Context-Aware Semantic Segmentation via Stage-Wise Attention

Antoine Carreaud
EPFL & HEIG-VD
antoine.carreaud@epfl.ch

Elias Naha
EPFL & HEIG-VD
elias.naha@epfl.ch

Arthur Chansel
EPFL & HEIG-VD
arthur.chansel@epfl.ch

Nina Lahellec
EPFL & HEIG-VD
nina.lahellec@epfl.ch

Jan Skaloud
EPFL
jan.skaloud@epfl.ch

Adrien Gressin
HEIG-VD
adrien.gressin@heig-vd.ch

Abstract

Semantic ultra-high-resolution (UHR) image segmentation is essential in remote sensing applications such as aerial mapping and environmental monitoring. Transformer-based models remain challenging in this setting because memory grows quadratically with the number of tokens, limiting either spatial resolution or contextual scope. We introduce ***CASWiT*** *(Context-Aware Stage-Wise Transformer), a dual-branch Swin-based architecture that injects low-resolution contextual information into fine-grained high-resolution features through lightweight stage-wise cross-attention. To strengthen cross-scale learning, we also propose a SimMIM-style pretraining strategy based on masked reconstruction of the high-resolution image. Extensive experiments on the large-scale FLAIR-HUB aerial dataset demonstrate the effectiveness of CASWiT. Under our RGB-only UHR protocol, CASWiT reaches 66.37% mIoU with a SegFormer decoder, improving over strong RGB baselines while also improving boundary quality. On the URUR benchmark, CASWiT reaches 49.2% mIoU under the official evaluation protocol, and it also transfers effectively to medical UHR segmentation benchmarks. Code and pretrained models are available at* `https://huggingface.co/collections/heig-vd-geo/caswit`.

1. Introduction

Semantic segmentation of remote sensing imagery is central to many geospatial applications, including land-use mapping, environmental monitoring, and disaster response. As these applications increasingly rely on ultra-high-resolution (UHR) aerial imagery, methods must preserve fine local structures while also exploiting broader spatial context.

Transformer-based architectures [7, 11, 31, 47] have advanced state-of-the-art results in vision tasks, but their application to UHR inputs remains challenging because of quadratic complexity and GPU memory constraints. Common workarounds such as downsampling or tiling either reduce useful context or sacrifice spatial resolution, impairing segmentation at both object and scene levels. Recent UHR approaches therefore combine high-resolution patch processing with explicit context modeling via multi-branch or cross-scale designs [2, 22, 37].

Our approach. **CASWiT**, a dual-branch hierarchical transformer, is introduced for RGB-only UHR segmentation. One branch processes high-resolution (HR) crops to preserve boundaries and small objects, while a second branch ingests wider low-resolution (LR) patches to encode global context. The two streams interact at multiple encoder stages through compact global cross-attention blocks (HR queries over LR keys/values), enabling early context injection while remaining compute-efficient. To further improve cross-scale learning, we adapt SimMIM-style [52] masked image modeling to this dual-stream setting and pretrain on large amounts of unlabeled orthophotos.

Benchmarks. The primary evaluation is conducted on FLAIR-HUB [14] using an RGB-only UHR protocol that exploits its geospatial structure to reconstruct large contiguous tiles with preserved long-range context. Compared with URUR [23], FLAIR-HUB provides a larger-scale and more carefully curated benchmark for cross-scale learning (see § 4.1), while URUR is retained for continuity with prior work. We also report transfer results on medical image segmentation benchmarks to assess whether the proposed design extends beyond remote sensing.

Contributions.

- We introduce CASWiT (Context-Aware Stage-Wise Transformer), a dual-branch architecture that injects LR context into HR features through stage-wise cross-attention while preserving fine-grained HR detail.

- We design a dual-stream SimMIM pretraining strategy that strengthens cross-scale learning and transfers effectively to large-scale UHR segmentation tasks.
- We establish an RGB-only UHR evaluation protocol on FLAIR-HUB and show consistent improvements over prior RGB-only state of the art on FLAIR-HUB-RGB and URUR, with additional transfer results on medical segmentation benchmarks.

2. Related Work

Dual-stream UHR segmentation. Processing ultra-high-resolution (UHR) imagery for semantic segmentation requires preserving fine details while aggregating long-range context. A widely adopted strategy is dual-stream fusion, with an HR stream for local structures and an LR/context stream for scene-level semantics. GLNet [5] popularized this formulation with CNN backbones and late fusion by concatenation. Subsequent works refine this template through alternative fusion schemes, backbones, or efficiency-oriented designs, including WSDNet [23], FCtL [36], GPWFormer [22], SGNet [45], STUNet [20], and DESformer [29]. Other UHR methods target iterative patching with global guidance, proposal-based computation, shallow all-pixel processing, or boundary refinement [3, 8, 16, 21–24, 26, 36, 39, 49, 55]. Most of these approaches still rely on mid/late fusion, where HR and LR features interact only after substantial single-stream processing, although earlier interaction can be beneficial when the two inputs share similar representational spaces [1].

Single-stream HR backbones. An alternative to dual-stream fusion is to rely on hierarchical vision transformers or multi-scale CNNs that capture locality and globality within a single stream. Representative examples include Swin Transformer [31, 32] and PVT [46, 47], alongside UHR-oriented single-stream variants [17, 30, 38, 40, 48, 53]. Recent efficient transformer variants also aim to mitigate the quadratic cost of self-attention through architectural approximations or more scalable attention patterns [9, 44, 54]. While these models improve scalability compared to vanilla ViT, balancing spatial resolution and memory for truly UHR inputs remains challenging.

Fusion mechanisms and module placement. Despite the intuitive complementarity of HR and LR signals, many dual-stream models still rely on concatenation or summation at mid or late stages. Attention-based alignment between heterogeneous resolutions is less explored across multiple depths, even though cross-attention provides a systematic way to condition HR features on LR context early in the hierarchy. DESformer [29] introduces a multi-depth feature interaction module inside the encoder, while CTCFNet [33] combines CNN-Transformer features through a mid-to-late aggregation module and a bi-directional decoder. More recently, Boosting Dual-Stream [37] revisits the dual-stream paradigm with uncertainty-guided HR/LR interaction. Compared with these works, CASWiT uses two hierarchical transformer branches and lightweight stage-wise cross-attention from the earliest encoder stages, and further couples this design with SimMIM-style pretraining [52].

Other UHR dense prediction tasks. Similar mechanisms have also been explored in other dense UHR tasks such as salient object detection or monocular depth estimation, where dual-stream designs, patch selection, and uncertainty estimation remain important [16, 27, 35, 42, 43, 49]. More broadly, the need to combine fine structures with wider context also arises beyond remote sensing, including medical image segmentation, which motivates our transfer experiments.

3. Method

CASWiT (Context-Aware Stage-Wise Transformer) is a dual-branch architecture that fuses high-resolution (HR) features with low-resolution (LR) contextual features through compact cross-attention blocks inserted after each encoder stage (Fig. 1). Each block applies HR-LR cross-attention followed by a residual MLP; we also consider an optional learned gate, but use the ungated variant by default. The network is trained with supervision on the HR output and an auxiliary LR loss weighted by α. At inference, the LR stream remains to provide context, but its decoder/head is removed.

3.1. Overview

CASWiT combines a HR Swin encoder [31] that preserves high resolution features with a LR Swin encoder that captures global contextual features from a larger field of view. Both encoders share identical hierarchical configurations (stages $\{1..4\}$, channel schedule C_s). Cross-attention modules are inserted after each stage to inject LR context into the HR stream. A UPerNet [50] or SegFormer [51] head processes the HR features to produce the final logits.

3.2. Dual-resolution encoder

Inputs. Given an HR crop $I^{\mathrm{HR}} \in \mathbb{R}^{H\times W\times 3}$ and a co-registered LR image I^{LR} (downsampled from a larger FoV), the two Swin encoders produce stage-wise feature maps:

$$X_s^{\mathrm{HR}} \in \mathbb{R}^{H_s\times W_s\times C_s}, \qquad X_s^{\mathrm{LR}} \in \mathbb{R}^{\hat{H}_s\times \hat{W}_s\times C_s}.$$

The HR and LR features may differ in spatial size; they are flattened into token sequences before fusion.

