Context-Aware Semantic Segmentation via Stage-Wise Attention

Supplementary Material

7. URUR: illustrative annotation mismatch

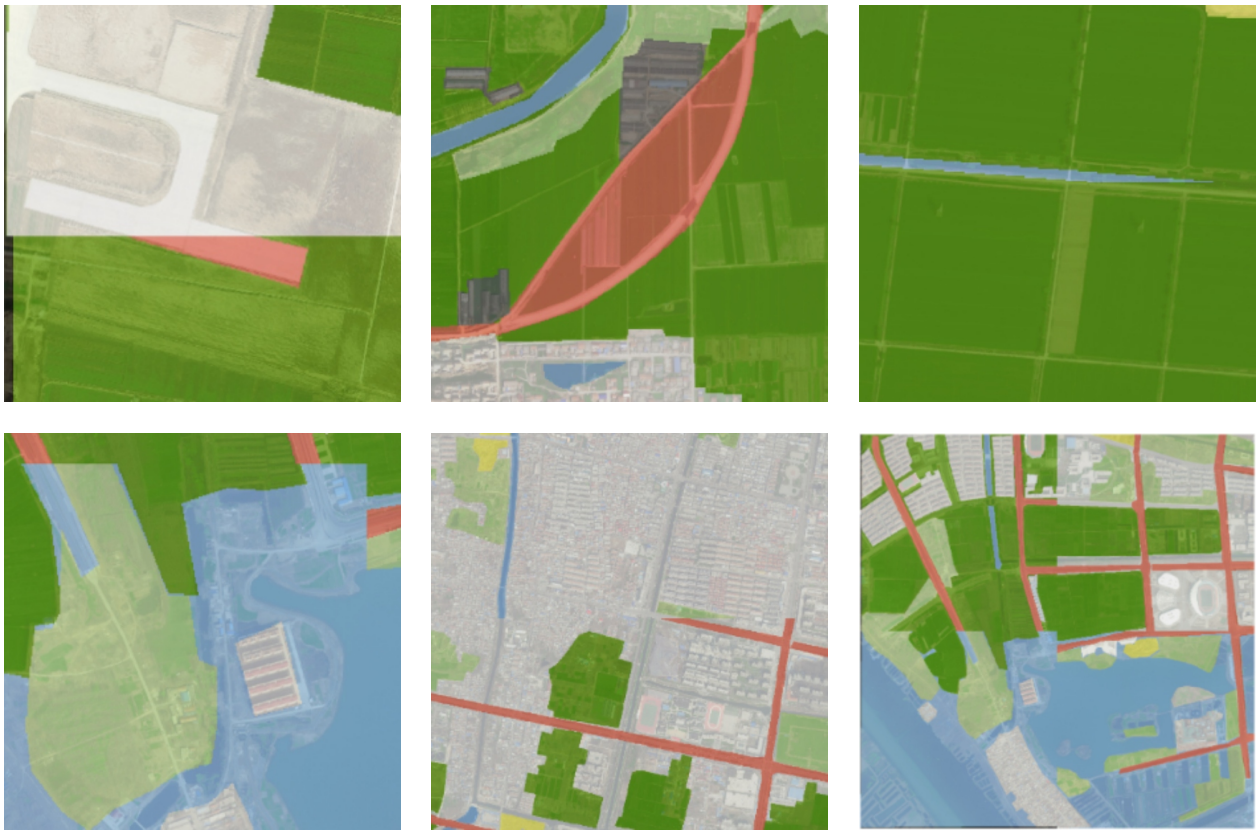
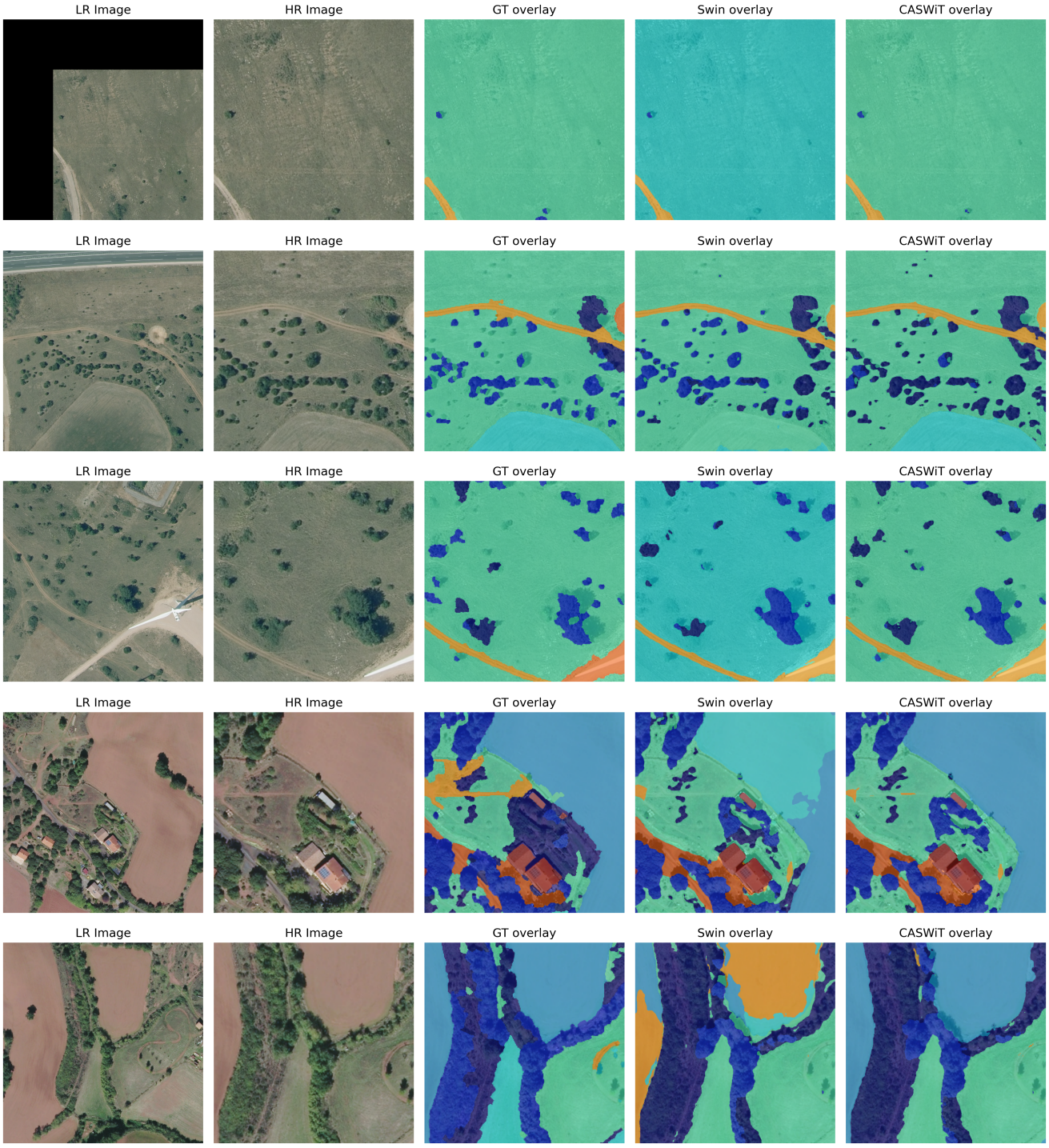


