
NASA
esa
International workshop on AI Foundation Model for EO
5-7 May 2025 | ESA ESRIN - Frascati, Italy

HiRes-FusedMIM: A High-Resolution RGB-DSM Pre-trained Model for Building-Level Remote Sensing Applications

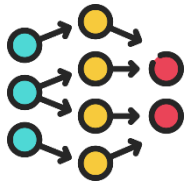
G. Mutreja[1] *, Dr. P. Schuegraf, Dr. K. Bittner[1]

[1] Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany

Why High-Res RGB + DSM?

- Existing pretrained models use Sentinel-2, 10–60 m resolution – insufficient for building detail.
- DSMs (elevation) rarely used, though critical for 3D urban structure.
- Urban tasks (roof segmentation, height estimation) demand sub-meter detail.

Key Contributions

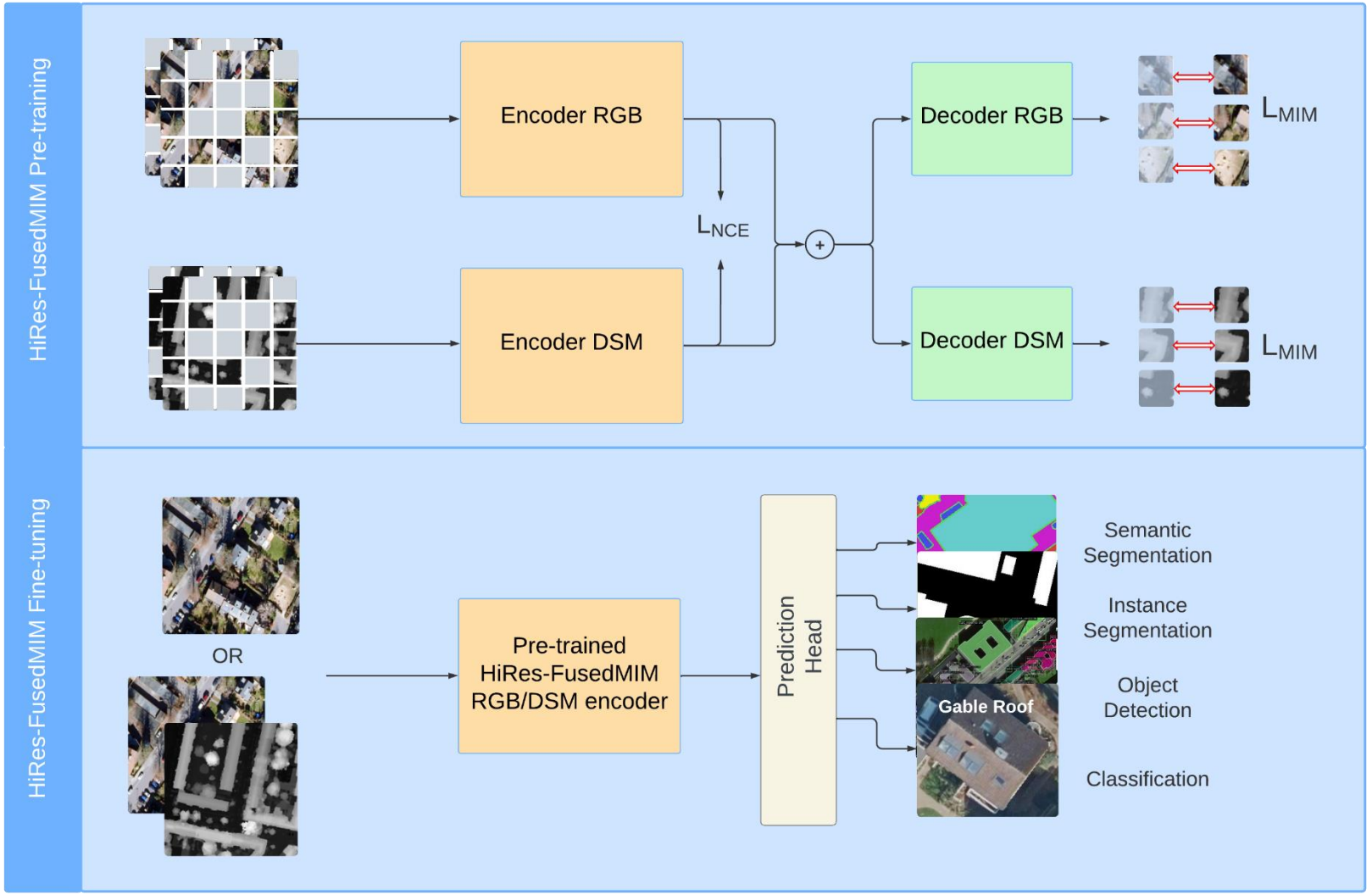


368 k paired RGB-DSM images @0.2–0.5 m

Dual-encoder SimMIM w/ contrastive loss

Strong performance on building-related segmentation and detection tasks

Model Architecture



Pre-training Objectives

- **MIM Loss**: L_1 reconstruction for RGB & DSM

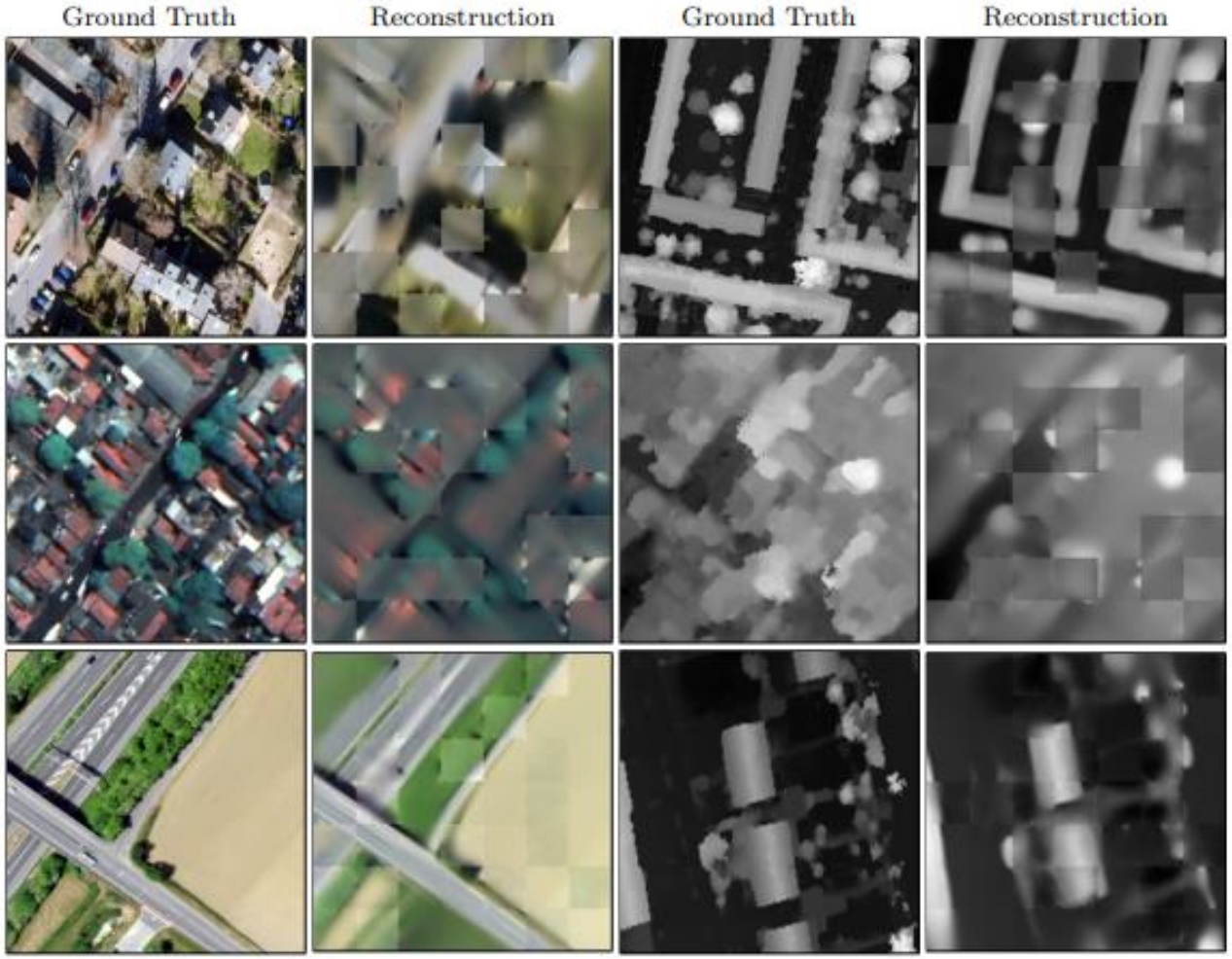
$$\mathcal{L}_{\text{MIM}} = \frac{||\boldsymbol{X}_{\text{RGB}} - \boldsymbol{R}_{\text{RGB}}||_1 + ||\boldsymbol{X}_{\text{DSM}} - \boldsymbol{R}_{\text{DSM}}||_1}{N}$$

