UNIVERSITY OF TARTU
Institute of Computer Science

BCV Lab revvity

CVPR Nashville JUNE 11-15, 2025

IAUNet: Instance-Aware U-Net

Yaroslav Prytula, Dmytro Fishman SUP

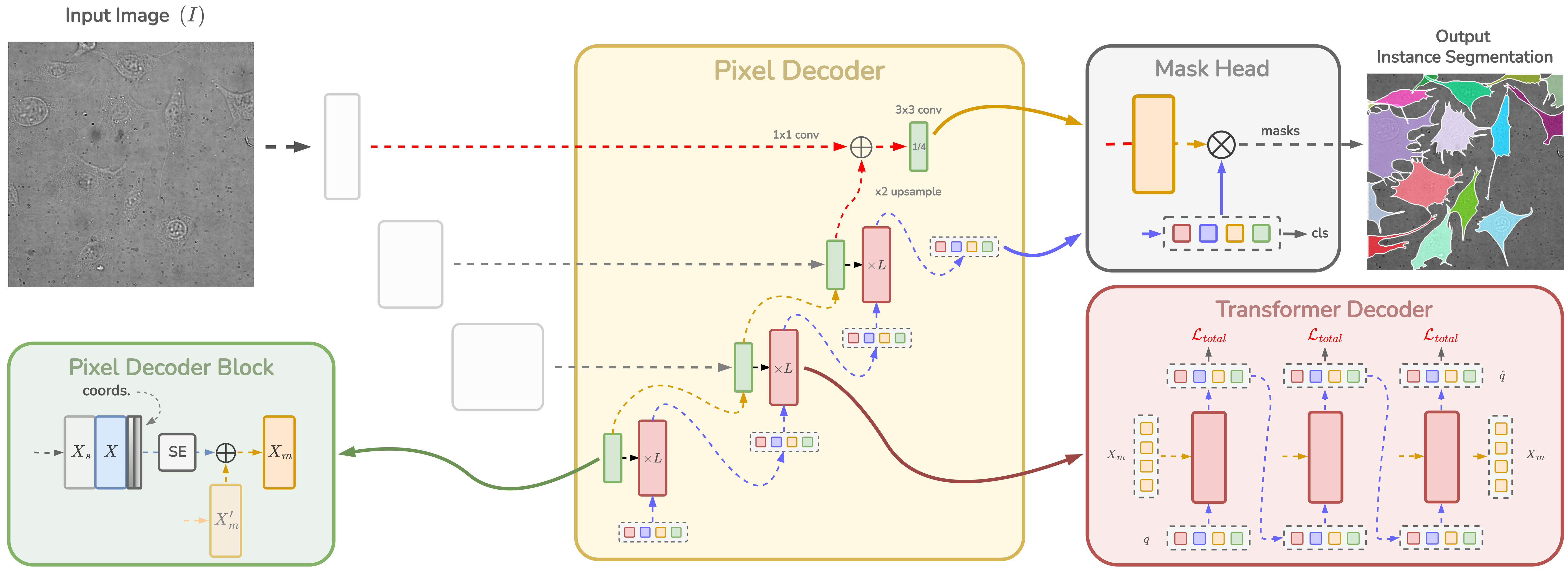


Figure 1. Model overview. IAUNet consists of a Pixel decoder and a Transformer decoder. The encoder extracts multi-scale features used as skip connections in the Pixel decoder. Each decoder block combines these features with CoordConv-based positional encodings and applies stacked depth-wise convolutions followed by a Squeeze-and-Excitation (SE) block to produce refined mask features. The Transformer decoder then refines learnable queries over multiple layers using these mask features with deep supervision.

Abstract

Instance segmentation is critical in biomedical imaging to distinguish individual objects like cells, which often overlap and vary in size. We propose IAUNet, a novel query-based U-Net architecture that retains the full U-Net design and adds a lightweight convolutional Pixel decoder for efficient multi-scale feature aggregation. To enhance instance segmentation, we incorporate a Transformer decoder with deep supervision that refines object queries across layers. We also introduce Revvity-25, a new 2025 dataset with detailed annotations of overlapping cell cytoplasm in brightfield images. IAUNet achieves strong results, outperforming existing convolutional, transformer-based, and query-based models.