Figure 6. Example where the provided mask (overlaid) locally diverges from the RGB content; such cases are occasional but can affect evaluation metrics. Visible classes include: **others**, **building**, **greenhouse**, **woodland**, **farmland**, **bareland**, **water**, **road**.

8. Qualitative analysis on FLAIR-HUB



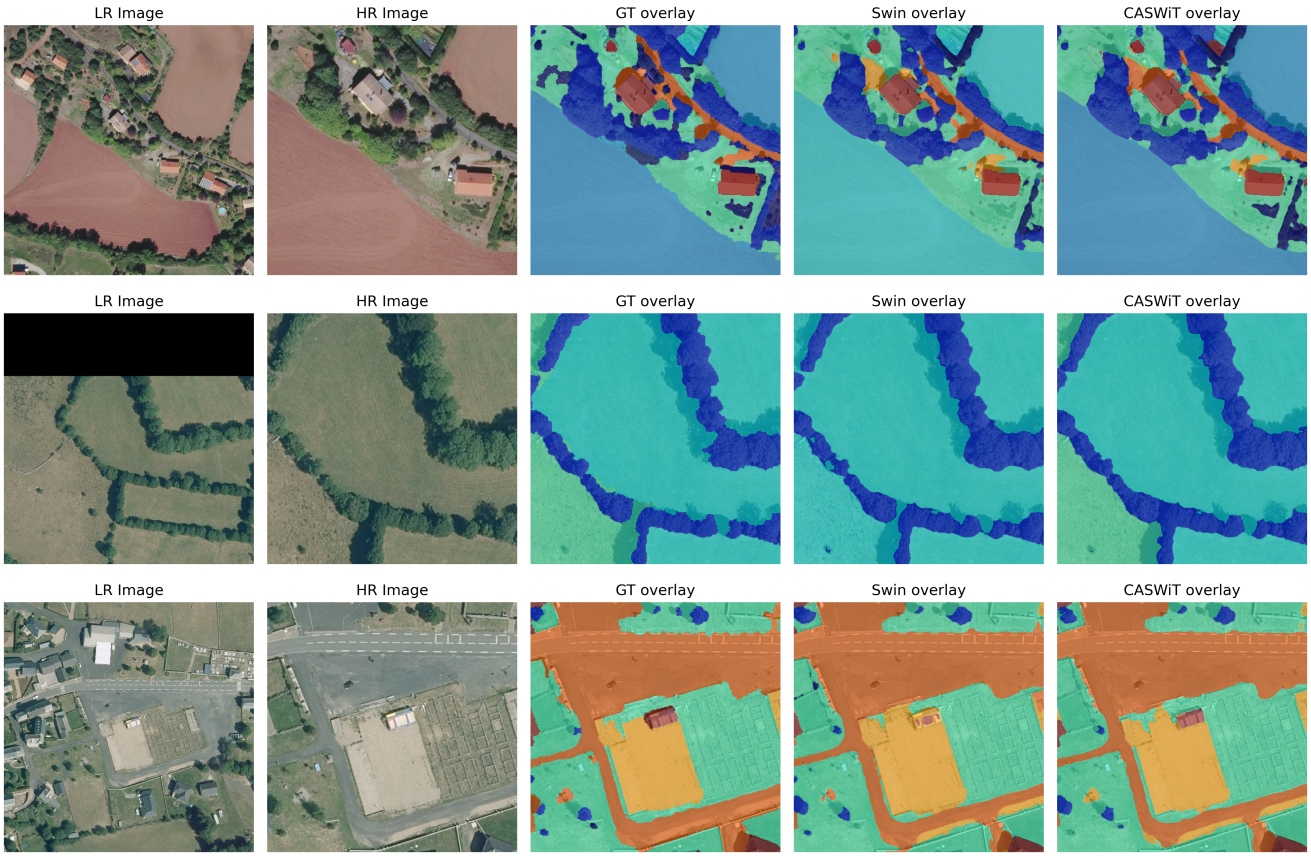


Figure 7. Comparison of LR/HR images, ground truth overlays, Swin Base predictions, and CASWiT predictions on eight test patches.

9. Supplementary results (IoUs)

Class	Swin-B [14]	CASWiT-B-SSL-aug + UPer	CASWiT-B-SSL-aug + SegF
Building	83.77	85.47	85.27
Greenhouse	77.89	79.46	80.67
Swimming pool	61.59	62.12	60.48
Impervious surface	75.03	76.78	76.94
Pervious surface	56.97	58.86	58.73
Bare soil	65.21	66.95	67.79
Water	90.08	90.65	90.35
Snow	67.77	66.59	75.95
Herbaceous vegetation	52.85	55.07	54.39
Agricultural land	56.53	60.38	60.63
Plowed land	37.34	38.20	38.49
Vineyard	78.88	80.71	80.71
Deciduous	70.07	71.47	70.89
Coniferous	58.95	62.89	62.73
Brushwood	30.97	31.79	31.61
mIoU	64.05	65.83	**66.37**

Table 5. Per-class IoU (%) on the FLAIR-HUB-RGB test set for Swin-B and our CASWiT variants.