- **Contrastive Loss**: InfoNCE between encoder outputs.

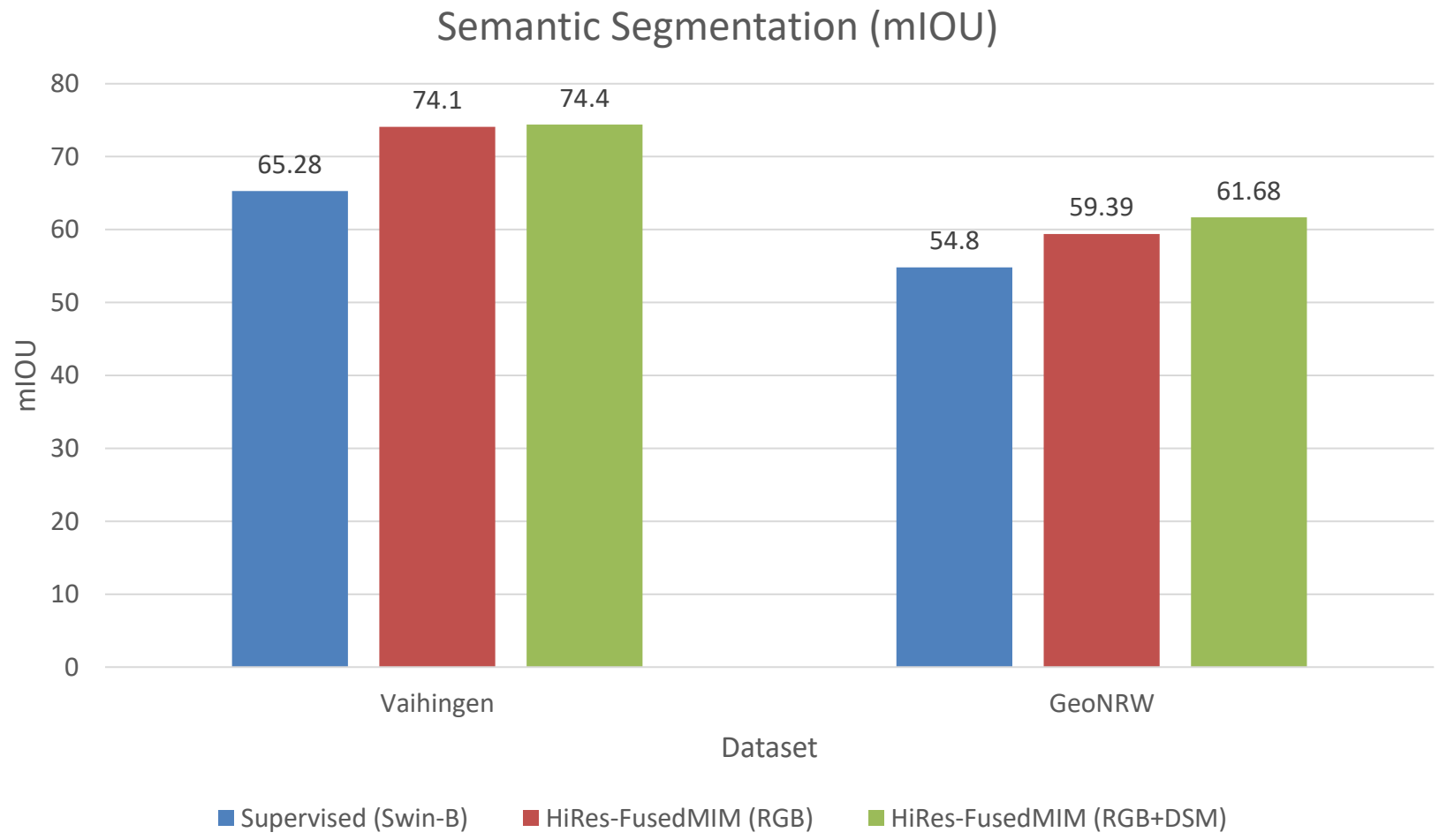
$$\mathcal{L}_{\text{InfoNCE}} = -\log\left(\frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{K}\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}\right)$$

