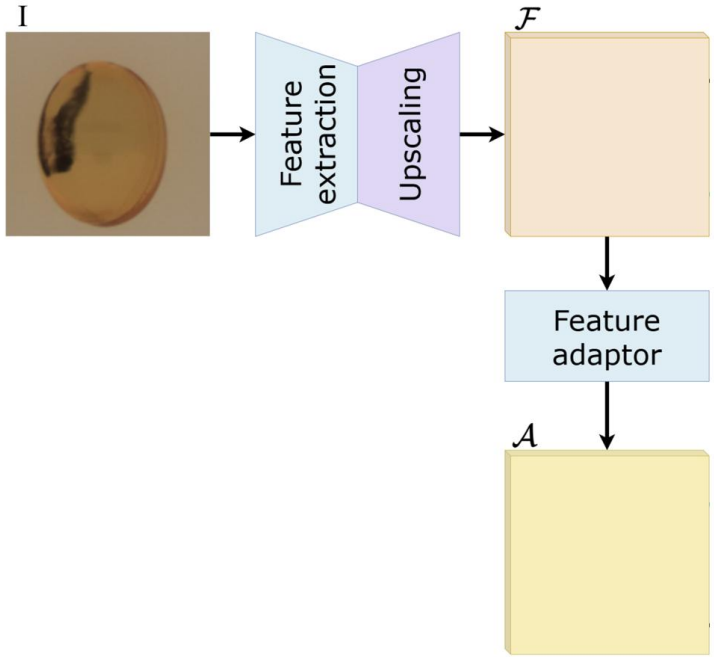
SuperSimpleNet



[3] Z. Liu, Y. Zhou, Y. Xu, Z. Wang, SimpleNet: A Simple Network for Image Anomaly Detection and Localization, CVPR 2023.
[4] B. Rolih, M. Fučka D. Skočaj, No Label Left Behind: A Unified Surface Defect Detection Model for all Supervision Regimes, JIM, 2024.

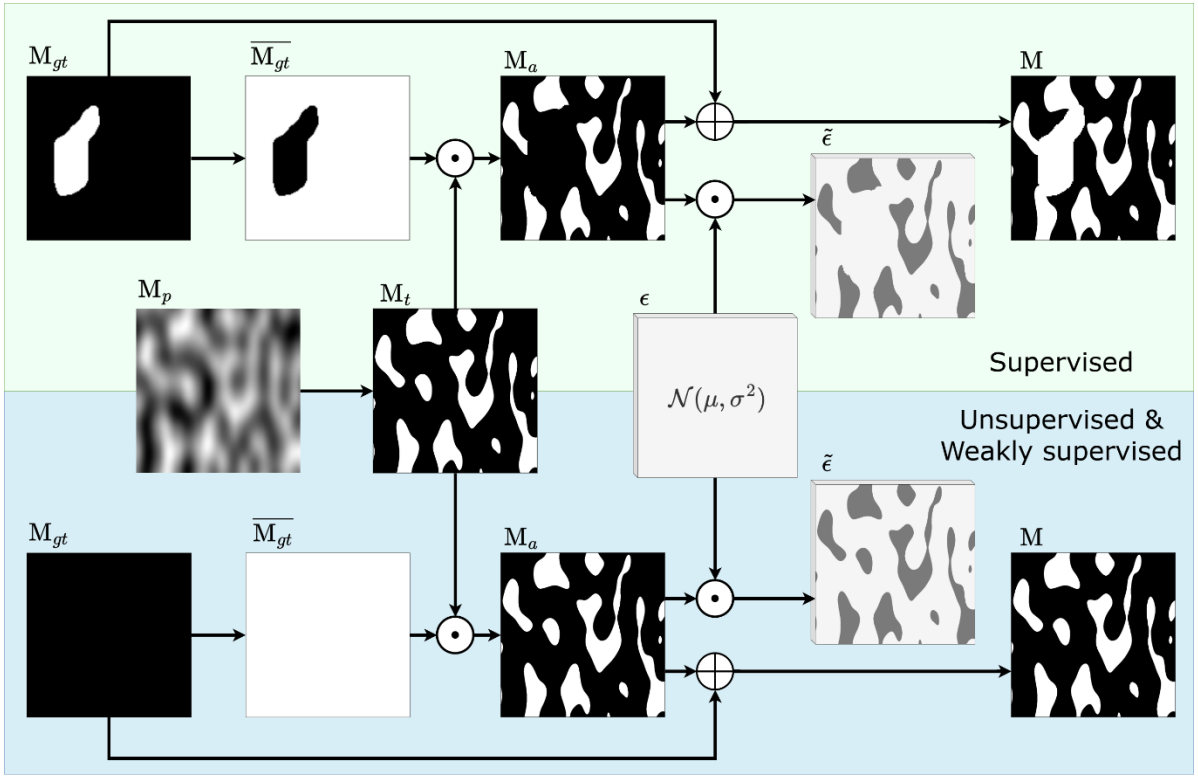
Features

- Pretrained network
- Upscaling
- Capturing neighbouring information
- Feature adaptation



Anomaly generation

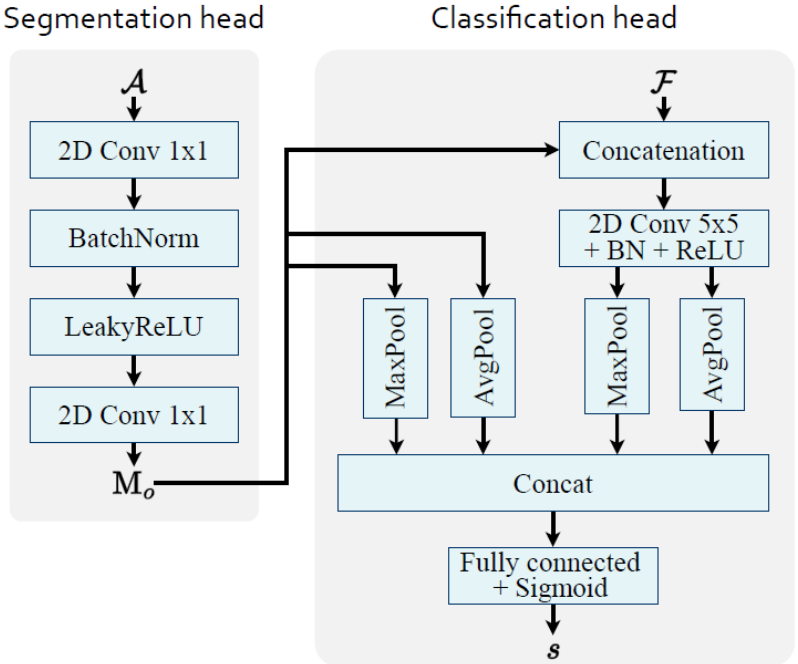
- True & synthetic anomalies
- Perlin noise
- Gaussian noise
- Label
  - Image level
  - Pixel level
- Weakly sup.
- Unsupervised



Segmentation-detection module

- Segmentation head
- Classification head
  - Capturing global context
  - Detection of small defects
  - Enables mixed supervision

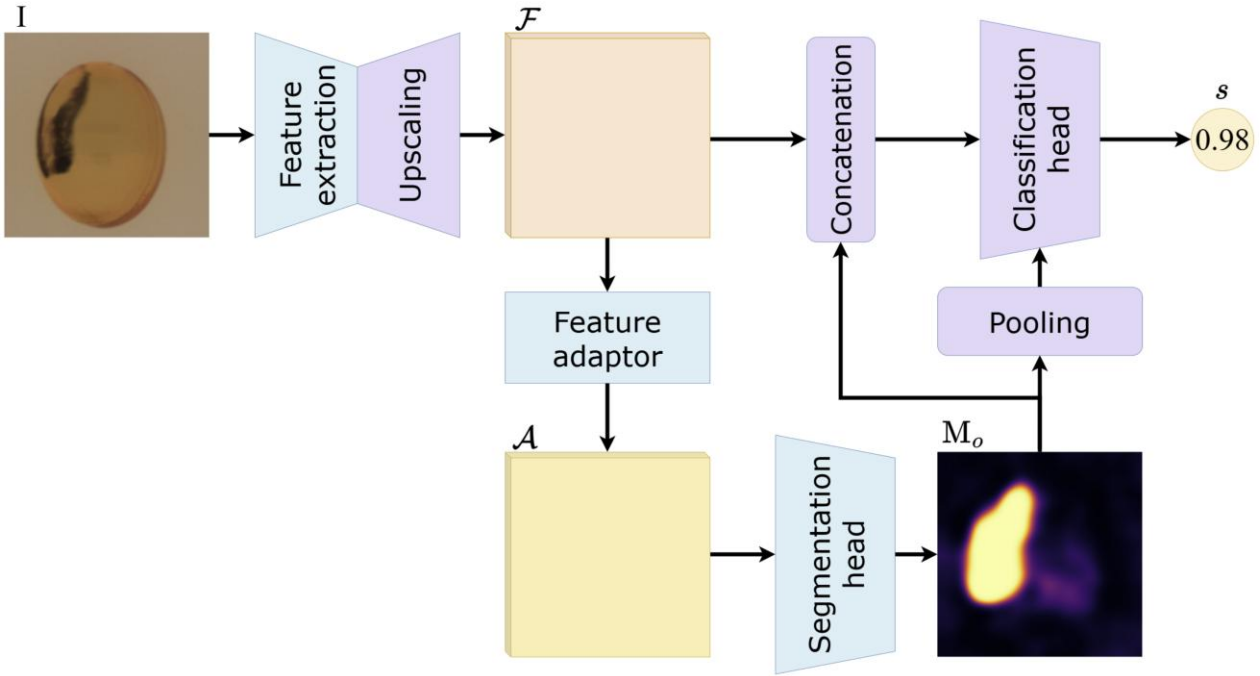
$$\mathcal{L}_{seg} = \mathcal{L}_{1t} + \mathcal{L}_{foc} \qquad \mathcal{L}_{cls} = \mathcal{L}_{foc}$$