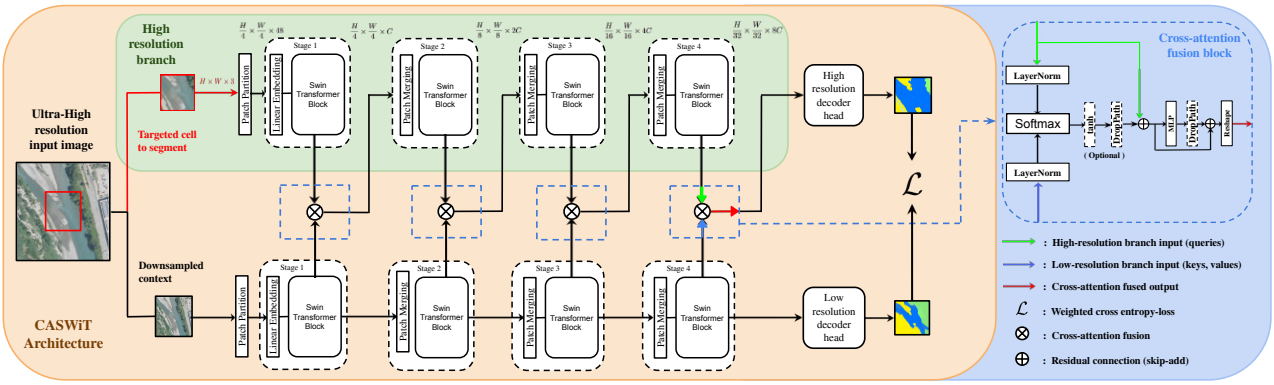


Figure 1. **CASWiT architecture.** A dual-branch encoder for ultra-high-resolution imagery: the HR branch processes the target tile, while the LR branch encodes a downsampled larger context. At each Swin stage (1→4), HR features provide queries and LR features provide keys/values to a cross-attention module with residual fusion and an optional gate (right). HR and LR decoder heads are jointly optimized during training with a weighted cross-entropy loss.

Cross-attention fusion block. At each stage s, we perform multi-head cross-attention (MHA) from HR queries to LR keys/values:

$$Q = \mathrm{LN}(X_s^{\mathrm{HR}})W_Q$$

$$K = \mathrm{LN}(X_s^{\mathrm{LR}})W_K, \quad V = \mathrm{LN}(X_s^{\mathrm{LR}})W_V,$$

$$A_s = \mathrm{MHA}\big(Q,\ K,\ V\big).$$

The final HR features at stage s, $\tilde{H}_s$, are obtained via a residual connection and an optional learned gate γ_s:

$$H'_s = X_s^{\mathrm{HR}} + \gamma_s \odot A_s, \qquad \tilde{H}_s = H'_s + \mathrm{MLP}(H'_s),$$

where $\gamma_s = \tanh(g_s)$ is a learned scalar stage-wise gate broadcast over HR tokens. The gate controls how much contextual information from the LR stream is injected into HR features. In practice, the ungated variant slightly outperforms the gated one, so we use the ungated version by default and report both settings in § 4.

3.3. Decoder and prediction heads

We consider two decoder variants for the HR branch: UPerNet [50] and SegFormer [51]. In the default setting, we adopt UPerNet, where stage features $\{\tilde{X}_s^{\mathrm{HR}}\}_{s=1}^4$ are fused by a Feature Pyramid Network (FPN) with a Pyramid Pooling Module (PPM) and then classified by the head. We also evaluate a SegFormer decoder on top of the same HR features. A LR decoder mirrors the same structure and is used only during training for auxiliary supervision; it can be removed at inference.

3.4. Supervised objectives

Let $\hat{Y}^{\mathrm{HR}} \in \mathbb{R}^{H\times W\times K}$ and $\hat{Y}^{\mathrm{LR}} \in \mathbb{R}^{\hat{H}\times\hat{W}\times K}$ be the logits from the HR and LR heads, and let $Y \in \{1,\dots,K\}^{H\times W}$ be the ground-truth labels. We compute standard pixel-wise cross-entropy on HR:

$$\mathcal{L}_{\mathrm{HR}} = -\frac{1}{HW}\sum_p \sum_k \mathbf{1}[Y_p{=}k]\ \log \mathrm{Softmax}(\hat{Y}_p^{\mathrm{HR}})_k.$$

For LR supervision, we use the downsampled label map $Y^{\downarrow} \in \{1,\dots,K\}^{\hat{H}\times\hat{W}}$ (nearest-neighbor):

$$\mathcal{L}_{\mathrm{LR}} = -\frac{1}{\hat{H}\hat{W}}\sum_p \sum_k \mathbf{1}[Y_p^{\downarrow}{=}k]\ \log \mathrm{Softmax}(\hat{Y}_p^{\mathrm{LR}})_k.$$

The total loss is the weighted sum

$$\mathcal{L} = \mathcal{L}_{\mathrm{HR}} + \alpha\,\mathcal{L}_{\mathrm{LR}},$$

where α controls the contribution of the LR auxiliary head (set to 0.5 in our experiments).

3.5. Self-supervised pretraining (SimMIM-style)

We adapt a simple framework for masked image modeling (SimMIM) [52] to the dual-stream encoder and keep the HR-LR fusion active throughout pretraining.

Masking strategy. On the HR stream, we apply random masking with ratio r_{HR} (default 0.75). On the LR stream, we apply centered masking with ratio r_{LR} (default 0.5), masking the LR region that is spatially aligned with the HR crop. This preserves the surrounding global layout while preventing trivial cross-scale copying from the LR center to the HR target region. In both cases, the masked tokens are replaced with a learnable mask token at the stage 1 embedding dimension (with no zeroing), as in many frameworks [19, 52]. Cross-attention therefore operates on LR features whose masked positions carry the learned mask token embedding.

Reconstruction head and objective. Only the HR branch is reconstructed. From the last HR stage (s=4), we use a 1×1 convolution producing $3s^2$ channels followed by a PixelShuffle with stride s (equal to the total downsampling factor of the HR encoder) to map tokens back to RGB at input resolution. Let $\hat{I}^{\mathrm{HR}}$ be the reconstruction. We minimize a masked ℓ_1 loss over the masked HR pixels only:

$$\mathcal{L}_{\mathrm{SSL}} = \frac{1}{3\,|M_{\mathrm{HR}}^{\mathrm{pix}}|} \sum_{p \in M_{\mathrm{HR}}^{\mathrm{pix}}} \left\| \hat{I}_p^{\mathrm{HR}} - I_p^{\mathrm{HR}} \right\|_1$$

where $M_{\mathrm{HR}}^{\mathrm{pix}}$ is obtained by upsampling the HR *patch* mask to pixel resolution using the stage-1 patch size. During SSL, masked tokens are replaced by a learnable mask embedding; fusion remains active so the encoder can leverage LR semantics to infer missing HR content. Fig. 2 illustrates this dual-stream masking strategy, with random HR masking and centered LR masking during self-supervised pretraining. After SSL, we discard the reconstruction head and fine-tune the dual-stream encoders with cross-attention under the supervised objective in § 3.4.

4. Experiments

4.1. Datasets

FLAIR-HUB. We use the FLAIR-HUB dataset [14], a large-scale multimodal extension of FLAIR [13], comprising 241,100 RGB patches of size 512×512 at 0.20 m GSD, annotated into 15 classes. To enable RGB-only UHR evaluation while remaining comparable to the official per-patch setting, we construct for each HR patch a geospatially aligned 3×3 context tile using its eight neighbors (GeoTIFF coordinates), yielding a 1024×1024 composite that we downsample by 2 to obtain a 512×512 LR input co-registered with the HR patch. When neighbors are missing at borders, we fill the gaps with black padding to keep the same dimensions for all patches. This protocol preserves long-range spatial context while keeping the input size compatible with standard backbones (see Fig. 3).

URUR. The URUR dataset [23] contains 3,008 UHR RGB images of size 5120×5120 from 63 cities with 8 land-cover classes. We follow the official split: 2,157 train, 280 val, 571 test. URUR has been influential for UHR evaluation, however, we observed occasional image-mask inconsistencies (e.g., local misalignment) that can affect evaluation metrics. We therefore report URUR results, but we advise interpreting URUR metrics with care: scores can be underestimated or display higher variance due to occasional image-mask non-conformities (illustrative examples are provided in the supplementary, Fig. 6), and, as discussed in [37], the handling of the other class may depress IoU (near zero) because it appears sparsely.