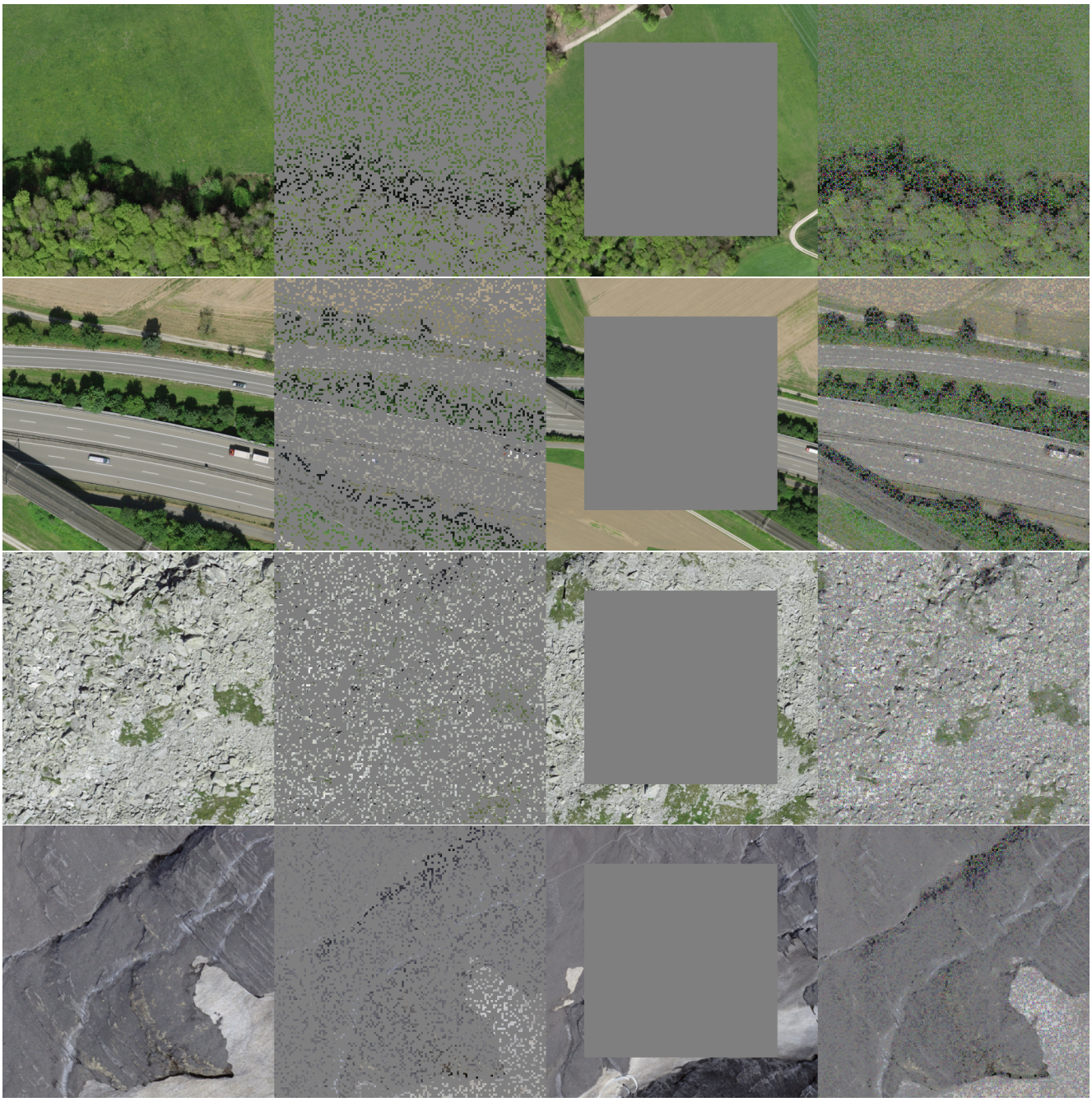
Class	WSDNET [23]	Boosting Dual-Stream [37]	CASWiT-B-SSL-aug + UPer	CASWiT-B-SSL-aug + SegF
Others	-	-	0.00	0.00
Building	-	-	75.07	74.42
Farmland	-	-	79.19	79.31
Greenhouse	-	-	46.51	46.50
Woodland	-	-	52.10	51.79
Bareland	-	-	31.64	32.31
Water	-	-	54.90	55.86
Road	-	-	53.33	53.67
mIoU	46.9	48.2	49.1	**49.2**

Table 6. Per-class IoU (%) on the URUR dataset test set for our CASWiT-Base variants.