$$\mathcal{L} = \gamma \cdot \mathcal{L}_{seg} + \mathcal{L}_{cls}$$

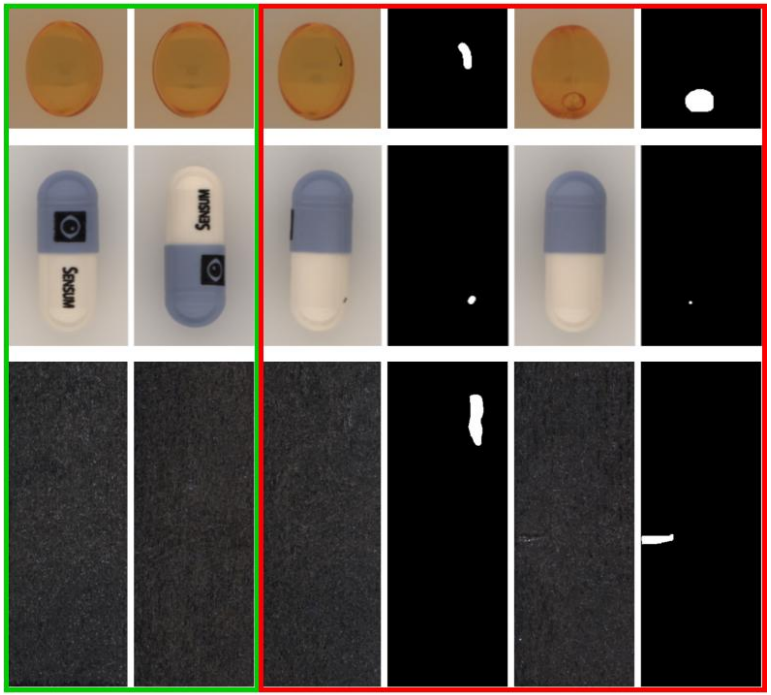
$$\gamma = \begin{cases} 1; \text{ if image is normal;} \\ 1; \text{ if image is anomalous a[illegible]} \\ 0; \text{ if image is anomalous and weakly labelled;} \end{cases}$$

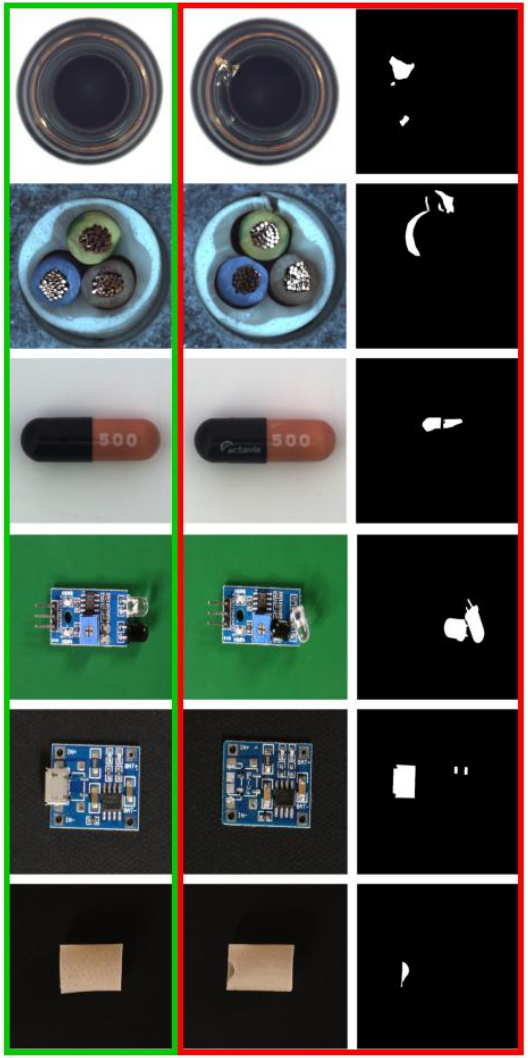
Inference


I
Feature extraction
Upscaling
$\mathcal{F}$
Concatenation
Classification head
s
0.98
Feature adaptor
Pooling
$\mathcal{A}$
M_o
Segmentation head

Datasets and performance metrics

- Supervised
  - SensumSODF
  - KolektorSDD2
- Unsupervised
  - MVTec AD
  - VisA
- Detection
  - AUC
  - AP_{det}
- Localisation
  - AUPRO
  - AP_{loc}





Experimental results

- Capabilities of SOTA methods:

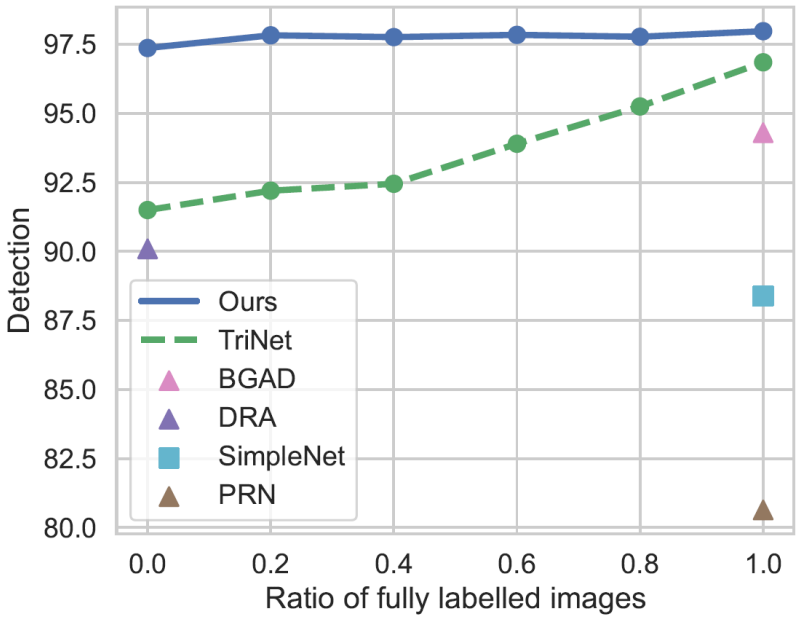
	US	WS	MS	FS	**Ours**	SDNet	TNet	MMNet	DRA	EAD	BGAD	DSR	SN	FF	PC	DRÆM	PRN
normal only	✓				✓					✓	✓	✓	✓	✓	✓	✓	
ano. image level		✓	✓		✓	✓	✓	✓	✓								
ano. pixel level			✓	✓	✓	✓	✓	✓			✓	✓					✓
Speed ($< 10ms$)					✓	✓	✓	✓	✓	✓							