Loss Component	Weight	Purpose
Reconstruction	95%	Learn fine-grained spatial features
Contrastive InfoNCE	5%	Align RGB and DSM latent spaces

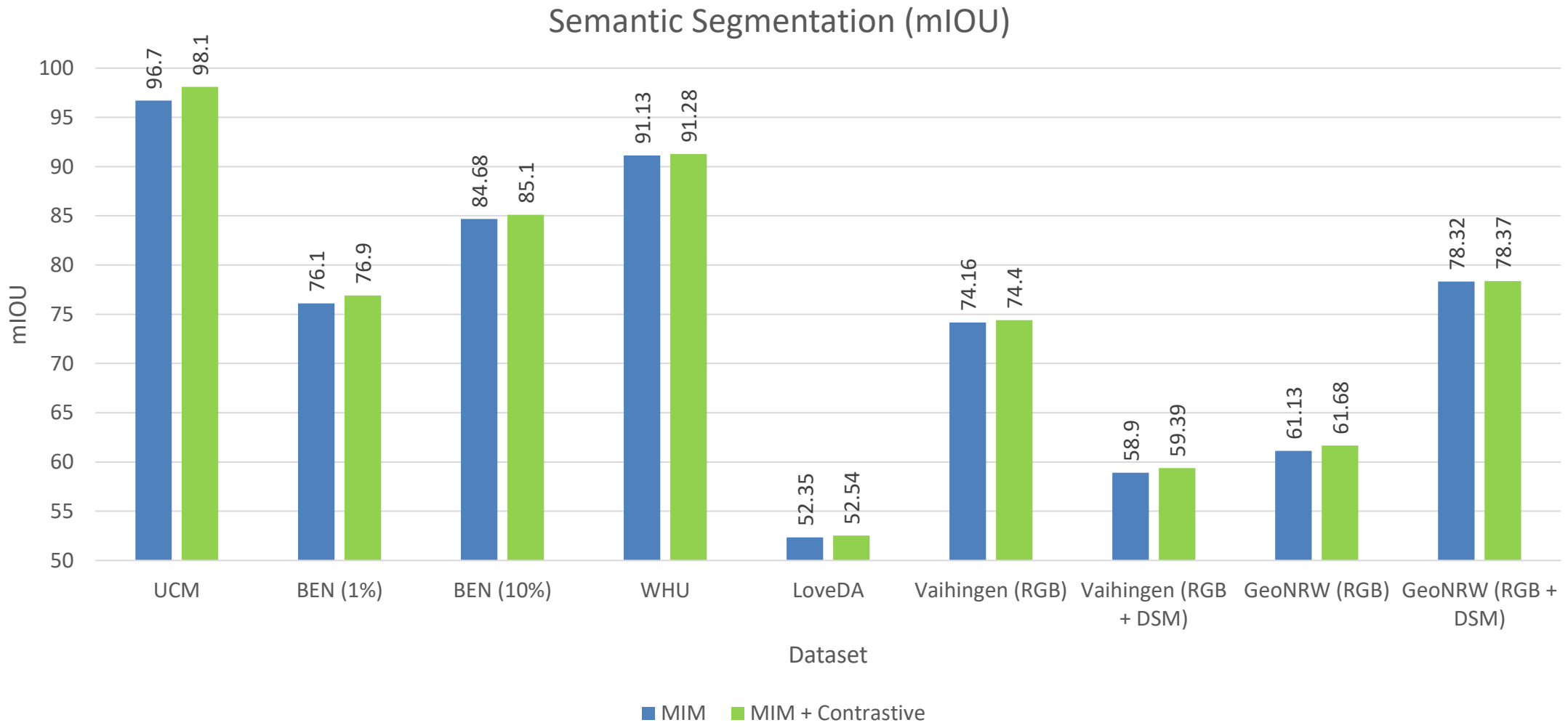
Reconstruction Capabilities



Impact of the DSM Modality



Impact of the Contrastive loss



Results: Semantic Segmentation (mIOU)