10. Dataset FLAIR-HUB merge

Figure 8. Examples of data pre-processing, on the left are the HR patches and on the right are the merged patches obtained from the available neighbors.

11. SSL results



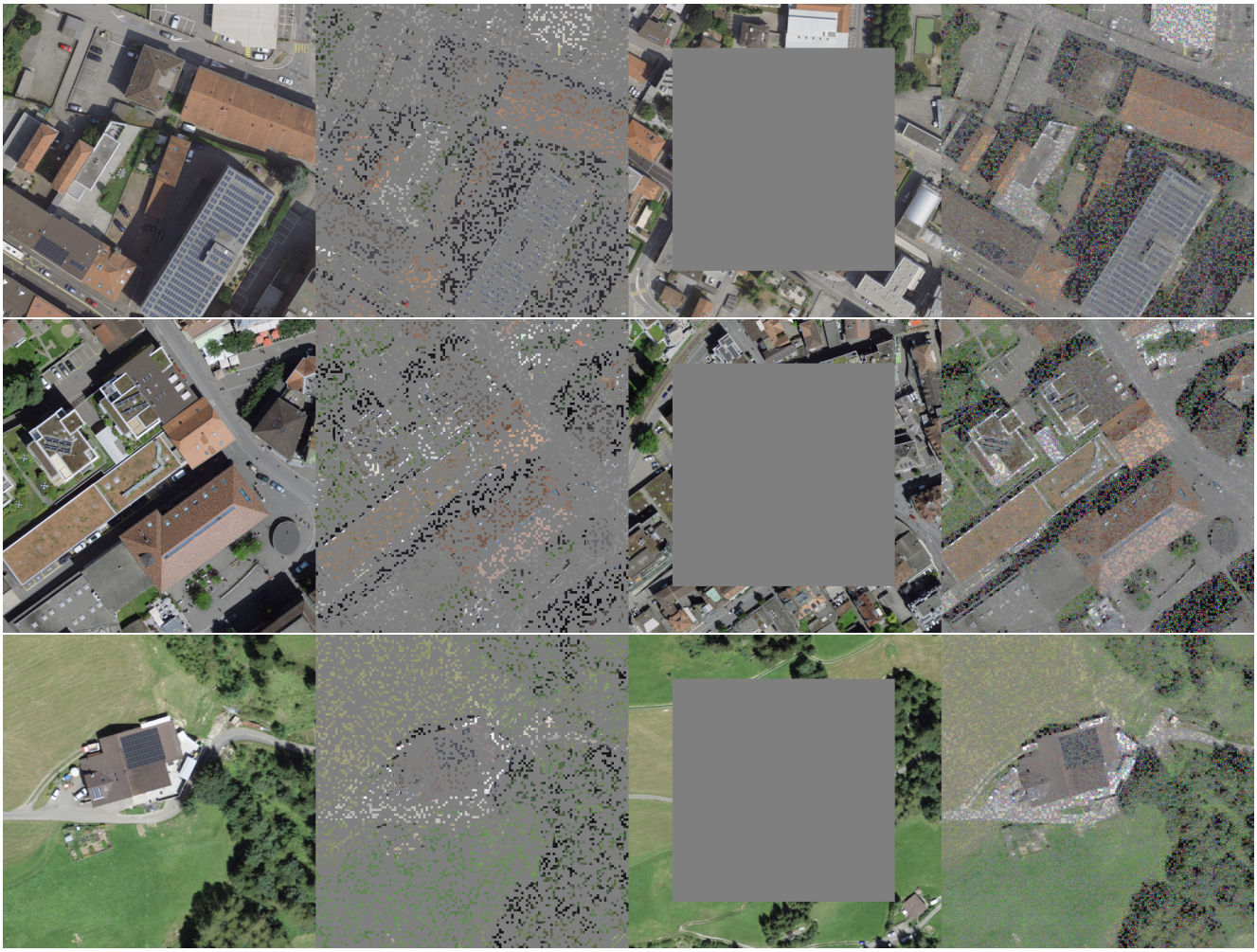


Figure 9. Self-supervised SimMIM-style inference results on the CASWiT-Base architecture. Each row (left to right) shows: original high-resolution image, high-resolution image with random masking, low-resolution image with central masking, and the reconstruction of the high-resolution image.