SWISSIMAGE (unlabeled, for SSL). For self-supervised pretraining, we use large-scale unlabeled orthophotos from the SWISSIMAGE archive at 0.20 m GSD (total ∼1067 Gpx; excluded from supervised splits). This corpus provides over 40× more pixels than labeled training data from official test split of FLAIR-HUB, enabling robust masked reconstruction pretraining.

Other UHR benchmarks. For completeness, we note that smaller remote sensing benchmarks such as INRIA Aerial [34] and DeepGlobe [10] are also commonly reported, but we focus on FLAIR-HUB and URUR because they provide a stronger setting for large-scale multi-class cross-scale evaluation. To assess transfer beyond aerial imagery, we additionally evaluate CASWiT on two medical UHR benchmarks, ISIC [41] and CRAG [15].

4.2. Evaluation Protocols

We report standard semantic segmentation metrics, including mean Intersection-over-Union (mIoU) and mean F1 score (mF1), computed over all non-void classes. For FLAIR-HUB, we follow the official split named "split_flairhub" and report per-class results in the supplementary. There is no geographic overlap between train/val/test splits. To avoid patch-boundary bias when using geospatially reconstructed neighborhoods, while remaining directly comparable to the original patch-based FLAIR-HUB protocol, evaluation is performed only on the center crop corresponding to the original HR patch. For URUR, we follow the original train/val/test protocol with the configuration of 8 classes (with class "other" for comparison with previous work) and also report per-class results in the supplementary. For ISIC and CRAG, we use the official splits and report mIoU on the test sets. Inferences are performed without overlapping sliding windows on FLAIR-HUB, URUR, ISIC and CRAG.

Mean Boundary IoU (mBIoU). Beyond region overlap, we evaluate boundary quality using the mean Boundary IoU metric (mBIoU) [6]. For each class c, we extract thin boundary bands from the ground truth (B_Y^c) and the prediction ($B_{\hat{Y}}^c$) by dilating their contours.

The boundary IoU for class c and mBIoU are:

$$\mathrm{bIoU}(c) = \frac{|B_Y^c \cap B_{\hat{Y}}^c|}{|B_Y^c \cup B_{\hat{Y}}^c|}, \quad \mathrm{mBIoU} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{bIoU}(c)$$

Compared to standard mIoU, mBIoU is insensitive to large homogeneous regions and focuses on how well object edges are localized.

4.3. Implementation Details

All experiments are implemented in PyTorch and trained on 4× NVIDIA L40S GPUs (48 GB each) using Distributed

Figure 2. Self-supervised inference results on the CASWiT architecture. Each image (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 after SimMIM-style pretraining.

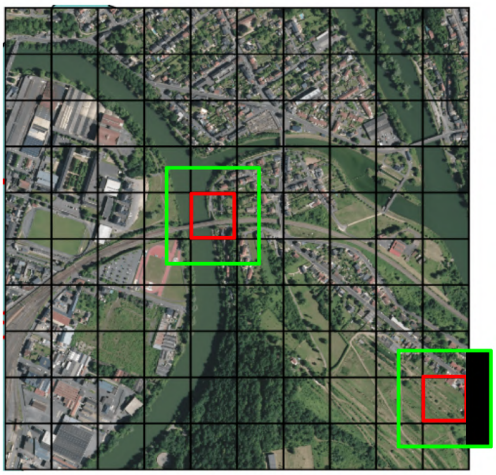


Figure 3. HR/LR construction on FLAIR-HUB. Red: original HR patch (512×512). Green: georeferenced 3×3 neighborhood assembled into a 1024×1024 context, then downsampled ×2 to form the LR input (512×512).

Data Parallel (DDP). We use the AdamW optimizer with an initial learning rate of 6×10^{-5}, decayed to 1×10^{-6} through a cosine annealing scheduler, and a weight decay of 0.01. Batch size is set to 20 (5 per GPU) for URUR and 16 (4 per GPU) for FLAIR-HUB. Training runs for 20 epochs with a crop size of 512×512 for both HR and LR inputs (LR initially 1024×1024 and subsampled to 512×512). No data augmentation is applied, unless specified, to ensure a controlled comparison across methods and datasets.

For all experiments, both HR and LR branches use identical backbones, CASWiT-Base means the use of 2 Swin-B backbones. Unless otherwise specified, the gating mechanism is disabled (see ablation in § 4.5), and the auxiliary LR supervision weight is set to $\alpha = 0.5$.

For capacity-controlled comparisons, we additionally evaluate a late-fusion dual-branch baseline with the same HR/LR inputs (using a simple sum). On the same dual-stream backbone, we report results for both decoder variants, UPerNet and SegFormer, and provide GFLOPs and inference speed (FPS) on FLAIR-HUB inputs.

Self-supervised pretraining. We perform SimMIM-style pretraining on the unlabeled SWISSIMAGE corpus for 100 epochs before fine-tuning. Masking ratios are set to 75% for HR (random) and 50% for LR (centered), the latter designed to maintain global layout while preventing trivial pixel copying across scales. Both streams and cross-attention blocks are optimized jointly during pretraining. The pretrained weights are then used for direct fine-tuning on FLAIR-HUB, URUR, ISIC and CRAG without intermediate adaptation.

4.4. Quantitative Results

FLAIR-HUB (RGB-only UHR protocol). Table 1 reports RGB-only segmentation performance on FLAIR-HUB under the proposed UHR protocol. Compared with the four official Swin+UPerNet RGB baselines released by the dataset authors [14], CASWiT consistently improves performance: CASWiT-Base + UPerNet reaches 65.11% mIoU, which further increases to 65.35% with self-supervised pretraining and to 65.83% with additional spatial and radiometric augmentations. Replacing UPerNet with SegFormer on the same CASWiT backbone yields the best result, **66.37%** mIoU and **78.58%** mF1, while also reducing compute from 489 to 298 GFLOPs and increasing inference speed from 15.4 to 17.9 FPS.

For reference, the best multimodal configuration reported in [25] reaches 65.9% mIoU, meaning that our best RGB-only variant surpasses this score while relying solely on RGB imagery. CASWiT also improves boundary quality, with mBIoU increasing from 32.57 for the retrained Swin-B + UPerNet baseline to 35.87 for CASWiT-Base + UPerNet and 36.90 for CASWiT-Base-SSL-aug + UPerNet.

Model	mIoU	mF1	mBIoU	GFLOPs	FPS
Official RGB baselines					
Swin-T + UPer [14]	62.01	75.27	-	237	69.2
Swin-S + UPer [14]	61.87	75.11	-	261	41.5
Swin-B + UPer [14]	64.05	76.88	-	306	36.2
↪ retrained	64.02	76.64	32.57	306	36.2
Swin-L + UPer [14]	63.36	76.35	-	420	27.8
Capacity-controlled					
ISDNet [16] (retrained)	52.77	-	-	-	-
Dual Swin-B (late fusion)	64.25	-	-	398	19.4
CASWiT (Ours)					
Base + UPer	65.11	77.71	35.87	489	15.4
Base-SSL + UPer	65.35	77.87	35.99	489	15.4
Base-SSL-aug + UPer	65.83	78.22	**36.90**	489	15.4
Base-SSL-aug + SegF	**66.37**	**78.58**	36.51	**298**	**17.9**

Table 1. FLAIR-HUB test results under the RGB-only UHR protocol. CASWiT improves over the official RGB baselines and a capacity-controlled late-fusion variant. The SegFormer head yields the best mIoU and mF1 while reducing GFLOPs. UPer = UPerNet, SegF = SegFormer.

To isolate the effect of stage-wise cross-attention from model capacity, we additionally compare against a dual-branch late-fusion baseline with the same HR/LR inputs and decoder, as well as against a larger single-stream Swin-L baseline. The late-fusion variant reaches 64.25 mIoU, remaining below CASWiT despite using the same dual-stream setting. As an additional reproducible reference, re-trained ISDNet reaches 52.77 mIoU on FLAIR-HUB-UHR, remaining well below all CASWiT variants.

URUR (legacy benchmark). Table 2 reports results on URUR. CASWiT improves over prior UHR-specific architectures such as WSDNet [23] and the recent Boosting Dual-Stream model [37]. It reaches 48.7% mIoU with CASWiT-Base + UPer, 49.1% with CASWiT-Base-SSL-aug + UPer, and 49.2% with CASWiT-Base-SSL-aug + SegF. These results support the benefit of stage-wise cross-attention for combining fine detail and large-scale context in UHR segmentation. As discussed in § 4.1, occasional annotation inconsistencies and the handling of the *other* class can lead to underestimated mIoU on URUR.

