

Hierarchical Adaptive Feature Refinement Network for VHR Remote Sensing Image Segmentation

Shuaishuai Cao, Meng Tang, Shuwei Peng, Xuan Liu, Min Huang, Jie Chen, Jiacheng Niu, Yong Chen, Edore Akpokodje, and Hui Lin, *Senior Member, IEEE*

Abstract—Semantic segmentation of very-high-resolution (VHR) remote sensing imagery increasingly benefits from strong pretrained hierarchical encoders, yet exploiting their multi-stage representations remains difficult. Nearby regions demand different balances between fine detail and semantic context, aggressive task-specific transformations perturb useful pretrained features, and conventional semantic supervision provides limited structural guidance. We present HAFR-Net, a progressive refinement framework that adaptively organizes and conservatively refines hierarchical representations instead of replacing them with a monolithic decoder transformation. Heterogeneity-Guided Stage-Adaptive Fusion (HG-SAF) predicts dense stage weights conditioned on local feature variation. A Frequency-Residual Adapter (FRA) then injects frequency information through a bounded, zero-initialized residual branch that keeps the fused representation as its reference. A Confusion-Aware Tri-Prior Decoder (CATP) finally regularizes the prediction with boundary, objectness, and training-derived class-relation cues. Under a matched Swin-B training and single-scale inference protocol, HAFR-Net attains 84.12%, 87.86%, 55.17%, and 67.70% mIoU on ISPRS Vaihingen, ISPRS Potsdam, LoveDA, and OpenEarthMap, improving the matched UPerNet baseline by 0.55, 0.95, 1.55, and 1.84 percentage points. Controlled analyses further show consistent spatial reweighting beyond content-only routing, improved boundary and thin-structure accuracy over matched spatial and spectral alternatives, and reduced confusion on pre-declared class pairs. The trained LoveDA model is released, with the source code to follow, at <https://github.com/anticipate218/HAFRNet>.

Index Terms—Frequency-domain refinement, hierarchical feature fusion, semantic segmentation, confusion-aware decoding, very-high-resolution (VHR) remote sensing.

I. INTRODUCTION

VERY-high-resolution (VHR) remote sensing imagery provides detailed observations of buildings, roads, vehicles, vegetation, and other land-cover elements at fine spatial scales [1]–[3]. Modern segmentation systems based on CNNs, Transformers, and state-space models have benefited substantially from large pretrained encoders and hierarchical feature representations [4]–[11]. As encoder capacity continues to improve, however, a growing part of the segmentation problem shifts from representation extraction to representation utilization: the decoder must determine how to combine spatially detailed shallow features with semantically strong deep features, and how to refine them without unnecessarily destroying information already encoded by pretraining.

A key difficulty in VHR imagery is that the preferred balance between spatial detail and semantic abstraction varies

strongly within the same image (Fig. 1). Small vehicles, narrow roads, roof edges, and object boundaries benefit from high-resolution stages, whereas large roofs, agricultural parcels, and homogeneous regions often require deeper contextual representations. Nevertheless, many commonly used decoders aggregate multi-stage features through fixed top-down connections, predefined summation, or image-level weighting [12]–[15]. Such strategies are effective in general, but they provide limited flexibility when the desired stage contribution changes from pixel to pixel within a highly heterogeneous scene.

A second consideration concerns how additional decoder capacity is introduced. Frequency-domain and global-context operators can enhance detail or long-range interactions [16]–[20], but unrestricted transformations may require the optimizer to simultaneously preserve pretrained representations and learn a new task-specific mapping. For VHR datasets, which are often considerably smaller than generic pretraining corpora, a more conservative strategy is attractive: rather than replacing the fused representation, the decoder should introduce bounded residual corrections that remain close to a known reference behavior at the beginning of fine-tuning.

Finally, accurate VHR segmentation is not determined by region classification alone. Errors frequently occur around object boundaries, small foreground regions, and semantically similar categories [21]–[24]. These patterns motivate the use of structural supervision and class-relation constraints during decoding. Rather than treating such auxiliary signals as independent tasks, they can be used as lightweight regularizers that shape the final feature space while leaving inference dominated by a single semantic prediction branch. When class-relation pairs are used, they are identified exclusively from a training-only pilot split and are fixed before final evaluation.