Revvity-25

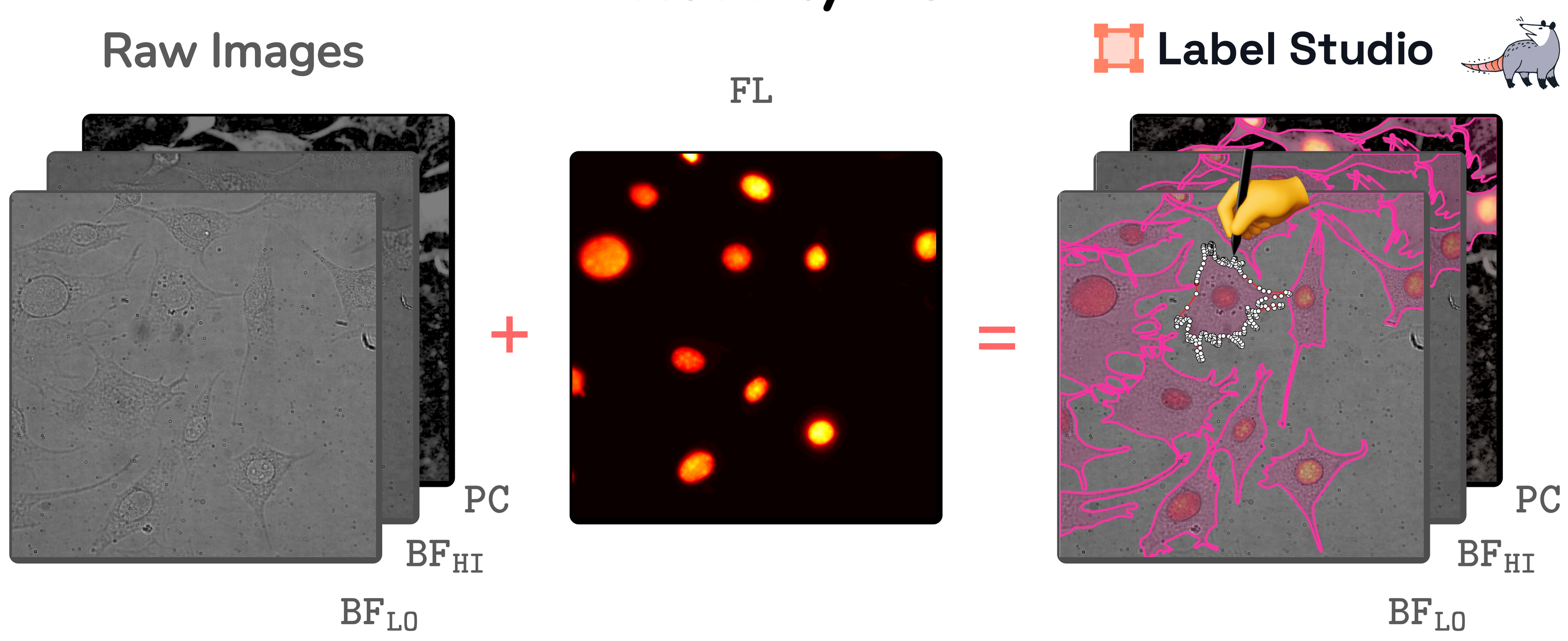


Figure 2. Multimodal annotation workflow for the Revvity-25 dataset.

Revvity-25 comprises 110 brightfield images with 2,937 expert-validated cell instances, each labeled with high-fidelity polygon masks averaging 60 points per cell (up to 400). It is the first public dataset to pair high-resolution brightfield images with precise instance-level annotations of overlapping cells.