Generalization beyond remote sensing: ISIC and CRAG. To assess whether CASWiT transfers beyond aerial imagery, we evaluate it on two medical UHR segmentation benchmarks: ISIC [41] and CRAG [15]. Table 3 shows that CASWiT generalizes well without architectural modification, outperforming recent dual-branch UHR baselines on ISIC and matching or improving upon them on CRAG. Using a SegFormer decoder further improves performance, reaching 86.5 mIoU on ISIC and 90.7 on CRAG. These results indicate that the proposed stage-wise context injection is not limited to aerial imagery and transfers effectively to other high-resolution segmentation domains.

Model	mIoU	Mem (MB)
Generic baselines		
PSPNet [8]	32.0	5482
ResNet18 [18] + DeepLabv3+ [4]	33.1	5508
STDC [12]	42.0	7617
UHR methods		
GLNet [5]	41.2	3063
FCtL [36]	43.1	4508
ISDNet [16]	45.8	4920
WSDNet [23]	46.9	4510
Boosting Dual-Stream [37]	48.2	3682
CASWiT-Base + UPer	48.7	2996
CASWiT-Base-SSL-aug + UPer	49.1	2996
CASWiT-Base-SSL-aug + SegF	**49.2**	2878

Table 2. URUR test results. CASWiT improves over prior UHR-specific methods while remaining memory-efficient. UPer = UPerNet, SegF = SegFormer.

Model	ISIC (mIoU)	CRAG (mIoU)
GPWFormer [22]	80.7	89.9
Boosting Dual-Stream [37]	83.4	90.3
CASWiT-Base-SSL-aug + UPer	85.4	90.3
CASWiT-Base-SSL-aug + SegF	**86.5**	**90.7**

Table 3. Results on the ISIC and CRAG test sets. CASWiT transfers effectively beyond remote sensing and benefits further from the SegFormer head. UPer = UPerNet, SegF = SegFormer.

4.5. Ablation Studies

We perform a series of ablation experiments on the FLAIR-HUB validation split to analyze the impact of the key design choices in CASWiT. Unless otherwise stated, all variants use a Swin-B backbone and are trained under identical conditions (20 epochs, crop size 512×512, no data augmentation). We systematically vary the cross-attention pattern, the auxiliary LR supervision weight α, the gating mechanism, and the SSL initialization. Results are summarized in Table 4.

Cross-attention. To isolate the effect of cross-scale fusion, we remove the LR/context branch and disable all cross-attention modules. This effectively reduces CASWiT to a standard Swin-B encoder with a UPerNet decoder, i.e., the RGB baseline used in the FLAIR-HUB paper. On the validation set, this single-stream baseline reaches 70.11 mIoU. When enabling all-stage cross-attention without LR supervision (α=0), performance increases to 70.30 mIoU, and further rises to 71.40 mIoU once auxiliary LR supervision is added (α=0.5). This corresponds to a gain of +1.29 mIoU over the re-trained Swin-B baseline, indicating that the improvement comes from explicit context-aware fusion. We also verified that our Swin-B reproduction closely matches the official FLAIR-HUB RGB results on the test set (see § 4.4), which validates our implementation and training setup.

Stage-wise fusion. We compare cross-attention applied only at the first stage, only at the last stage, and at all four encoder stages. Using cross-attention exclusively at the last stage (Stage-4 only) already provides a strong improvement over the single-stream baseline (71.32 vs. 70.11 mIoU), confirming that injecting LR context at a high semantic level is beneficial. Relying on the first stage alone is less effective (69.89 mIoU), suggesting that early low-level interaction is not sufficient by itself. Our full CASWiT-Base model, which performs stage-wise fusion at all levels, achieves the best overall result (71.40 mIoU), indicating that combining early and late cross-scale interactions yields the most balanced trade-off between fine detail and global coherence.

Auxiliary LR supervision. We vary the auxiliary loss weight α from 0 to 0.5 in Table 4, and additionally observe that performance remains robust for moderate values around $\alpha = 0.5$. In particular, values in the range $\alpha \in [0.3, 0.7]$ yield comparable validation performance (70.7–71.3 mIoU), whereas more extreme settings reduce performance (below 68.6 mIoU for $\alpha = 0.1$ and $\alpha = 0.9$). Comparing the all-stage variants, adding LR supervision (α=0.5) improves mIoU from 70.30 to 71.40 and also leads to smoother validation curves (not shown), suggesting that lightweight LR guidance regularizes the shared representation and facilitates optimization.

Gating mechanism. We evaluate the optional learned gate g_s used to scale the cross-attention residuals. With all-stage fusion and α=0, enabling gating yields 70.15 mIoU, slightly below the ungated counterpart (70.30 mIoU). We did not observe consistent benefits in terms of stability or final accuracy, and therefore keep gating disabled in the main configuration (equivalent to $\gamma_s = 1$).

Model size. We additionally evaluate a lighter variant, CASWiT-Tiny, which uses the same cross-attention design but a reduced-capacity backbone. CASWiT-Tiny reaches 70.91 mIoU, only 0.49 mIoU below CASWiT-Base despite its smaller parameter budget. This indicates that the proposed fusion strategy can still provide benefits in a more compact setting.

Self-supervised pretraining. Finally, we compare models trained from scratch to those initialized with our SimMIM-style pretraining on SWISSIMAGE. On the FLAIR-HUB validation set, pretraining increases performance from 71.40 to 71.55 mIoU (+0.15), and we observe larger gains on the test set (see § 4). Qualitatively, the pretrained model produces sharper boundaries and more coherent large structures, indicating that masked reconstruction on large-scale orthoimagery exposes the network to structural cues beneficial for UHR semantic segmentation. The centered LR masking is designed so that the model cannot rely on trivial cross-scale copying from the LR region spatially aligned with the HR crop, and must instead exploit the surrounding LR context, which better matches the downstream role of cross-attention.

Variant	Cross-Attn	α	mIoU (%)↑	mF1 (%)↑
Baseline (no CA)	None	0.0	70.11	81.72
All-stage + gating on	(All, gated)	0.0	70.15	81.78
All-stage	(All)	0.0	70.30	81.89
Stage-1 fusion only	(Stage 1)	0.5	69.89	81.56
Last-stage fusion only	(Stage 4)	0.5	71.32	82.56
CASWiT-Tiny	(All)	0.5	70.91	82.24
CASWiT-Base	(All)	0.5	71.40	82.62
CASWiT-Base-SSL	(All)	0.5	71.55	82.78

Table 4. Ablation study on the FLAIR-HUB validation set. Each component is varied independently; CA denotes cross-attention and α the LR supervision weight. All models use a Swin-B backbone, except CASWiT-Tiny, and are trained without augmentation.

4.6. Qualitative Analysis

We provide qualitative visualizations to further illustrate the behavior of CASWiT and the role of cross-scale context injection.

Segmentation results on FLAIR-HUB. Fig. 4 shows a representative example from the FLAIR-HUB test set comparing CASWiT-Base-SSL + UPerNet to the RGB Swin-B + UPerNet baseline (more examples are provided in the supplementary). Our model produces cleaner boundaries and better preserves fine structures, such as narrow roads and building outlines. It also reduces semantic bleeding between adjacent classes.

Cross-attention visualization. To understand how context propagates across scales, we visualize the attention maps from HR queries to LR keys at different encoder stages (Fig. 5 and additional examples in the supplementary). Early stages (stages 1-2) tend to concentrate on fine local structures and boundaries, whereas later stages (stages 3-4) attend more broadly to the semantic layout, such as roads, vegetation, and water. This progressive behavior suggests that multi-stage cross-attention integrates contextual information across different levels of the hierarchy.

Self-supervised reconstruction. Fig. 2 illustrates the masked reconstruction process during self-supervised pretraining. Random HR masks (75%) and centered LR masks (50%) are applied jointly; the network must reconstruct missing HR pixels using both visible HR content and surrounding LR context. CASWiT successfully recovers fine-grained textures and object geometry, indicating that the dual-stream fusion effectively learns cross-scale correspondences.