Based on these observations, we propose HAFR-Net, a Hierarchical Adaptive Feature Refinement Network for VHR remote sensing image segmentation. The framework follows a progressive refinement philosophy. A pretrained Swin-B encoder first provides four hierarchical feature stages. Heterogeneity-Guided Stage-Adaptive Fusion (HG-SAF) adaptively aggregates these features with pixel-wise weights conditioned on local heterogeneity. Frequency-Residual Adapter (FRA) then performs a bounded residual correction in the frequency domain instead of replacing the fused representation with a full spectral transform. Finally, the Confusion-Aware Tri-Prior Decoder (CATP) regularizes semantic decoding using structural and class-relation cues. The resulting decoder explicitly separates adaptive hierarchical

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

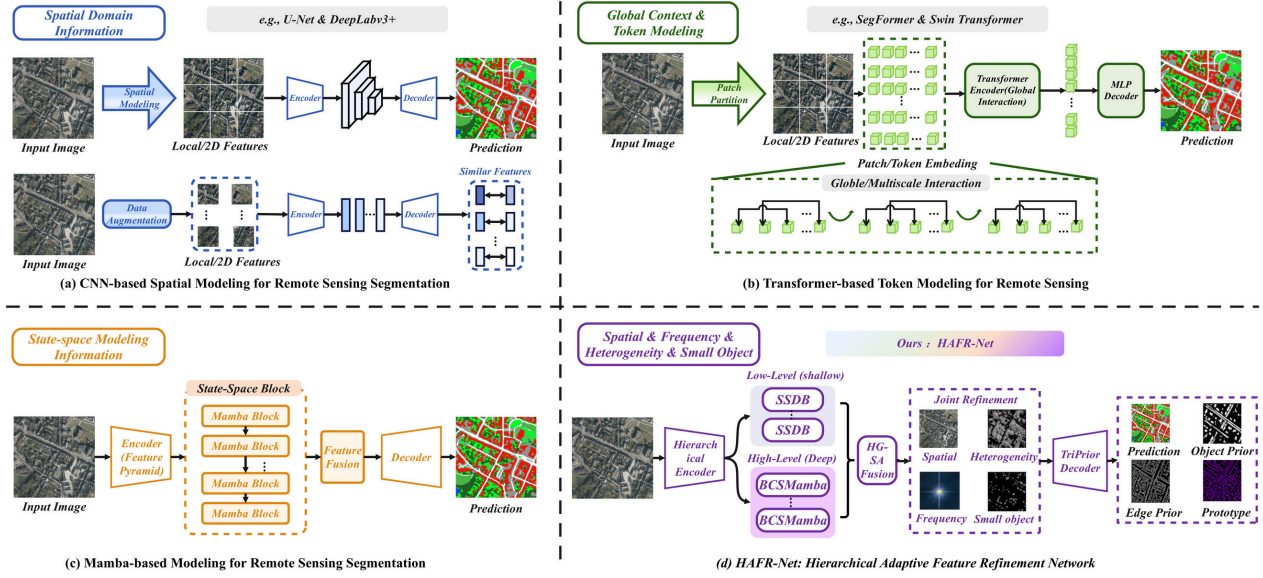


Fig. 1. Positioning of HAFR-Net. (a)–(c) CNN, Transformer and state-space decoders combine the encoder hierarchy through a fixed aggregation rule. (d) HAFR-Net keeps the pretrained hierarchy and refines it progressively, through stage-specialized preparation, heterogeneity-guided stage-adaptive fusion, bounded residual frequency adaptation, and confusion-aware tri-prior decoding.

fusion, conservative feature refinement, and final tri-prior decoding.

The main contributions are as follows.