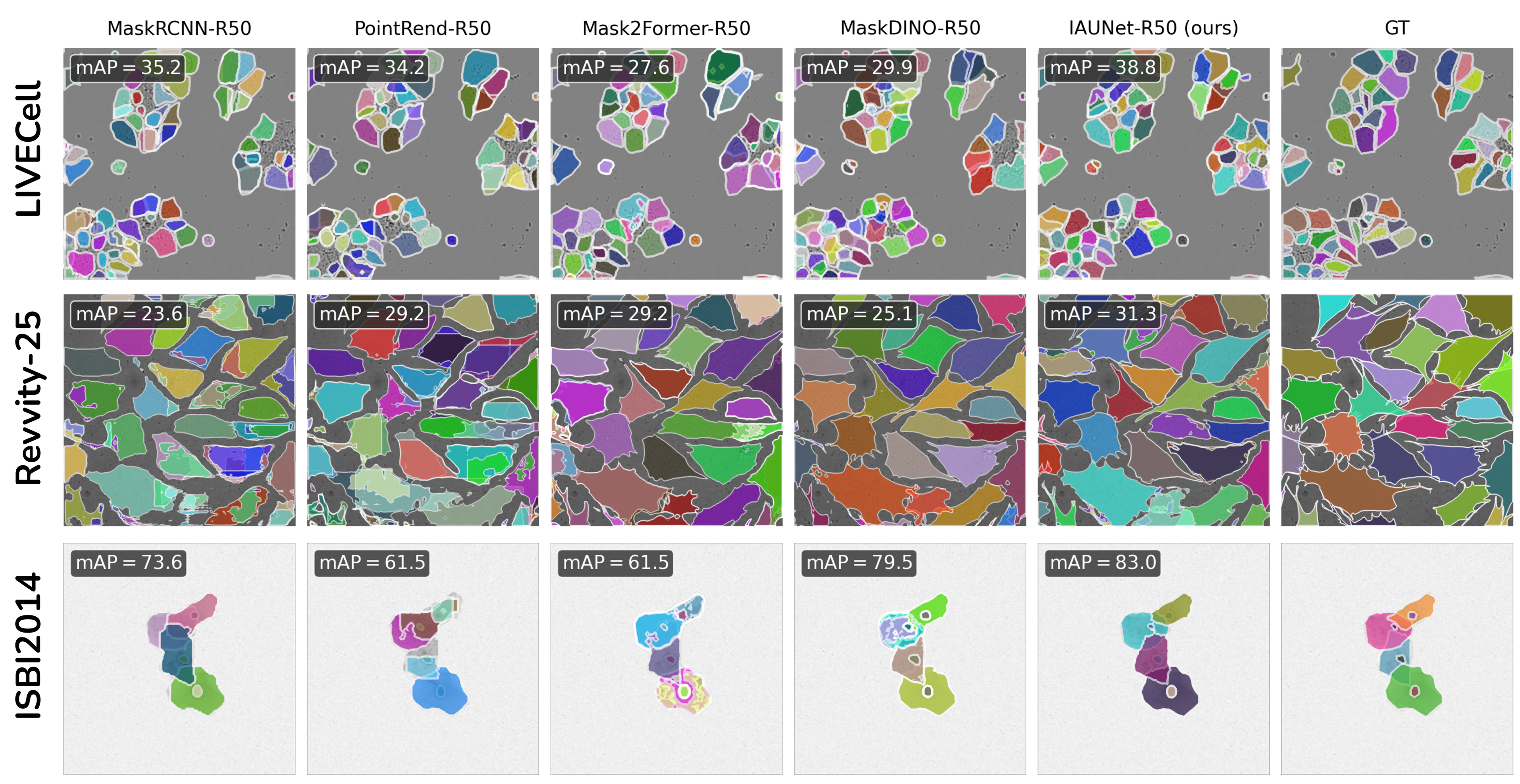


Figure 3. Visualization of instance segmentation predictions across different state-of-the-art models (using ResNet50 backbone). We also report per-image AP score.

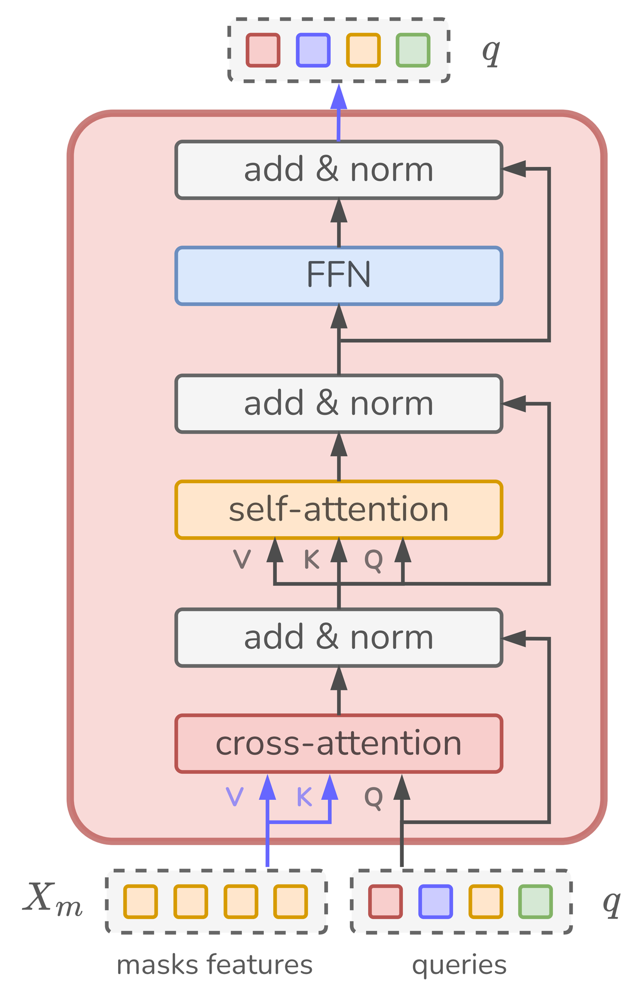
Acknowledgments

We express gratitude to the Armed Forces of Ukraine and the bravery of the Ukrainian people for enabling a secure working environment, without which this work would not have been possible. Special thanks to my supervisor, Dmytro Fishman, for invaluable discussions, and to the Biomedical Computer Vision Lab for their continuous support and collaboration. We thank the High-Performance Computing Cluster at the University of Tartu for providing the computational resources.