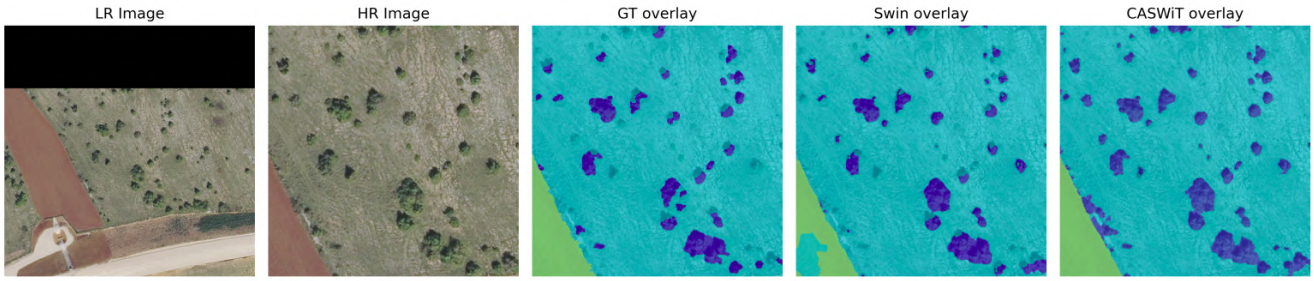


Figure 4. Qualitative comparison on FLAIR-HUB. From left to right: LR image (note the missing band at the top), HR crop, ground-truth overlay, RGB baseline (Swin-B + UPerNet) overlay, and CASWiT overlay. CASWiT better recovers small vegetation patches and yields sharper boundaries, while reducing false positives on bare soil and road areas. Despite the LR artifact, CASWiT remains stable.

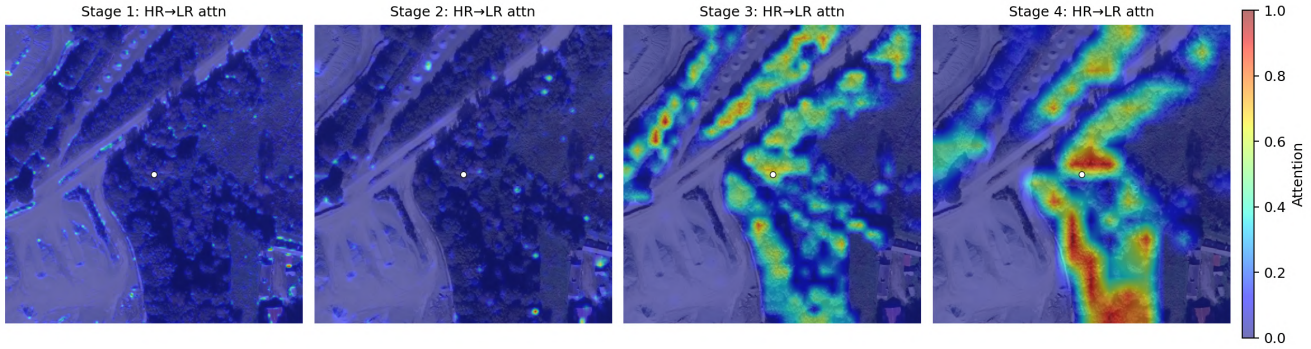


Figure 5. Cross-attention maps after self-supervised pretraining and supervised fine-tuning. Visualization of HR-to-LR cross-attention at each encoder stage of CASWiT. The queried HR pixel is marked by a white dot, and attention weights are reprojected onto the LR token grid and overlaid on the LR image.

5. Conclusion

5.1. Conclusion and Limitations

We introduced **CASWiT**, a cross-attentive dual-branch backbone for ultra-high-resolution RGB segmentation that fuses HR detail with LR context via lightweight, stage-wise cross-attention. We also proposed an RGB-only FLAIR-HUB-UHR evaluation protocol that leverages geospatial structure while remaining comparable to the original patch-based setting of FLAIR-HUB [14]. On this benchmark, CASWiT reaches 66.37 mIoU with a SegFormer decoder and improves boundary quality by +4.33 mBIoU over the retrained Swin-B + UPerNet baseline. This best RGB-only variant also exceeds the 65.9 mIoU reported on FLAIR-HUB by the multimodal Maestro configuration [25]. On URUR, CASWiT reaches 49.2 mIoU, improving over prior UHR-specific methods under the official protocol. CASWiT also transfers effectively to medical UHR segmentation benchmarks, reaching 86.5 mIoU on ISIC [41] and 90.7 mIoU on CRAG [15]. Overall, these results suggest that the gains stem primarily from explicit cross-scale fusion and are further reinforced by SimMIM-style pretraining on large-scale orthophotos. A current limitation is that CASWiT still relies on a dual-branch design, which increases architectural and computational complexity relative to simpler single-stream baselines.

5.2. Perspectives

Our core contribution is the backbone: CASWiT delivers stronger, context-aware features while remaining compatible with different decoder heads. A natural next step is to extend the evaluation of CASWiT to broader high-resolution vision settings. In particular, competitive results obtained with a CASWiT-based model in the NTIRE 2026 Remote Sensing Infrared Image Super-Resolution challenge suggest that improved context-aware feature extraction is also a promising direction for super-resolution [28].

6. Acknowledgements

We would like to thank Fabien Délèze for his careful proofreading, and the ESO team for their support. We also thank Shanci Li for his valuable assistance with the dataset, as well as Amir Zamir and his team for their constructive feedback and insightful discussions as part of the Visual Intelligence course. This research was supported by the Canton of Vaud and the INSIT Institute at HEIG-VD.

References

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. 2

[2] Chun-Fu (Richard) Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 357–366, 2021. 1

[3] Lijia Chen, Honghui Chen, Yanqiu Xie, Tianyou He, Jing Ye, and Yushan Zheng. An efficient and light transformer-based segmentation network for remote sensing images of landscapes. *Forests*, 14(11), 2023. 2

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 6

[5] Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning Qian. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 6

[6] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C. Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation, 2021. 4

[7] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation, 2022. 1

[8] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 6

[9] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers, 2022. 2

[10] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018. 4

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1

[12] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9716–9725, 2021. 6

[13] Anatol Garioud, Nicolas Gonthier, Loic Landrieu, Apolline De Wit, Marion Valette, Marc Poupée, Sebastien Giordano, and boris Wattrelos. Flair : a country-scale land cover semantic segmentation dataset from multi-source optical imagery. In *Advances in Neural Information Processing Systems*, pages 16456–16482. Curran Associates, Inc., 2023. 4

[14] Anatol Garioud, Sébastien Giordano, Nicolas David, and Nicolas Gonthier. Flair-hub: Large-scale multimodal dataset for land cover and crop mapping, 2025. 1, 4, 5, 6, 8, 3

[15] Simon Graham, Hao Chen, Jevgenij Gamper, Qi Dou, Pheng-Ann Heng, David Snead, Yee Wah Tsang, and Nasir Rajpoot. Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images. *Medical Image Analysis*, 52:199–211, 2019. 4, 6, 8

[16] Shaohua Guo, Liang Liu, Zhenye Gan, Yabiao Wang, Wuhao Zhang, Chengjie Wang, Guannan Jiang, Wei Zhang, Ran Yi, Lizhuang Ma, and Ke Xu. Isdnet: Integrating shallow and deep networks for efficient ultra-high resolution segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4361–4370, 2022. 2, 6

[17] Renlong Hang, Ping Yang, Feng Zhou, and Qingshan Liu. Multiscale progressive segmentation network for high-resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 3

[20] Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, and Yu Qiao. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training, 2023. 2

[21] Chuong Huynh, Anh Tuan Tran, Khoa Luu, and Minh Hoai. Progressive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16755–16764, 2021. 2

[22] Deyi Ji, Feng Zhao, and Hongtao Lu. Guided patch-grouping wavelet transformer with spatial congruence for ultra-high resolution segmentation, 2023. 1, 2, 6

[23] Deyi Ji, Feng Zhao, Hongtao Lu, Mingyuan Tao, and Jieping Ye. Ultra-high resolution segmentation with ultra-rich context: A novel benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23621–23630, 2023. 1, 2, 4, 6

[24] Yuyang Ji and Lianlei Shan. Ldnet: Semantic segmentation of high-resolution images via learnable patch proposal and dynamic refinement. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2024. 2

[25] Antoine Labatie, Michael Vaccaro, Nina Lardiere, Anatol Garioud, and Nicolas Gonthier. Maestro: Masked autoencoders for multimodal, multitemporal, and multispectral earth observation data, 2025. 5, 8