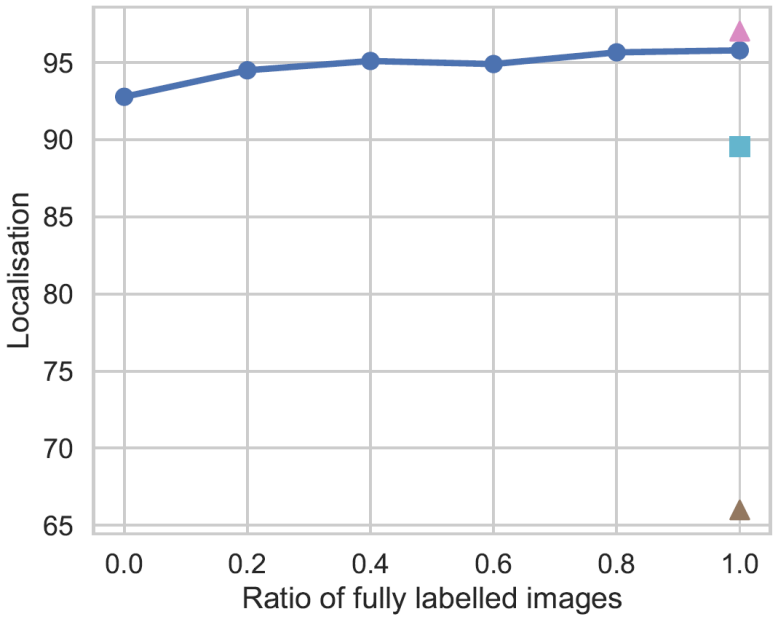
- Unsupervised learning results

	MVTec AD		VisA	
	Det.	Loc.	Det.	Loc.
AST (Rudolph et al., 2023)	98.9	81.2	94.9	81.5
DSR (Zavrtanik et al., 2022)	98.1	90.8	91.8	68.1
EfficientAD (Batzner et al., 2024)	99.1	93.5	98.1	94.0
FastFlow (Yu et al., 2021)	96.9	92.5	93.9	86.8
PatchCore (Roth et al., 2022)	98.7	92.7	94.3	79.7
DRÆM (Zavrtanik et al., 2021a)	98.0	92.8	91.5	78.0
SimpleNet (Liu et al., 2023)	97.6 (± 0.40)	90.5 (± 0.75)	91.2 (± 1.08)	88.0 (± 0.87)
Ours	98.3 (± 0.14)	91.2 (± 0.14)	93.6 (± 0.77)	87.4 (± 0.98)

Experimental results – mixed supervision

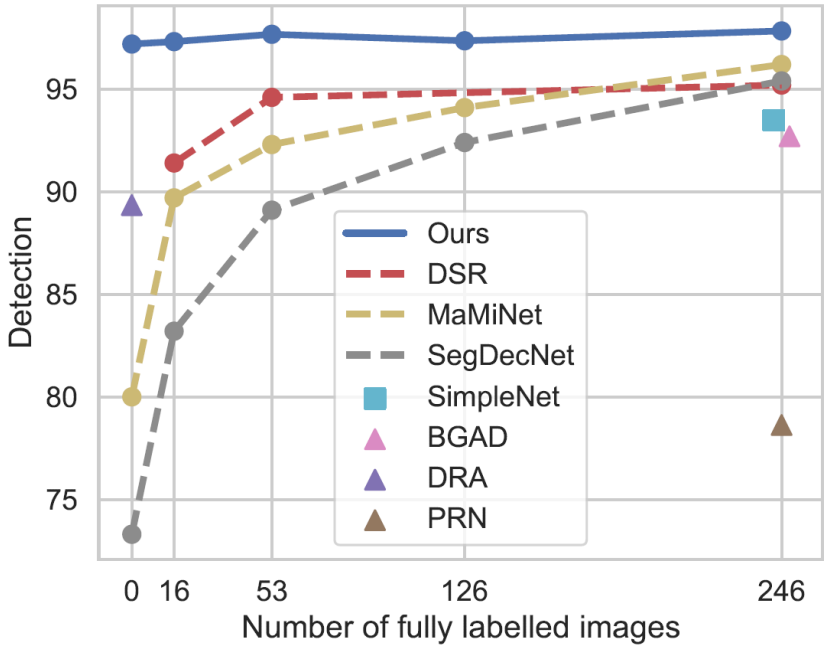
- Weakly supervised -> Mixed supervision -> Fully supervised
- SensumSODF:

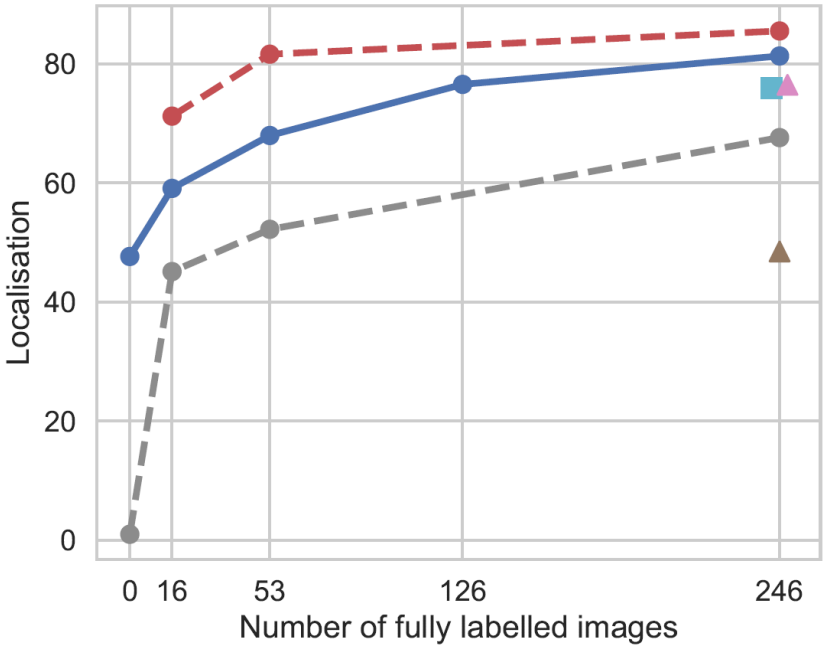




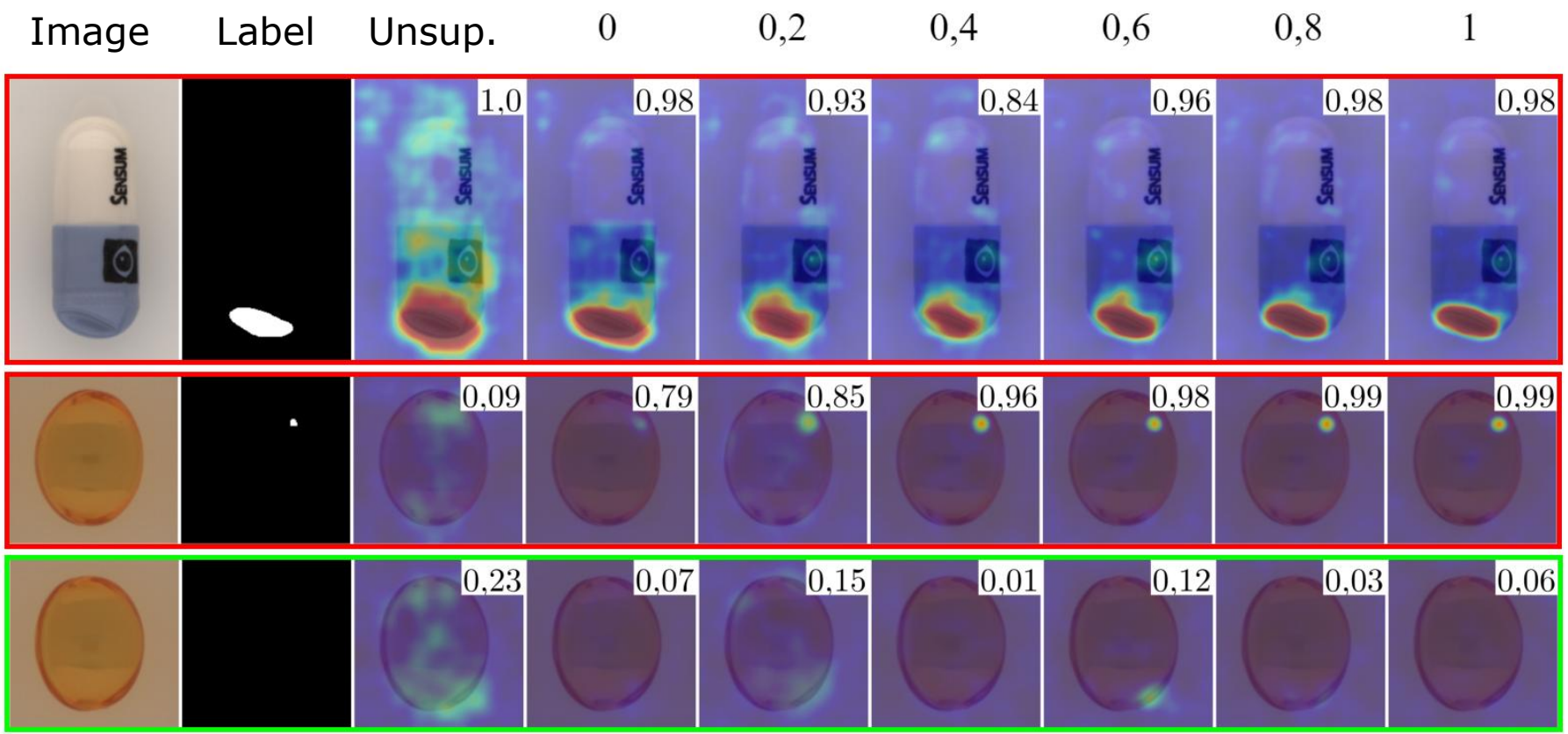
Experimental results – mixed supervision