- We formulate VHR decoding as hierarchical adaptive refinement rather than generic feature aggregation. HG-SAF predicts pixel-wise stage weights from local feature heterogeneity, enabling the contribution of shallow spatial detail and deep semantic context to vary across a scene.
- We introduce a bounded residual refinement strategy for fused hierarchical representations. FRA performs channel-bottlenecked spectral modulation with an explicitly bounded gate and residual initialization, refining detail while starting from the pretrained spatial representation rather than an unconstrained frequency transform.
- We combine adaptive fusion and conservative refinement with confusion-aware tri-prior decoding and validate each design through matched controls. Experiments on four VHR benchmarks include controlled decoder comparisons, direct spatial and frequency alternatives, stage-weight analysis, boundary metrics, class-confusion analysis, statistical uncertainty, and efficiency measurements.

II. RELATED WORK

A. Hierarchical Feature Fusion for Remote Sensing Segmentation

Encoder–decoder models from FCN [4] and U-Net [5] to PSPNet [6], DeepLabv3+ [7], HRNet [25] and UPerNet [13] established multi-level aggregation as a standard recipe for dense prediction. The same idea has been specialized for remote sensing through CNN and Transformer decoders such as ABCNet [26], BANet [27], MANet [28], A²-FPN [29], UNetFormer [14], the FT-UNetFormer implementation in GeoSeg [30] and DCSwin [15]. These methods differ mainly in how encoder stages are combined: some use fixed top-down

laterals, others add global pooling or channel attention [12], [31], [32], and a smaller subset predicts spatially varying weights.

Existing hierarchical decoders therefore already demonstrate the value of multi-level aggregation, while different designs vary in how strongly they adapt fusion to local content. HG-SAF focuses specifically on dense spatial routing conditioned by a lightweight local heterogeneity statistic. It is evaluated against mean fusion, globally pooled gates, and content-adaptive pixel-wise alternatives under the same backbone, so that the contribution of the heterogeneity cue can be isolated from the mere use of a pixel-wise gate.

B. Spatial-Frequency Feature Refinement

Generic spectral mixers including GFNet [16], FNet [33] and FFC [17], together with Fourier operator learning [34], show that Fourier-domain modulation can complement spatial convolution. Recent remote sensing models go further. SFFNet [18] fuses wavelet and spatial features, FNet [19] decouples frequency bands, FGNet [35] inserts FFT-guided filters into a Swin backbone, AFENet [20] modulates bands according to image content, FreDNet [36] performs frequency-guided denoising, and WgANet [37] uses wavelet-guided attention. These works establish frequency modeling as a competitive tool for VHR segmentation rather than as a new primitive.

The remaining design question is how a spectral branch should be attached to an already strong fused representation. HAFR-Net does not use frequency modeling to replace the fused spatial features. FRA restricts frequency modeling to a bounded residual correction inside a channel bottleneck and is directly compared with standard spatial residuals and spectral alternatives at the same insertion point.

C. Structure-Aware Semantic Decoding

Boundary supervision [21], [22], [38], objectness or imbalance losses [39]–[41], and pixel-level contrast or prototype learning [23], [24] are established tools for sharpening contours and separating similar classes. Individual boundary, objectness, and prototype mechanisms are not claimed as novel primitives. CATP uses them as a compact final decoder that follows hierarchical adaptive refinement, and matched controls test whether the combination improves on conventional auxiliary losses.

Recent Transformer and Mamba backbones [8], [9], [11], [42]–[44] strengthen long-range context, and a parallel line keeps a pretrained model fixed and adapts it with lightweight modules, as in the adapter and LoRA fine-tuning of SAM for multimodal remote sensing segmentation [45]. Neither line by itself specifies how the hierarchical features of such an encoder should be organized and refined for VHR scenes. HAFR-Net keeps Swin-B as the representation backbone and places the emphasis on progressive decoding.

III. METHOD

A. Overview

HAFR-Net is designed around a single principle: progressively refine pretrained hierarchical representations rather than replace them with an increasingly complex task-specific decoder. Given an input image $\mathbf{I}$, a pretrained Swin-B encoder [8] produces four stage features $\{F_s\}_{s=1}^4$. Each stage first passes a fixed preparation block T_s , which prepares the stage feature and is kept identical across the module studies of Section VI (Section III-B). HG-SAF then projects the stages to a common space and adaptively aggregates them into F_{fuse} . FRA computes a bounded residual correction to obtain F_{ref} . CATP finally maps F_{ref} to semantic logits, while auxiliary structural signals are used only during training.