[26] Qi Li, Jiaxin Cai, Jiexin Luo, Yuanlong Yu, Jason Gu, Jia Pan, and Wenxi Liu. Memory-constrained semantic segmentation for ultra-high resolution uav imagery. *IEEE Robotics and Automation Letters*, 9(2):1708–1715, 2024. 2

[27] Hongyu Liu, Runmin Cong, Hua Li, Qianqian Xu, Qingming Huang, and Wei Zhang. ESNet: Evolution and succession network for high-resolution salient object detection. In *Forty-first International Conference on Machine Learning*, 2024. 2

[28] Kai Liu, Haoyang Yue, Zeli Lin, Zheng Chen, Jingkai Wang, Jue Gong, Radu Timofte, Yulun Zhang, et al. The First Challenge on Remote Sensing Infrared Image Super-Resolution at NTIRE 2026: Benchmark Results and Method Overview . In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. 8

[29] Wenshu Liu, Nan Cui, Luo Guo, Shihong Du, and Weiyin Wang. Desformer: A dual-branch encoding strategy for semantic segmentation of very-high-resolution remote sensing images based on feature interaction and multiscale context fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024. 2

[30] Yatong Liu, Yu Zhu, Ying Xin, Yanan Zhang, Dawei Yang, and Tao Xu. Mestrans: Multi-scale embedding spatial transformer for medical image segmentation. *Computer Methods and Programs in Biomedicine*, 233:107493, 2023. 2

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 1, 2

[32] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, 2022. 2

[33] Chen Lu, Xian Zhang, Kaile Du, Han Xu, and Guangcan Liu. Ctcfnet: Cnn-transformer complementary and fusion network for high-resolution remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024. 2

[34] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3226–3229, 2017. 4

[35] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H.S. Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24384–24394, 2023. 2

[36] Qi, Lin Xindai, Yang Weixiang, He Shengfeng, Yu Yuanlong Liu Wenxi, and Li. Ultra-high resolution image segmentation via locality-aware context fusion and alternating local enhancement. *International Journal of Computer Vision*, 132:5030–5047, 2024. 2, 6

[37] Rong Qin, Xingyu Liu, Jinglei Shi, Liang Lin, and Jufeng Yang. Boosting the dual-stream architecture in ultra-high resolution segmentation with resolution-biased uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25960–25970, 2025. 1, 2, 4, 6

[38] Yanzhou Su, Jian Cheng, Haiwei Bai, Haijun Liu, and Changtao He. Semantic segmentation of very-high-resolution remote sensing images via deep multi-feature learning. *Remote Sensing*, 14, 2022. 2

[39] Yihao Sun, Mingrui Wang, Xiaoyi Huang, Chengshu Xin, and Yinan Sun. Fast semantic segmentation of ultra-high-resolution remote sensing images via score map and fast transformer-based fusion. *Remote Sensing*, 16(17), 2024. 2

[40] Loic Themyr, Clément Rambour, Nicolas Thome, Toby Collins, and Alexandre Hostettler. Full contextual attention for multi-resolution transformers in semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3224–3233, 2023. 2

[41] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):180161, 2018. 4, 6, 8

[42] Matias Valdenegro-Toro. Sub-ensembles for fast uncertainty estimation in neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 4119–4127, 2023. 2

[43] Hongzhen Wang, Ying Wang, Qian Zhang, Shiming Xiang, and Chunhong Pan. Gated convolutional neural network for semantic segmentation in high-resolution images. *Remote Sensing*, 9(5), 2017. 2

[44] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020. 2

[45] Sai Wang, Yutian Lin, Yu Wu, and Bo Du. Toward real ultra image segmentation: Leveraging surrounding context to cultivate general segmentation model. In *Advances in Neural Information Processing Systems*, pages 129227–129249. Curran Associates, Inc., 2024. 2

[46] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 568–578, 2021. 2

[47] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 1, 2

[48] Honglin Wu, Peng Huang, Min Zhang, Wenlong Tang, and Xinyu Yu. Cmtfnet: Cnn and multiscale transformer fusion network for remote-sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–12, 2023. 2

[49] Tong Wu, Zhenzhen Lei, Bingqian Lin, Cuihua Li, Yanyun Qu, and Yuan Xie. Patch proposal network for fast semantic segmentation of high-resolution images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12402–12409, 2020. 2

[50] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding, 2018. 2, 3

[51] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, pages 12077–12090. Curran Associates, Inc., 2021. 2, 3

[52] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, 2022. 1, 2, 3

[53] Fei Yang, Fenlong Jiang, Jianzhao Li, and Lei Lu. Mstrans: Multi-scale transformer for building extraction from hr remote sensing images. *Electronics*, 13, 2024. 2

[54] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. In *Advances in Neural Information Processing Systems*, pages 4203–4217. Curran Associates, Inc., 2022. 2

[55] Zhan Zhang, Daoyu Shu, Guihe Gu, Wenkai Hu, Ru Wang, Xiaoling Chen, and Bingnan Yang. Ringformer-seg: A scalable and context-preserving vision transformer framework for semantic segmentation of ultra-high-resolution remote sensing imagery. *Remote Sensing*, 17:3064, 2025. 2

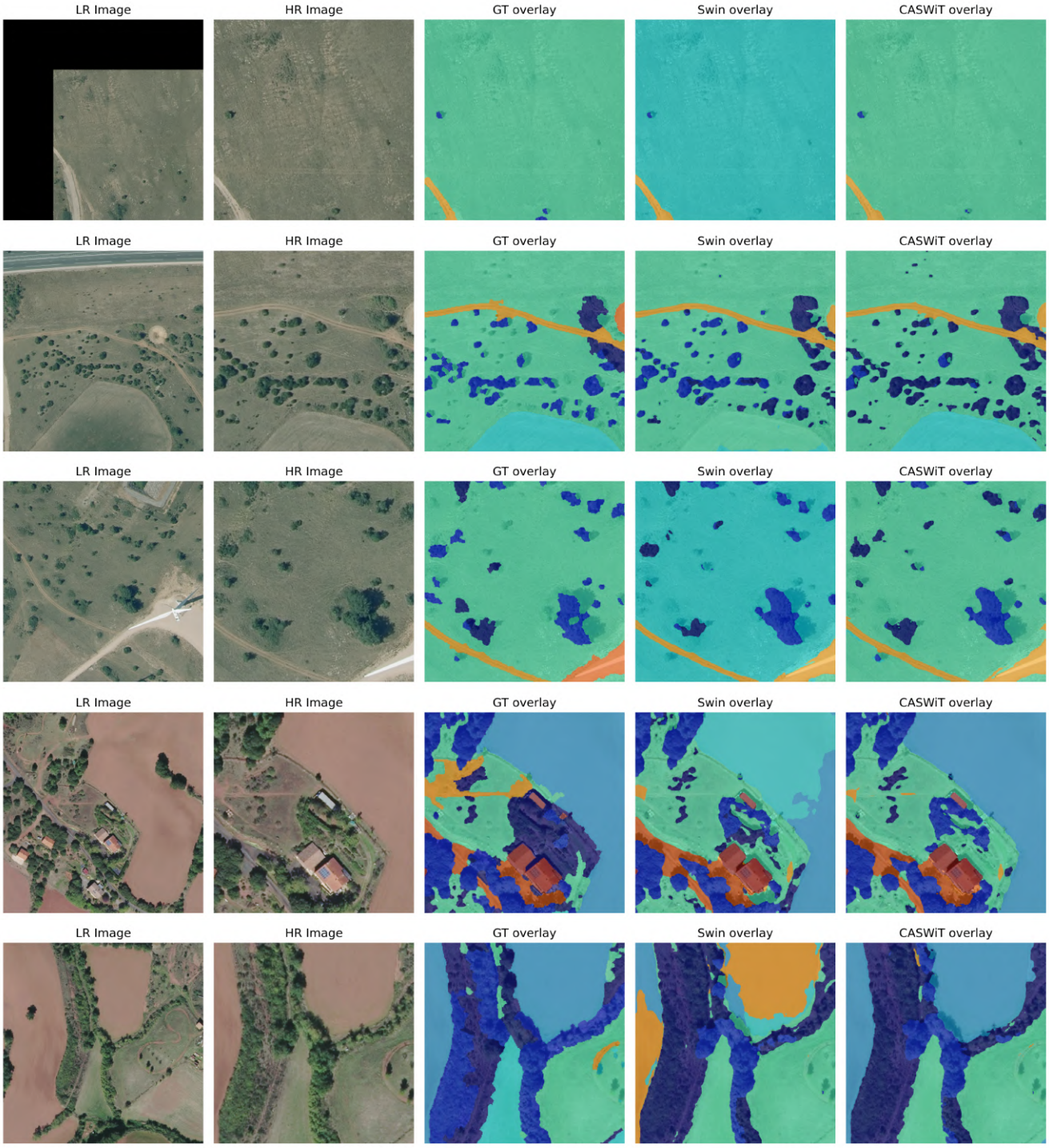
Context-Aware Semantic Segmentation via Stage-Wise Attention