Method	Backbone	WHU Aerial	LoveDA	Vaihingen	SpaceNetV1
SeCO	ResNet50	86.7	43.63	68.9	77.09
SatMAE	ViT-L	82.5	-	70.6	78.07
GFM	Swin-B	90.7	-	**75.2**	72.81
MAE+MTP	ViT-B+RVSA	-	52.39	-	**79.63**
HiRes-FusedMIM	Swin-B	**91.28**	**52.54**	74.16	78.37

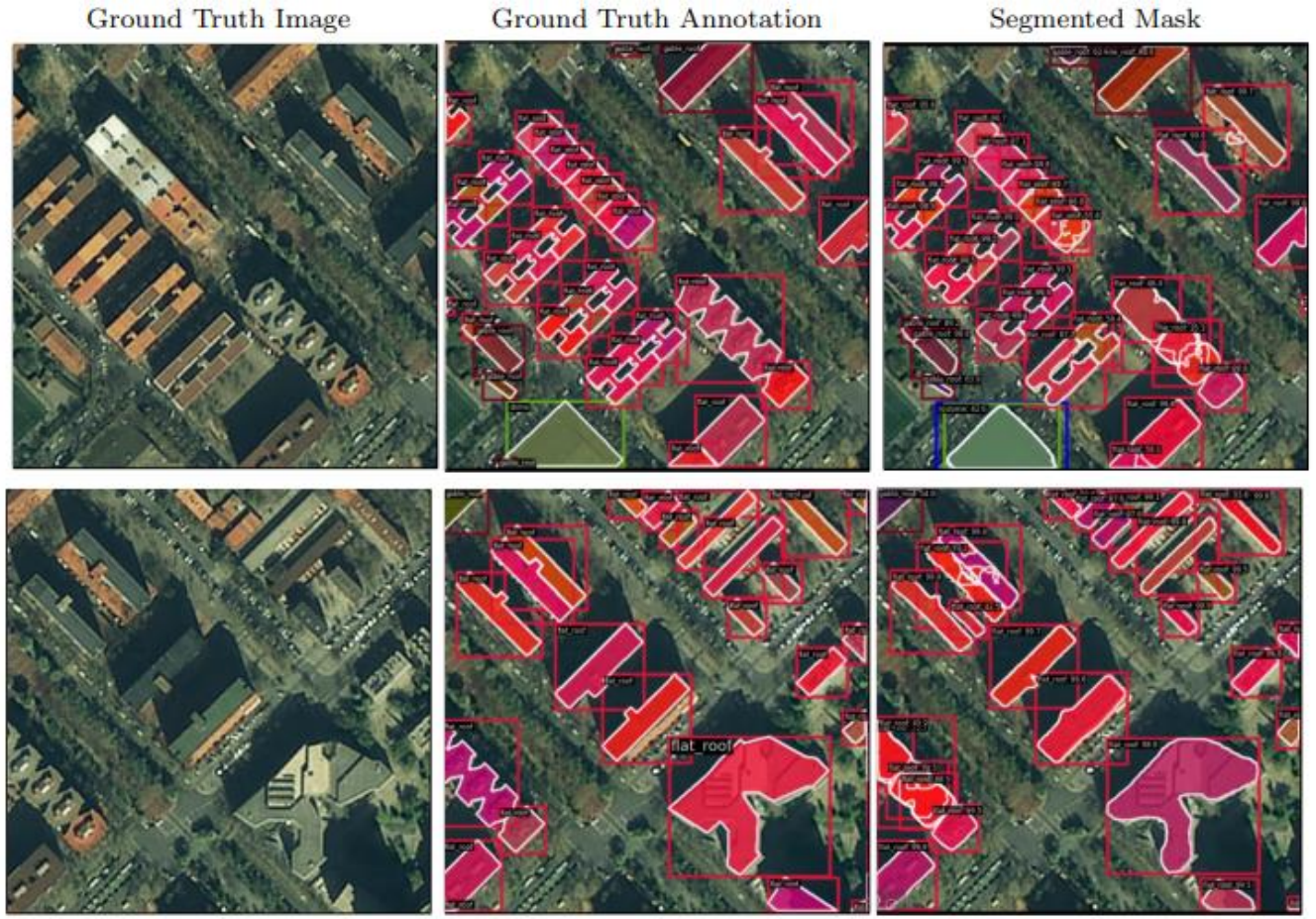
Results: Semantic Segmentation (visuals)