The encoder is Swin-B pretrained on ImageNet-22k and fine-tuned on ImageNet-1k [46], whose four stages have widths (128, 256, 512, 1024); a 1×1 projection maps them to (96, 192, 384, 768) before the preparation blocks. Each stage is then aligned to width $C_e=384$ and spatial size $(H_1, W_1) = (H/4, W/4)$ before fusion. As illustrated in Fig. 2, the inference path is

$$P_{\text{sem}} = \text{CATP}\left(\text{SOE}\left(\text{FRA}\left(\text{HG-SAF}\left(\{T_s(F_s)\}_{s=1}^4\right)\right)\right)\right). \quad (1)$$

The decomposition in (1) separates three functions that are commonly entangled in a decoder: selecting hierarchical information, refining the aggregated representation, and imposing the tri-prior decoding constraints. The remainder of this section first records the fixed preparation blocks, then describes the three refinement stages in that order, and closes with their initialization and the training objective.

B. Stage-Specialized Preparation

Three blocks are inherited from our implementation and are not claimed as contributions: SSDB and BCS-Mamba prepare the encoder stages, and SOE acts on the refined feature just

before decoding. They are held identical in every module-level comparison of Section VI, so they cannot account for a difference between the rows of those studies, and Section VI-E measures what they contribute to the margin of Table II. The two stage-preparation blocks are shown in Fig. 3.

The two high-resolution stages use a spectral-spatial decoupled block (SSDB, Fig. 3(a)). A normalized stage feature is processed in parallel by a learnable per-channel complex filter applied to its orthonormal two-dimensional real FFT and by a depthwise 3×3 convolution followed by a pointwise convolution; a gate conditioned on both branches mixes them as a residual update, and a convolutional feed-forward tail closes the block. The two low-resolution stages use a bi-directional cross-scan Mamba block (BCS-Mamba, Fig. 3(b)): rows and columns are scanned as independent sequences, which shortens each scan from $H_1 W_1$ to $\max(H_1, W_1)$ steps, two orthogonal depthwise bridges (3×1 and 1×3) re-inject the cross-axis context that axial scanning removes, and a direction-adaptive weighting merges the horizontal and vertical outputs. The third fixed block, small-object enhancement (SOE), combines parallel dilated depthwise convolutions at rates 2, 4 and 8 with channel attention and a residual scale initialized to 0.1; it is not a stage-preparation block but acts on F_{ref} immediately before decoding, identically in every reported model. Section VI-E measures each of the three on the reference decoder; all three effects are below +0.2pp, so the gains analyzed in this paper are not produced by these fixed blocks.

C. Heterogeneity-Guided Stage-Adaptive Fusion

Hierarchical encoders expose features with progressively lower spatial resolution and stronger semantic abstraction. In VHR imagery, the relative usefulness of these stages can vary considerably within one tile. HG-SAF therefore performs dense rather than globally shared stage weighting.

Each stage is projected to width C_e and upsampled to (H_1, W_1) , giving $\tilde{F}_s$. Local feature heterogeneity is a lightweight statistic of variation in the deepest projected representation,

$$\mathcal{H}(\mathbf{x}) = \frac{1}{C_e} \sum_{c=1}^{C_e} \sqrt{\text{Var}_{k \times k}(\tilde{F}_4^{(c)})(\mathbf{x}) + \varepsilon}, \quad (2)$$

with a $k \times k$ box filter ($k=5$) and $\varepsilon=10^{-5}$. A convolutional gate g_θ , a 3×3 convolution to $C_e/4$ channels followed by normalization, a GELU and a 1×1 convolution to the four stage logits, predicts pixel-wise stage weights $\mathbf{w} = (w_1, \dots, w_4)^\top$, normalized over the stage axis, from the mean stage context $B = \frac{1}{4} \sum_s \tilde{F}_s$ and $\mathcal{H}$ (Fig. 4(a)),

$$\mathbf{w}(\mathbf{x}) = \text{softmax}_s(g_\theta(\text{cat}[B, \mathcal{H}])(\mathbf{x})) \in \Delta^3 \subset \mathbb{R}^4, \quad (3)$$

$$F_{\text{fuse}}(\mathbf{x}) = \sum_{s=1}^4 w_s(\mathbf{x}) \tilde{F}_s(\mathbf{x}). \quad (4)$$