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



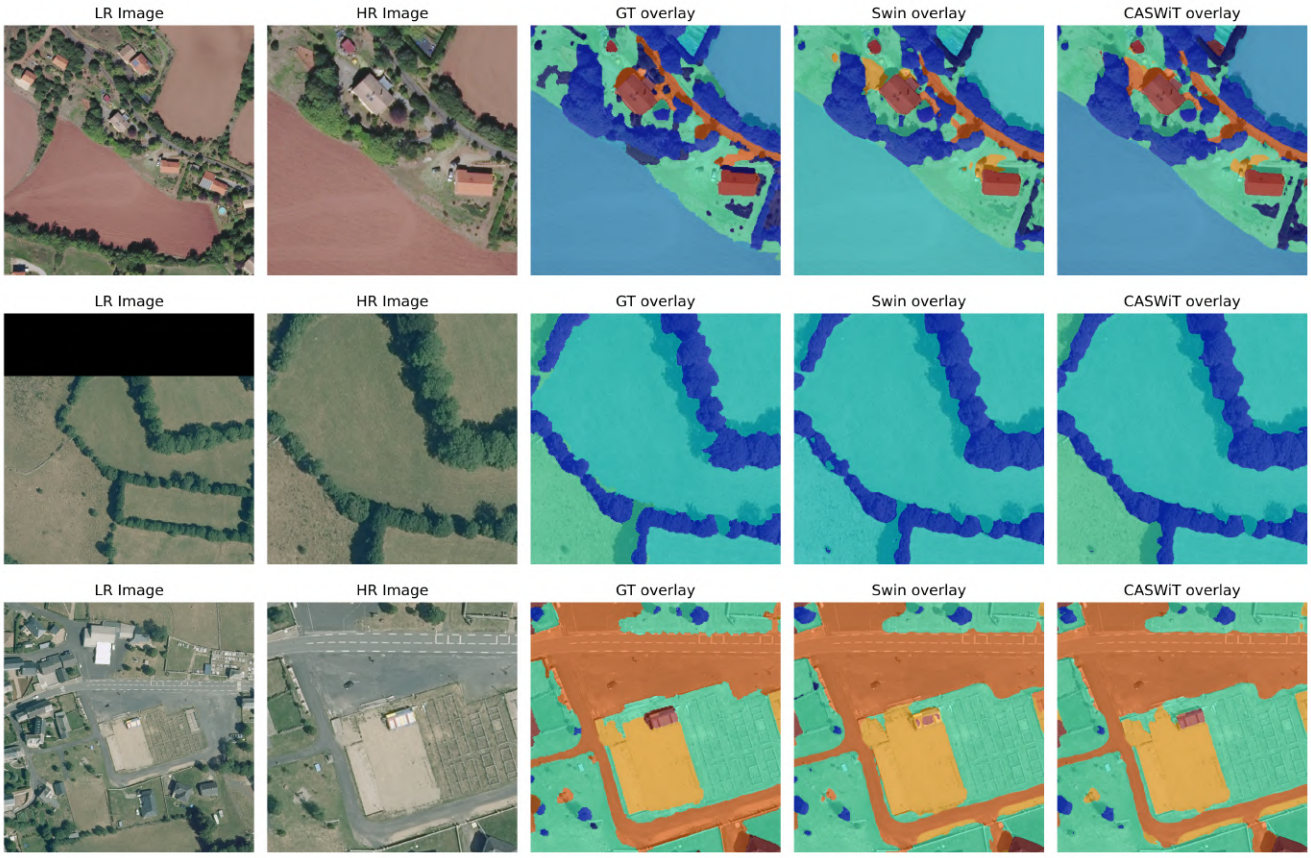


Figure 7. Comparison of LR/HR images, ground truth overlays, Swin Base predictions, and CASWiT predictions on eight test patches.

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.

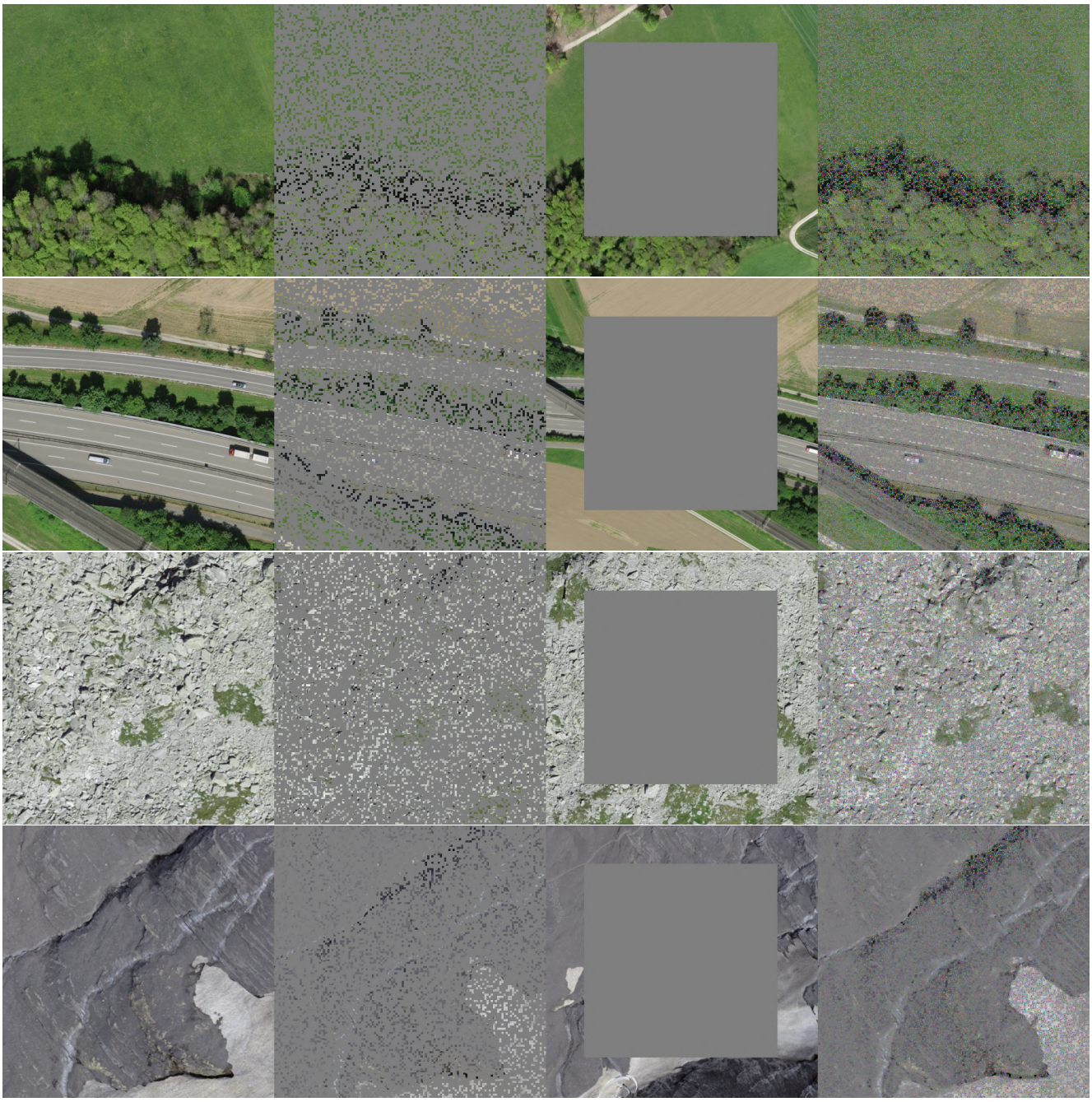
Class	WSDNET [23]	Boosting Dual-Stream [37]	CASWiT-B-SSL-aug + UPer	CASWiT-B-SSL-aug + SegF
Others	-	-	0.00	0.00
Building	-	-	75.07	74.42
Farmland	-	-	79.19	79.31
Greenhouse	-	-	46.51	46.50
Woodland	-	-	52.10	51.79
Bareland	-	-	31.64	32.31
Water	-	-	54.90	55.86
Road	-	-	53.33	53.67
mIoU	46.9	48.2	49.1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