- Weakly supervised -> Mixed supervision -> Fully supervised
- KolektorSDD2

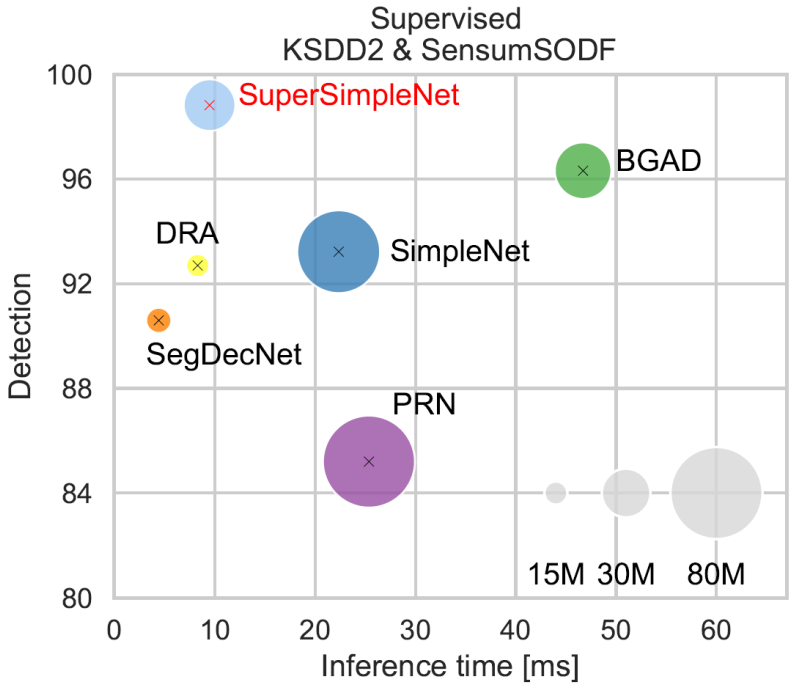


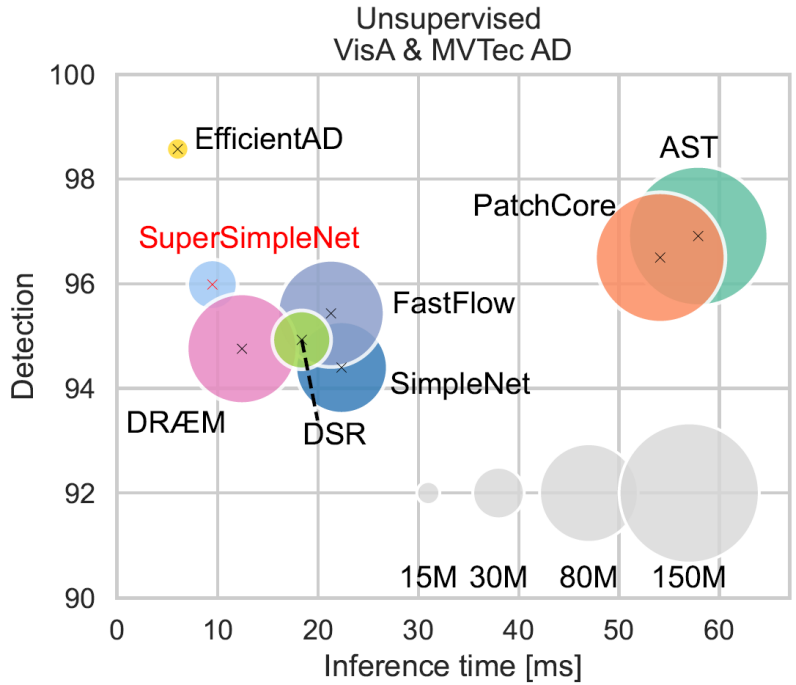


Qualitative experimental results

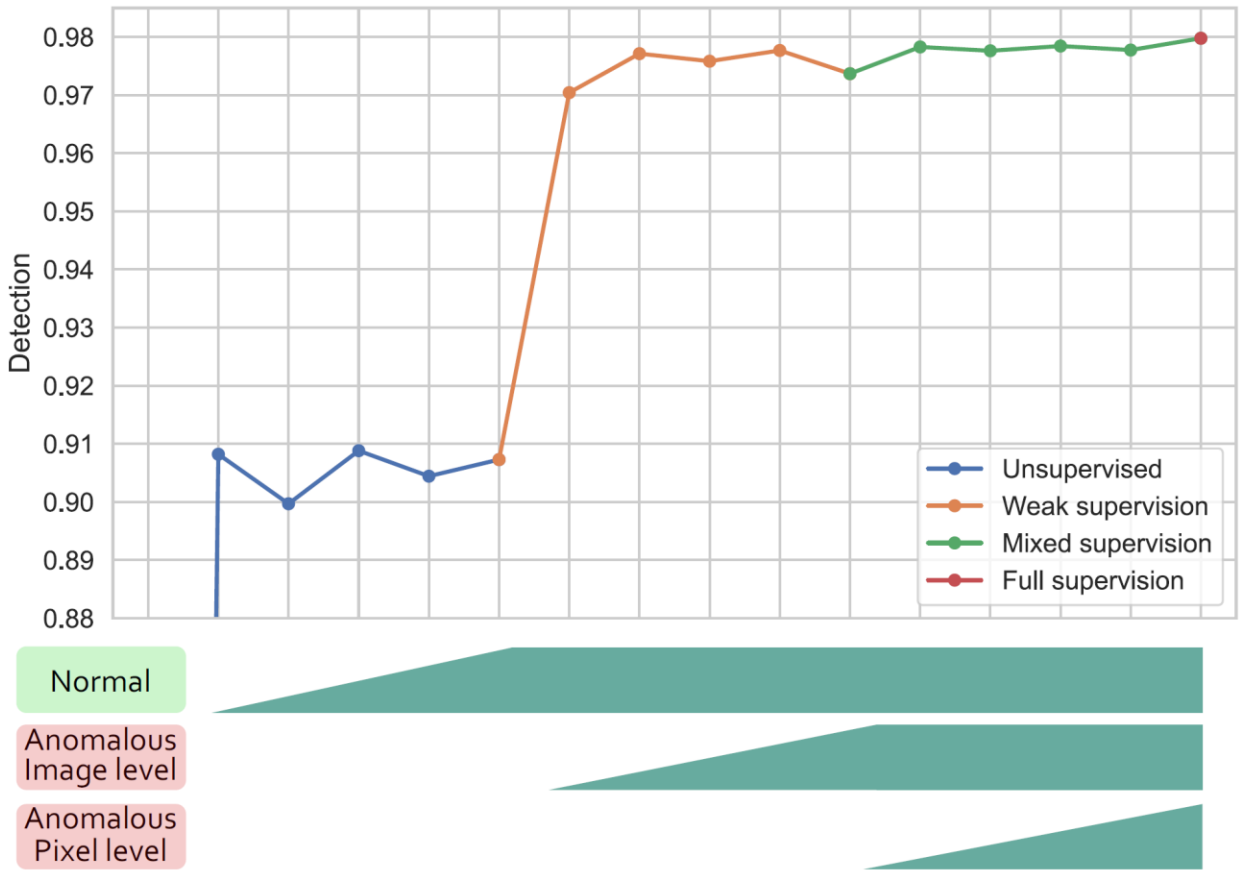


Efficiency

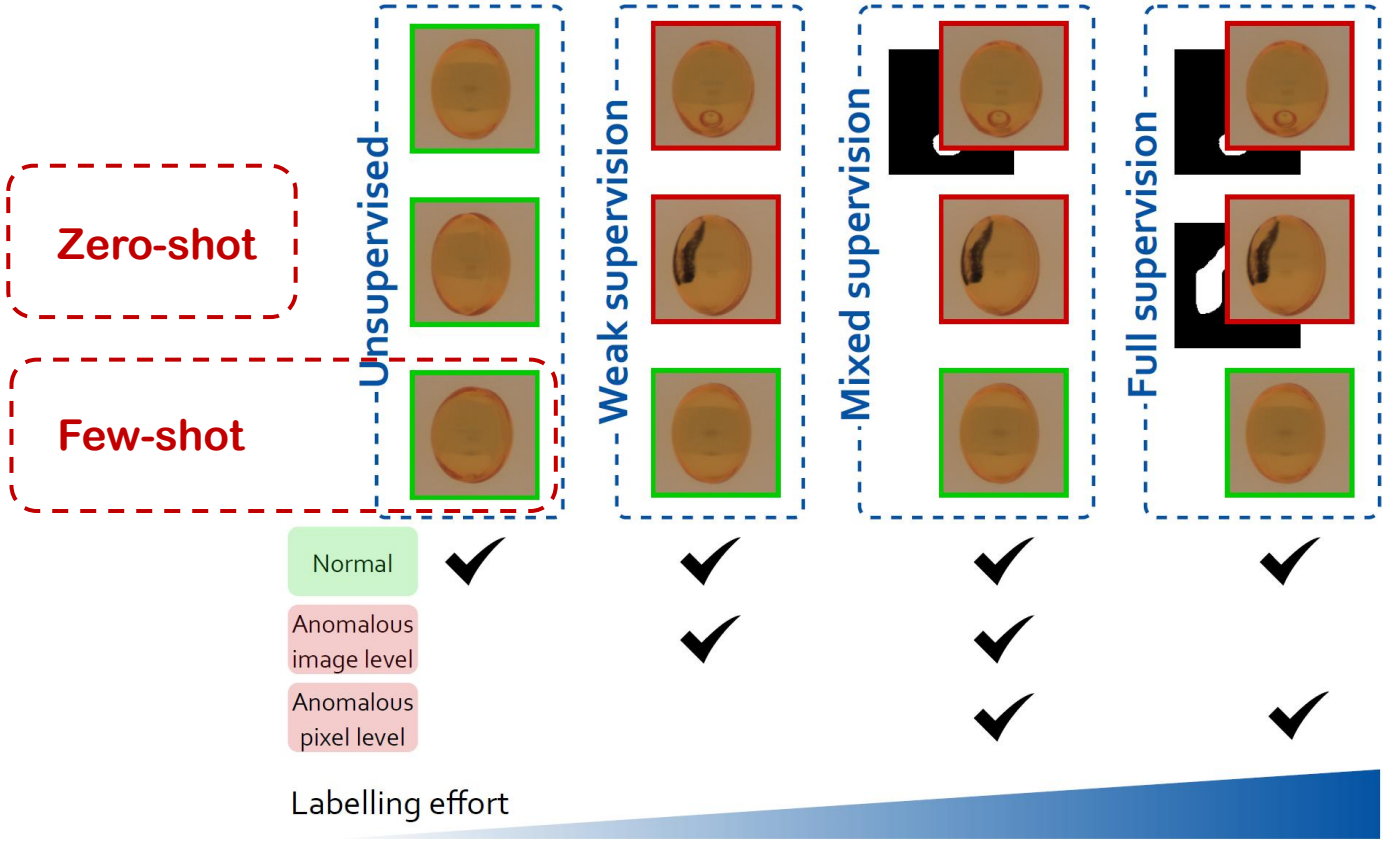

Supervised
KSDD2 & SensumSODF
Detection
Inference time [ms]
SuperSimpleNet
BGAD
DRA
SimpleNet
SegDecNet
PRN
15M
30M
80M


Unsupervised
VisA & MVTec AD
Detection
Inference time [ms]
EfficientAD
AST
PatchCore
SuperSimpleNet
FastFlow
SimpleNet
DRÆM
DSR
15M
30M
80M
150M

Spectrum of learning regimes

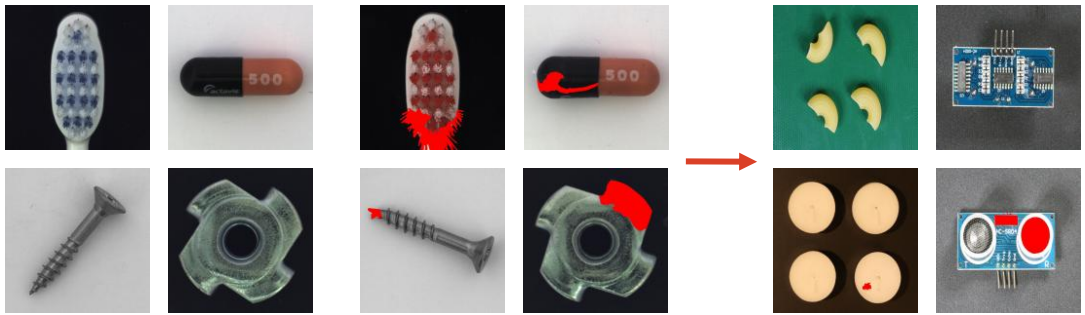


Learning regimes



Zero-shot anomaly detection

- The problem is in fact **General Defect Detection / Segmentation**
- Training on Testset A → Testing on Testset B
- Leveraging knowladge encoded in LLMs (& VLMs)



Train Test

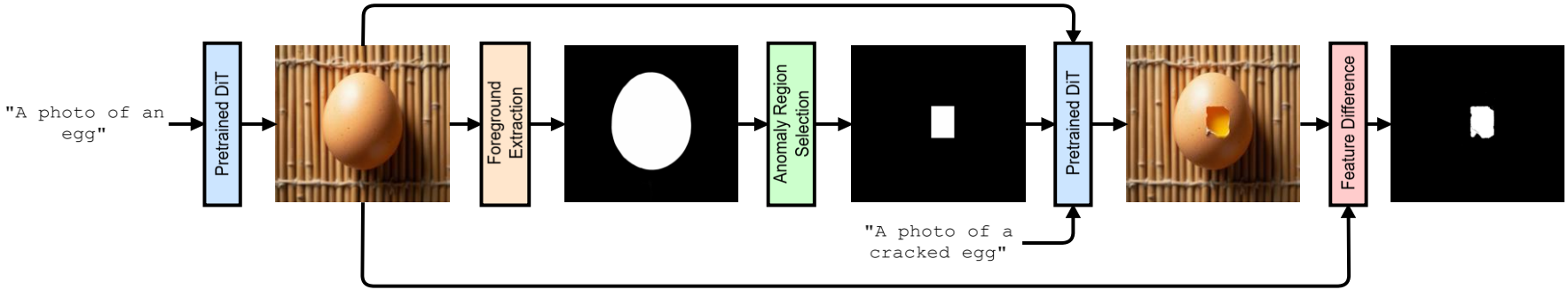