Results: Instance Segmentation (AP)

Method	AP	AP_{50}	FL	GA	GM	RO	ME	H1	H2	MA	PY	AR	RE	OT
SOLOv2	14.2	24.0	24.6	21.3	27.7	5.8	9.5	7.6	35.0	3.6	4.6	14.8	11.2	6.1
QueryInst	15.3	25.2	25.5	23.4	28.6	5.1	24.2	8.9	33.9	7.0	4.8	13.9	4.5	5.5
Mask R-CNN	15.5	25.9	25.8	24.7	29.2	5.1	12.3	10.0	39.0	5.4	5.6	16.8	5.8	7.2
Cascade Mask R-CNN	16.5	26.9	26.5	24.5	29.1	4.7	19.7	10.9	38.6	6.0	7.0	18.5	6.7	6.4
Mask R-CNN + CGT	16.3	26.3	25.2	25.1	27.4	9.0	21.6	9.8	39.6	5.7	5.8	17.2	3.2	6.8
Cascade Mask R-CNN + CGT	17.1	27.1	27.2	25.0	28.6	6.6	24.9	11.1	39.8	7.7	5.8	16.4	5.7	7.2
Mask R-CNN (SWIN-B)	14.8	24.5	43.8	40.8	49.4	4.3	30.2	8.7	47.2	6.7	7.7	13.3	**30.5**	11.7
HiRes-FusedMIM	**17.7**	**28.5**	**47.32**	**45.4**	**54.9**	**13.7**	**27.8**	**16.1**	**59.0**	**12.6**	**12.7**	**20.9**	18.6	**12.4**

AP (%) on UBCv2 Test set with FineGrained roof classes: FL-Flat, GA-Gable, GM-Gambrel, RO- Row, ME- Multiple Eave, H1- Hipped V1, H2- Hipped V2, MA- Mansard, PY- Pyramid, AR- Arched, RE- Revolved and OT-Other

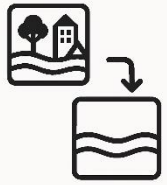
Results: Instance Segmentation (visuals)