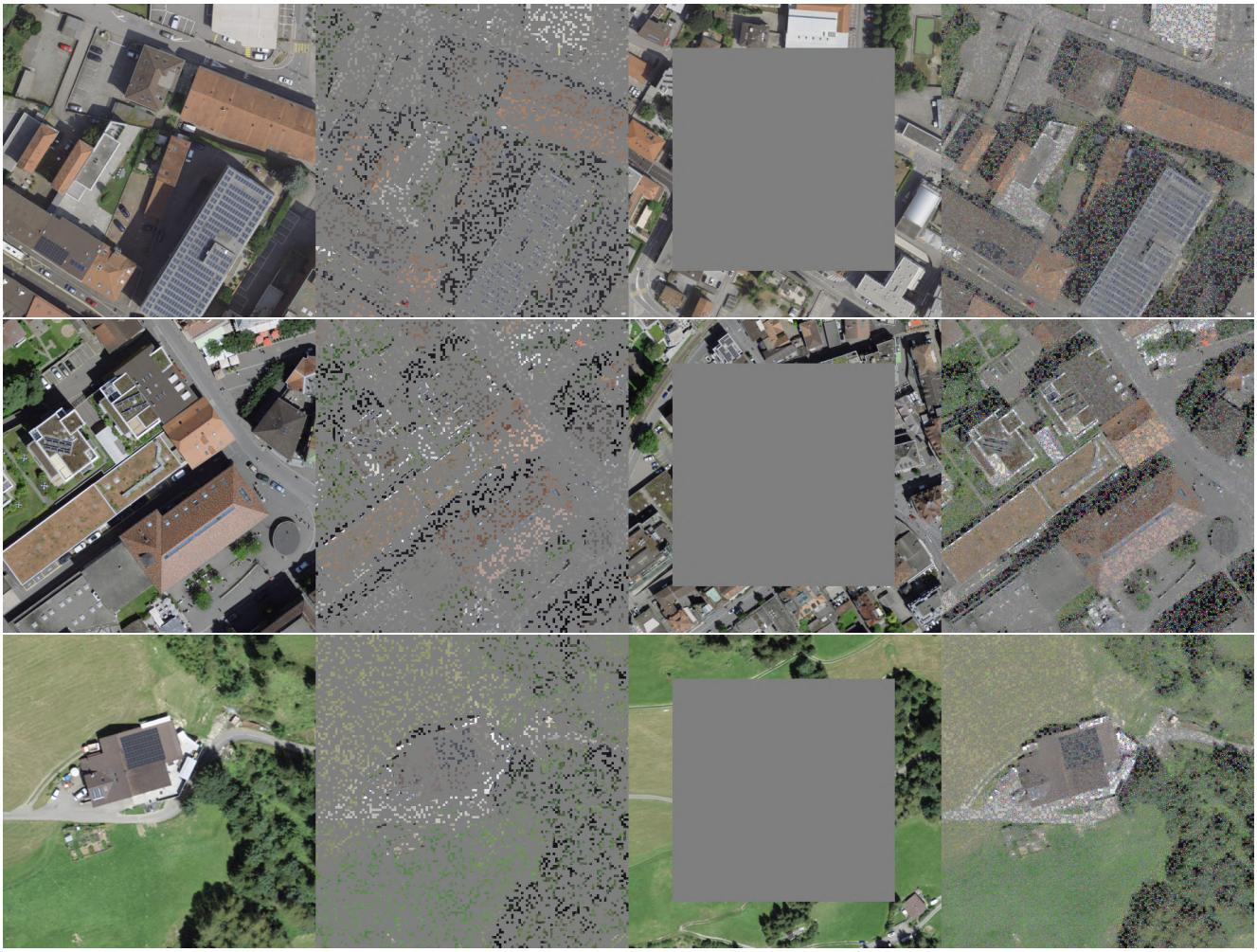


Figure 9. Self-supervised SimMIM-style inference results on the CASWiT-Base architecture. Each row (left to right) shows: original high-resolution image, high-resolution image with random masking, low-resolution image with central masking, and the reconstruction of the high-resolution image.

12. Cross-attention visualization

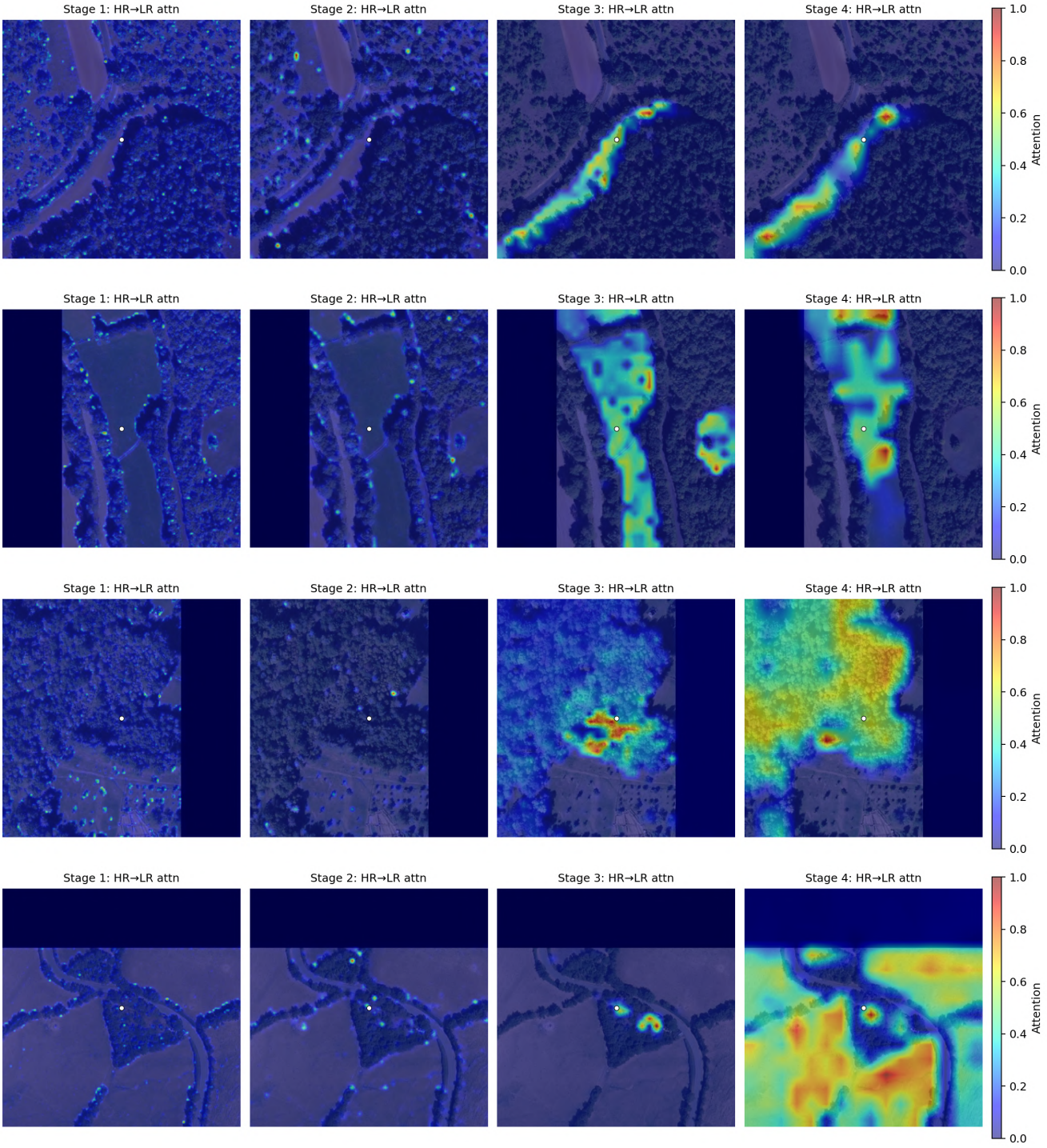


Figure 10. Visualization of cross-attention maps for each stage of the model on four test patches.