- Current SOTA
  - Use CLIP as a backbone
  - Use trainable text embeddings to encode generic notions of normality and abnormality

Research questions and proposed solution

- Why VLMs? Is text really needed? Why not VFMs?

- Problems:
  - Existing datasets have low data diversity
    → Most backbones do not generalize
  - Current methods finetune only later layers
    → Features are suboptimal for zero-shot anomaly detection
- Proposed solution:
  - **Synthetic dataset generation** scheme
    → high data diversity
  - **AnomalyVFM**: Rather than complex additions add simple parameter efficient drop-ins across all layers
    → features adapted for anomaly detection

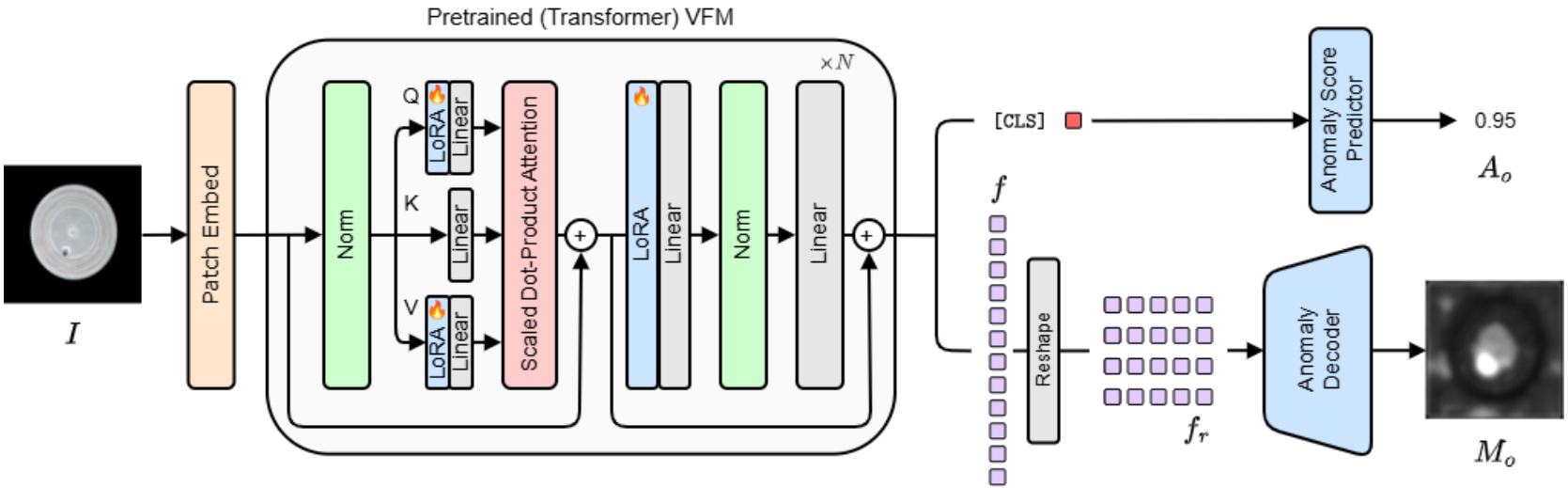
Synthetic Dataset Generation



- Features extracted from images with a pretrained model (DINOv2)
- Anomaly mask approximated as the feature difference
- If no pixel is positive, we discard the sample