12. Cross-attention visualization

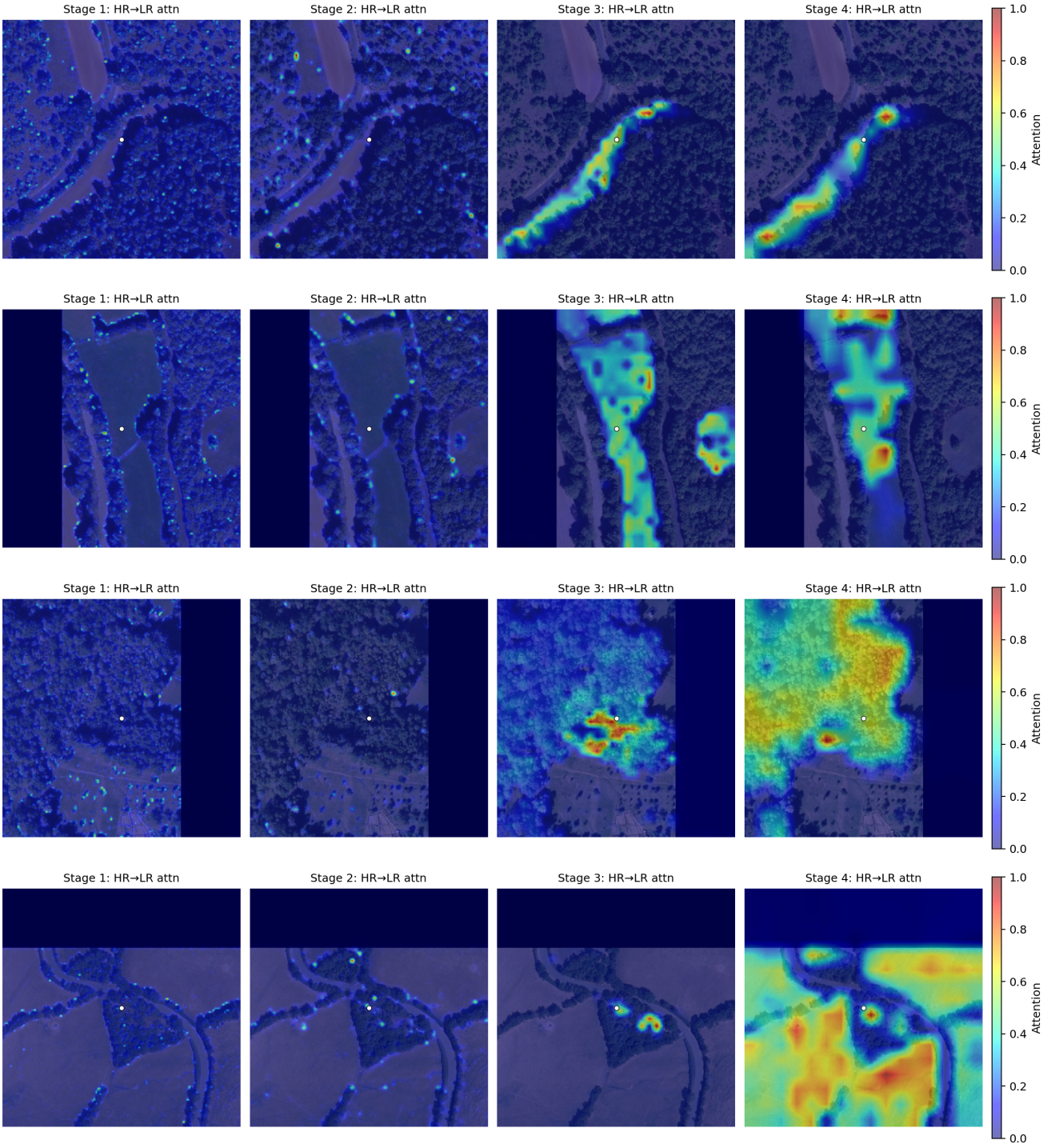


Figure 10. Visualization of cross-attention maps for each stage of the model on four test patches.