Design

Pixel Decoder

- A lightweight Pixel decoder is designed to refine multi-scale features.
- Features are processed through lightweight depth-wise convolutions.
- CoordConv injects explicit positional information into the decoder without increasing computational complexity.
- Squeeze-and-Excitation (SE) block enhances feature refinement for better instance separation.



Transformer Decoder

- Transformer decoder learns instance-level representations.
- Uses learnable queries for potential objects.
- Queries attend to mask features via cross- and self-attention.
- Three blocks per layer refine semantic and spatial content.

Revvity-25

Models	backbones	num_queries	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	#params.	FLOPs
Models with Convolution-Based Backbones										
Mask R-CNN [14]	R50	100	39.7	77.2	37.4	0.6	19.0	44.6	44M	115G
PointRend [34]	R50	100	42.2	79.4	40.9	0.4	21.7	47.3	56M	66G
Mask2Former [19]	R50	100	46.4	79.8	49.9	0.7	25.7	52.8	44M	67G
MaskDINO [20]	R50	100	45.6	80.4	48.2	**1.8**	22.3	51.8	44M	64G
IAUNet (ours)	R50	100	**49.7**	**82.1**	**54.8**	0.6	**27.3**	**56.0**	39M	49G
Mask R-CNN [14]	R101	100	40.7	77.5	39.9	0.4	20.1	45.8	63M	134G
PointRend [34]	R101	100	42.9	79.3	42.5	0.0	18.4	48.9	75M	86G
Mask2Former [19]	R101	100	47.2	80.1	51.8	**1.7**	25.7	53.3	63M	86G
MaskDINO [20]	R101	100	47.3	81.0	50.4	0.9	23.0	53.5	63M	84G
IAUNet (ours)	R101	100	**51.5**	**84.7**	**56.1**	**1.7**	**29.2**	**57.8**	58M	69G
Models with Transformer-Based Backbones										
Mask R-CNN [14]	Swin-S	100	24.7	63.4	12.5	0.0	7.3	28.9	69M	141G
PointRend [34]	Swin-S	100	43.6	80.0	43.0	0.5	21.5	48.9	81M	93G
Mask2Former [19]	Swin-S	100	51.2	83.3	56.4	2.7	27.7	58.0	69M	93G
MaskDINO [20]	Swin-S	100	50.3	83.2	53.9	**4.7**	27.6	56.1	71M	181G
MaskDINO [20]	Swin-S	300	49.4	83.6	53.3	2.9	25.8	55.3	71M	187G
IAUNet (ours)	Swin-S	100	53.0	85.7	57.0	1.3	**29.7**	59.1	64M	76G
IAUNet (ours)	Swin-S	300	**53.3**	**86.0**	**59.6**	1.6	29.4	**59.8**	64M	87G
Mask R-CNN [14]	Swin-B	100	27.1	64.9	17.2	0.1	9.7	31.2	107M	186G
PointRend [34]	Swin-B	100	45.2	80.1	47.9	0.1	23.0	50.9	119M	137G
Mask2Former [19]	Swin-B	100	52.0	83.6	58.4	1.1	27.8	59.0	107M	138G
MaskDINO [20]	Swin-B	100	50.5	83.5	54.9	**2.0**	27.1	56.4	110M	226G
MaskDINO [20]	Swin-B	300	50.4	84.3	54.8	0.8	26.3	56.6	110M	232G
IAUNet (ours)	Swin-B	100	53.5	86.1	**59.4**	0.8	**30.5**	59.7	102M	120G
IAUNet (ours)	Swin-B	300	**53.7**	**86.5**	**59.4**	1.0	30.0	**60.3**	102M	132G

Table 1. Instance segmentation on our Revvity-25 dataset. IAUNet outperforms strong query-based baselines as well as other state-of-the-art models when training with fewer parameters

Conclusions

We introduce IAUNet, a query-based U-Net with a lightweight Pixel decoder and a Transformer decoder for efficient cell instance segmentation. IAUNet achieves strong performance with low computational cost. We also present Revvity-25, a high-resolution microscopy dataset with expert-labeled cell masks for modal and amodal segmentation. This work sets a strong baseline for future research and will be presented at CVPR 2025 at the CVMI Workshop in Nashville.