Current Limitations

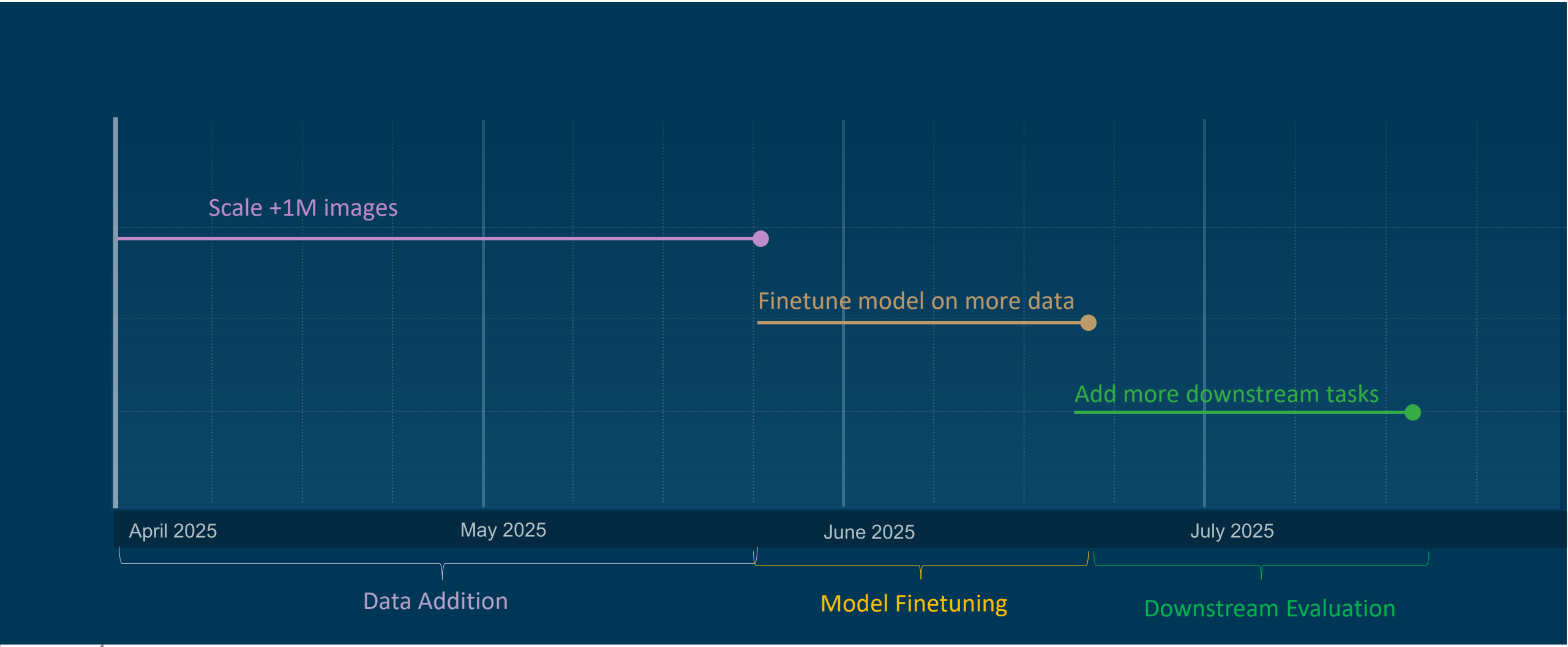
Dataset licensing – cannot fully open 368 k images.

Modest DSM gains (e.g., +0.2–2% mIoU).



Unaddressed tasks: height estimation, DSM→DTM, change detection.

Next Steps


Scale +1M images
Finetune model on more data
Add more downstream tasks
April 2025
May 2025
June 2025
July 2025
Data Addition
Model Finetuning
Downstream Evaluation

Conclusion & Collaboration

- **HiRes-FusedMIM** advances building-level tasks via high-res RGB+DSM.
- Strong results on classification, segmentation, detection.
- **Next**: open data, new regression tasks, advanced fusion.
- **Let’s collaborate**

Thank you

G. Mutreja[1] *, Dr. P. Schuegraf, Dr. K. Bittner[1]

[1] Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany

Guneet.mutreja@dlr.de