Examples of synthetic images

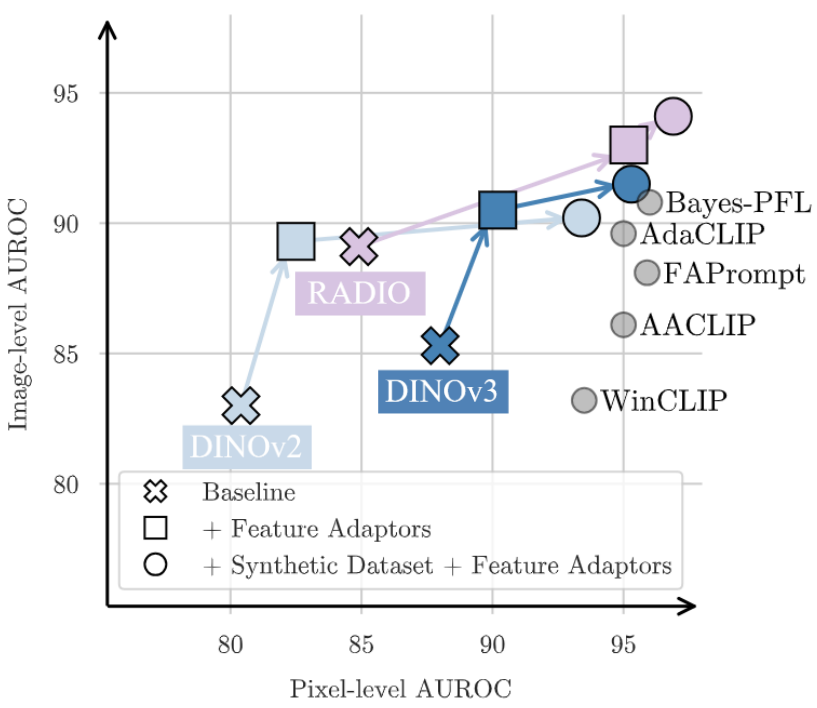
AnomalyVFM



- Add LoRA to each layer of VFM
- Add Decoder and Predictor
- Optimize with Confidence weighted loss (akin to DUST3R)
- **Strong VFM for anomaly detection**

Experimental results – Zero-shot AD

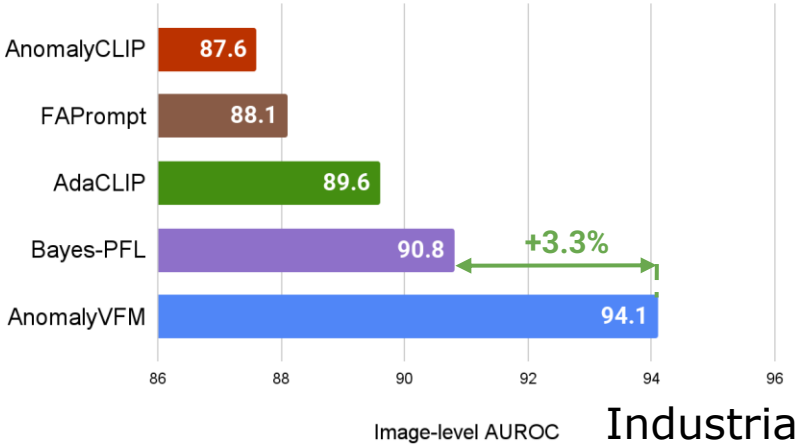
- 9 industrial and 9 medical datasets
- Detection and localisation
- Performance metrics:
  - AUROC
  - F1-max

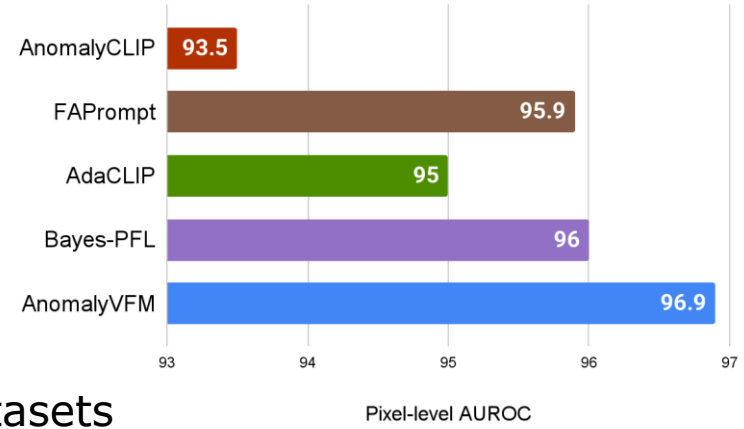


Zero-shot AD – quantitative results

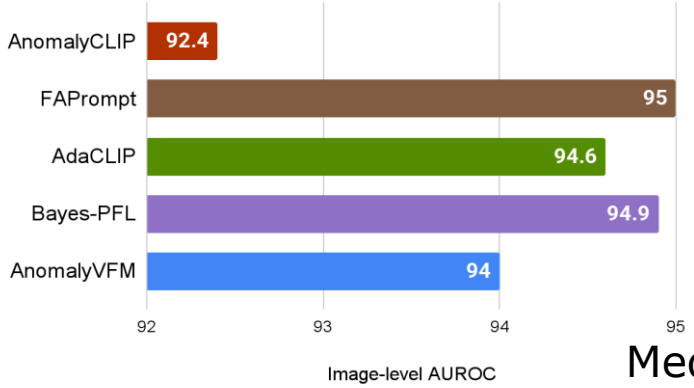
Metric	Dataset	SAA [8] ToC'25	WinCLIP [25] CVPR'23	AnomalyCLIP [80] ICLR'24	AdaCLIP [9] ECCV'24	AACLIP [46] CVPR'25	Bayes-PFL [52] CVPR'25	FAPrompt [84] ICCV'25	*AnomalyVFM*
Image-level (AUROC, max-F1)	MVTec AD	(63.5, 87.4)	(91.8, 92.9)	(91.6, 92.7)	(89.2, 90.6)	(90.5, 90.4)	(92.3, 93.1)	(91.1, 92.2)	(94.9, 94.1)
	VisA	(67.1, 75.9)	(78.1, 80.7)	(82.0, 80.4)	(85.8, 83.1)	(84.6, 78.8)	(87.0, 84.1)	(82.8, 81.3)	(93.6, 90.1)
	BTAD	(59.0, 89.7)	(68.2, 67.8)	(88.2, 83.8)	(88.6, 88.2)	(94.8, 93.7)	(93.2, 91.9)	(90.7, 88.1)	(96.0, 91.0)
	MPDD	(42.7, 73.9)	(61.4, 77.5)	(77.5, 80.4)	(76.0, 82.5)	(75.1, 79.8)	(81.2, 83.5)	(76.6, 80.4)	(85.5, 87.8)
	RealIAD	(51.4, 64.6)	(74.7, 69.8)	(78.7, 80.0)	(79.2, 73.5)	(81.3, 76.4)	(85.2, 78.7)	(81.6, 75.2)	(88.0, 81.6)
	KSDD	(68.6, 37.6)	(93.3, 79.0)	(84.5, 71.1)	(97.1, 90.7)	(69.3, 57.1)	(88.2, 56.0)	(81.3, 71.1)	(92.5, 69.7)
	KSDD2	(91.6, 67.0)	(94.2, 71.5)	(94.1, 80.0)	(95.9, 86.7)	(95.9, 84.4)	(97.3, 87.6)	(95.6, 84.8)	(97.1, 79.2)
	DAGM	(87.1, 88.8)	(91.8, 87.6)	(97.7, 90.1)	(99.1, 97.5)	(93.2, 79.4)	(97.7, 95.7)	(97.3, 89.3)	(99.6, 95.8)
	DTD	(94.4, 93.5)	(95.1, 94.1)	(93.9, 93.6)	(95.5, 94.7)	(90.4, 92.8)	(95.1, 95.1)	(95.9, 94.7)	(99.4, 99.0)
	Average	(69.5, 75.4)	(83.2, 80.1)	(87.6, 83.6)	(89.6, 87.5)	(86.1, 81.4)	(90.8, 85.1)	(88.1, 84.1)	(94.1, 87.6)
Pixel-level (AUROC, max-F1)	MVTec AD	(75.5, 38.1)	(88.7, 43.4)	(91.1, 39.1)	(88.7, 43.4)	(91.4, 46.4)	(91.8, 49.0)	(90.8, 39.3)	(92.7, 45.2)
	VisA	(76.5, 31.6)	(95.5, 37.7)	(95.5, 28.3)	(95.5, 37.7)	(94.8, 30.2)	(95.6, 34.3)	(95.6, 27.6)	(96.2, 31.2)
	BTAD	(65.8, 14.8)	(92.1, 51.7)	(94.2, 49.7)	(92.1, 51.7)	(97.3, 55.1)	(93.9, 52.0)	(95.8, 52.6)	(92.3, 49.7)
	MPDD	(81.7, 18.9)	(96.1, 34.9)	(96.5, 34.2)	(96.1, 32.8)	(96.7, 30.0)	(97.8, 35.0)	(95.5, 31.9)	(97.0, 38.1)
	RealIAD	(73.5, 4.5)	(87.2, 10.8)	(96.3, 39.0)	(97.2, 43.0)	(96.2, 40.2)	(97.2, 41.2)	(96.2, 38.3)	(96.4, 40.4)
	KSDD	(78.8, 6.6)	(97.7, 54.5)	(90.6, 42.5)	(97.7, 54.5)	(87.1, 28.0)	(96.5, 6.6)	(93.1, 47.2)	(99.0, 10.1)
	KSDD2	(79.9, 63.4)	(94.4, 23.9)	(98.5, 59.8)	(98.5, 67.0)	(99.5, 63.4)	(97.0, 62.0)	(99.1, 60.4)	(99.3, 55.9)
	DAGM	(91.5, 57.5)	(91.5, 57.5)	(95.6, 58.9)	(91.5, 57.5)	(96.2, 53.3)	(95.9, 49.8)	(98.6, 60.2)	(99.4, 61.3)
	DTD	(97.9, 71.6)	(97.9, 71.6)	(97.9, 62.2)	(97.9, 71.6)	(95.8, 59.6)	(98.4, 65.2)	(98.1, 61.9)	(99.4, 66.5)
	Average	(80.1, 34.1)	(93.5, 42.9)	(95.1, 46.0)	(95.0, 51.0)	(95.0, 45.1)	(96.0, 43.9)	(95.9, 46.6)	(96.9, 44.3)

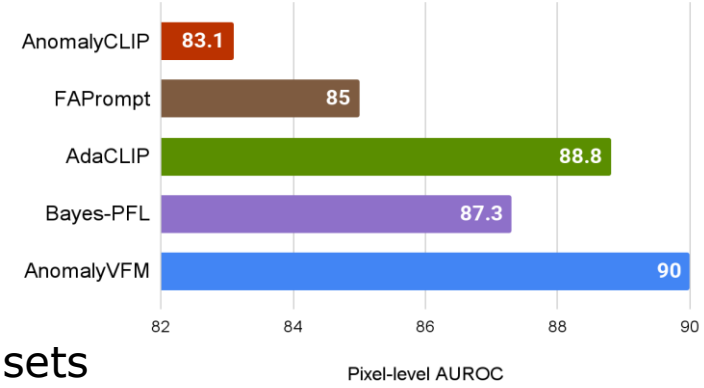
Zero-shot Anomaly Detection





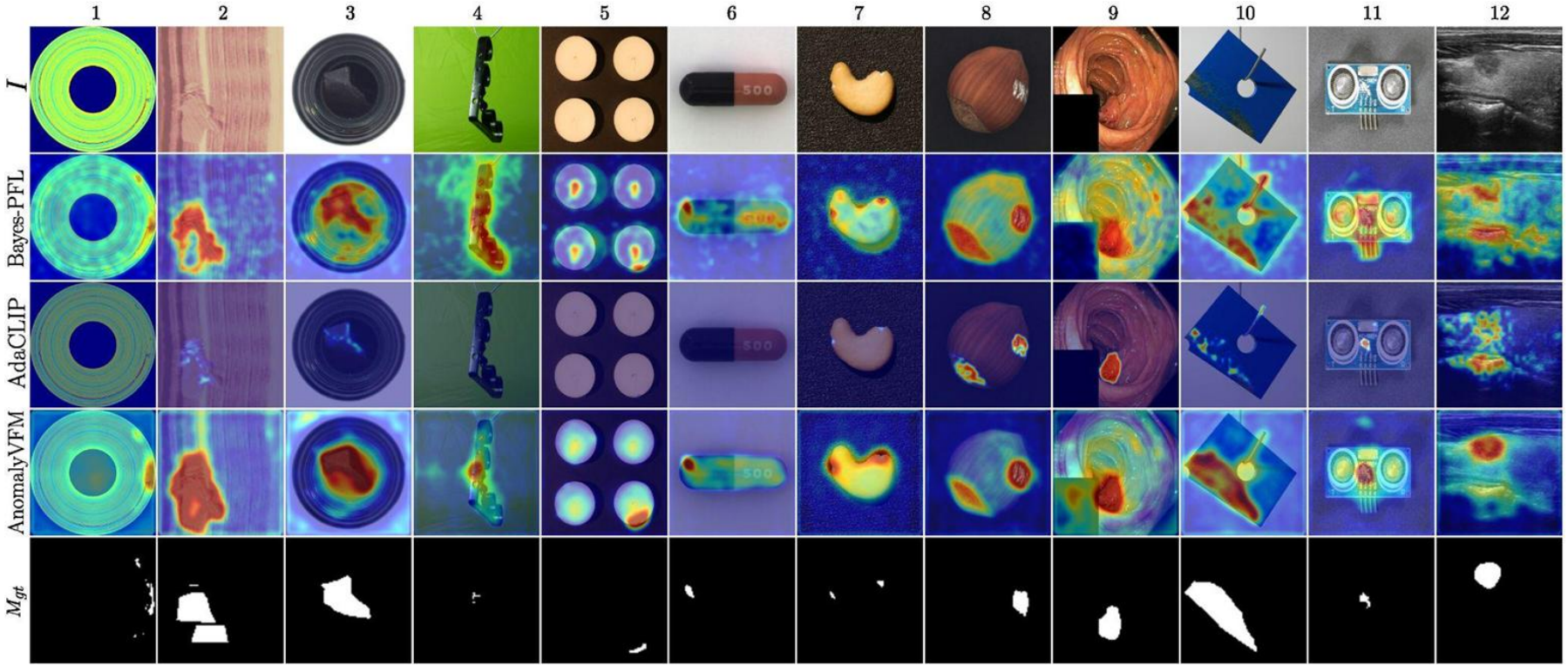
Industrial datasets





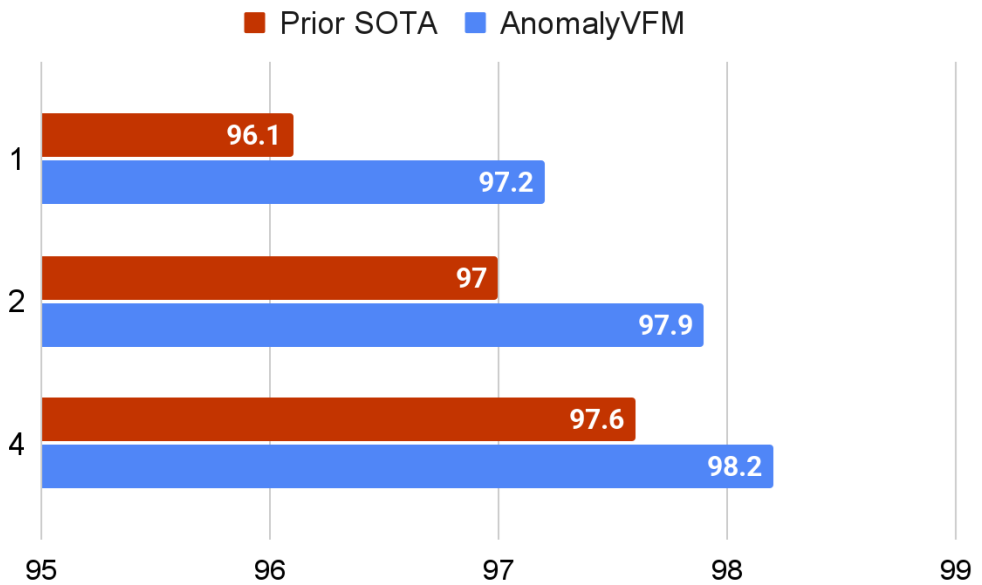
Medical datasets

Zero-shot AD – qualitative results

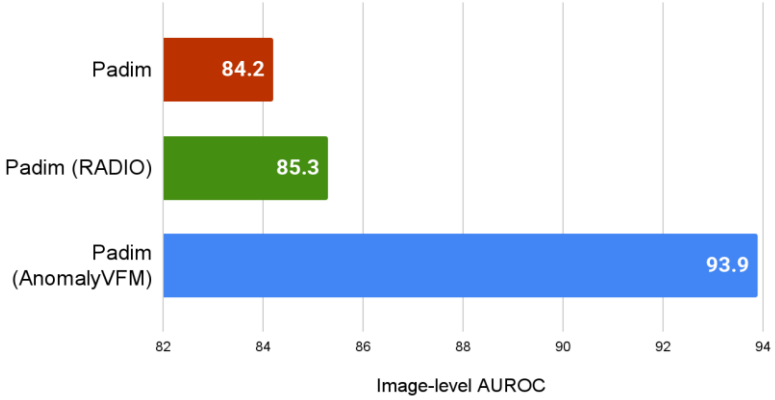


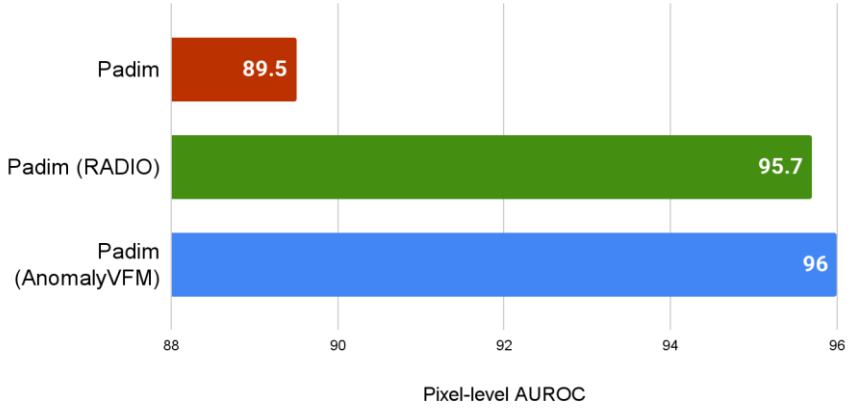
Few-shot anomaly detection

- Simply fine-tune zero shot model using few normal samples
- Results on MVTec AD:
- Better than few-shot SOTA!
- AnomalyVFM **strong backbone** for anomaly detection



VFM for AD – using it as a backbone

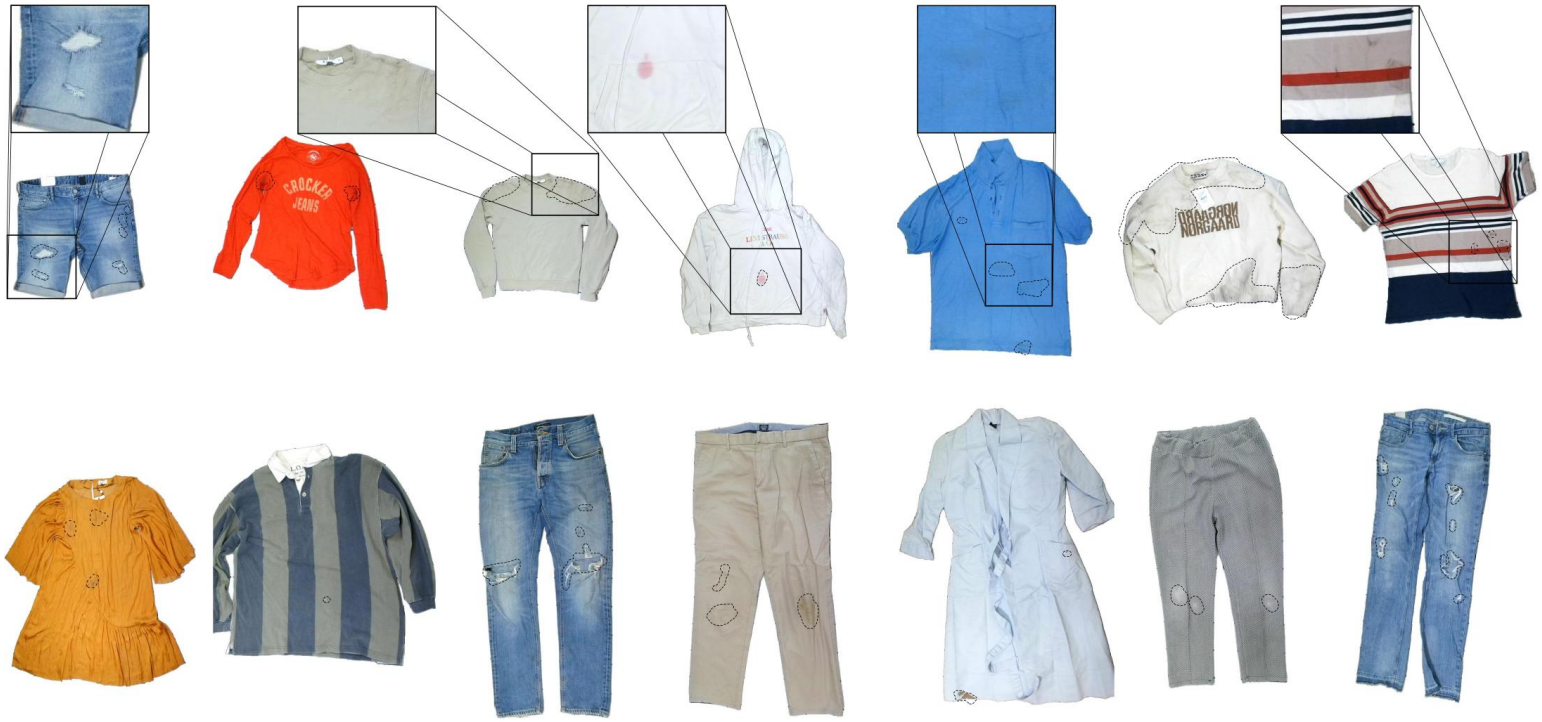




- Preliminary results
- Results show great potential
- MVTec AD under the multi-class setting (one model for all categories)

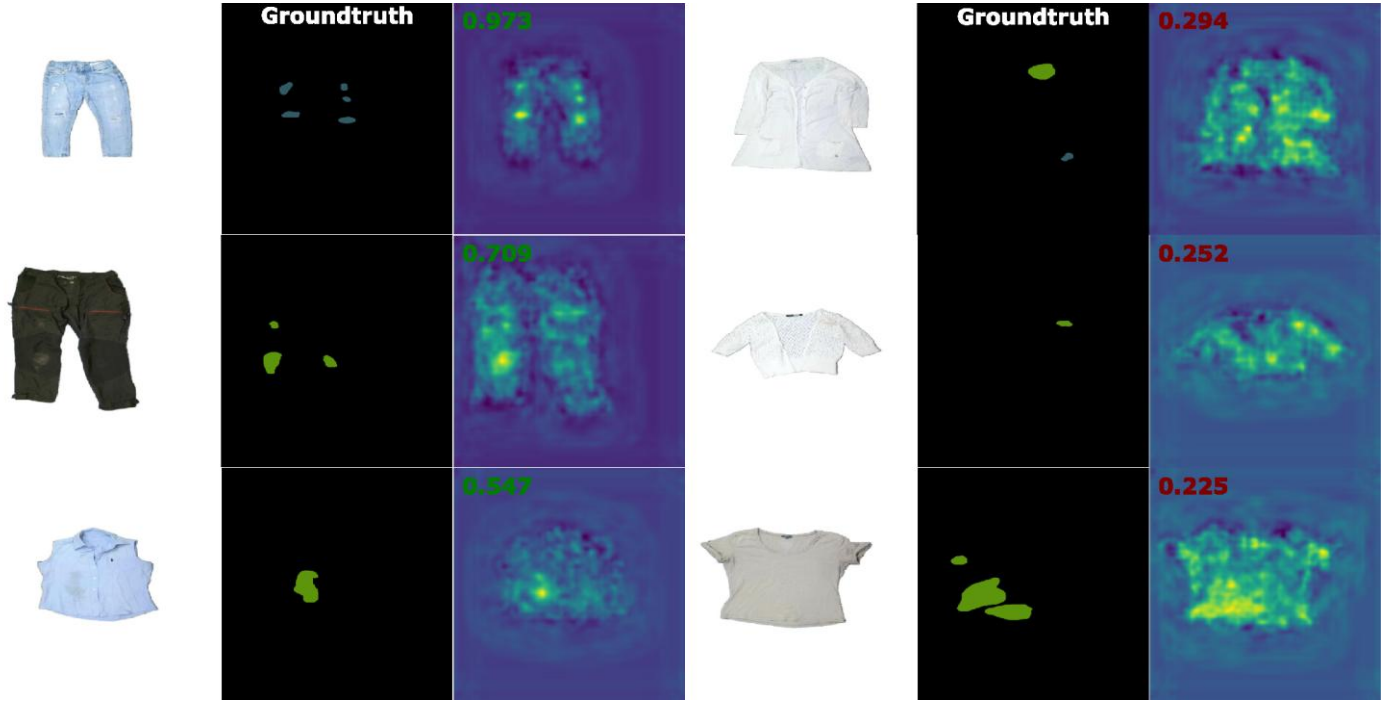
Anomaly detection on textile

- Challenging domain: high variabilty, wrinkles, subtle defects



[5] F. Nauman, “Clothing Dataset for Second-Hand Fashion”. Zenodo, Jun. 24, 2024

Anomaly detection on textile - SuperSimpleNet



Conclusion

- Data-driven deep-learning-based solutions
- AI/DL/CV/MV – key enabling technologies
  - Also for robotics!
- Wide applicability, interdisciplinarity
- Robustness
- New challenges, new opportunities
- Collaboration between academia and industrial partners
- Use all data available ;-)