The final projection of g_θ is zero-initialized, so (3) returns uniform stage weights at the first forward pass and (4) reduces to the mean of the projected stages. HG-SAF therefore starts from uniform mean fusion of the projected stages, a non-adaptive reference, and learns spatially varying routing during

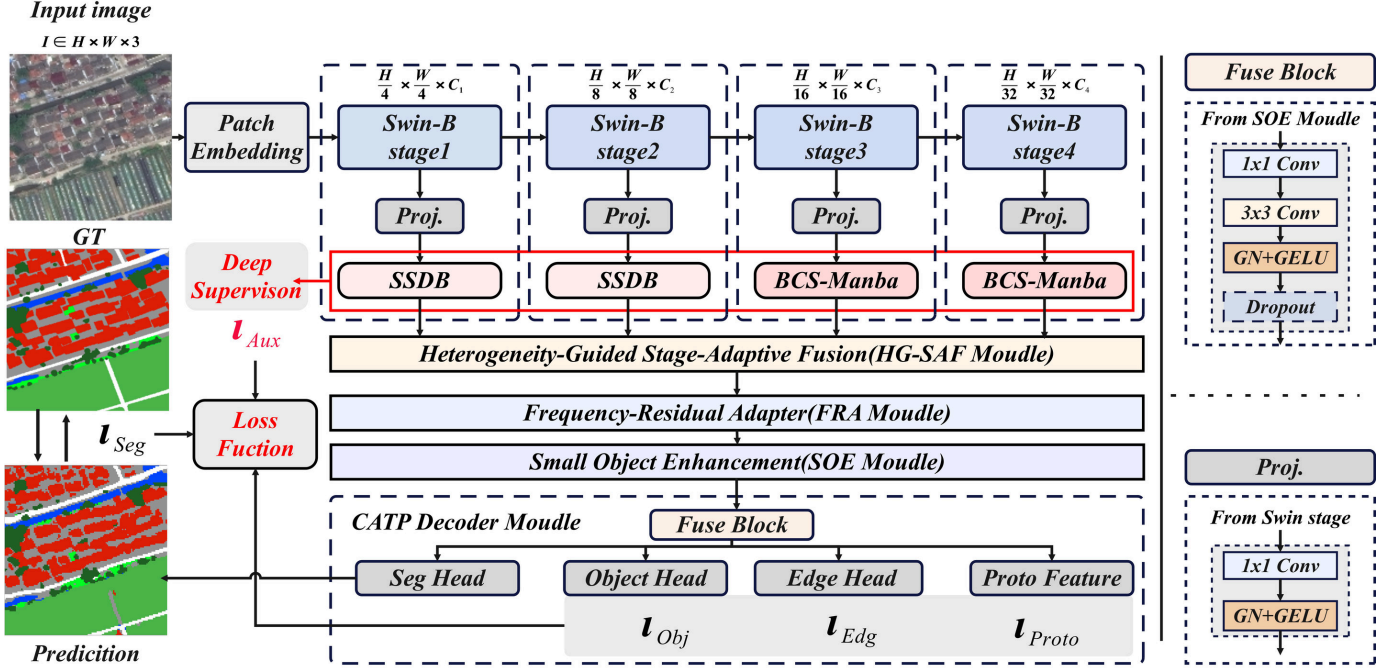


Fig. 2. Overall architecture of HAFR-Net. A pretrained Swin-B encoder produces four hierarchical stages, which pass fixed preparation blocks (SSDB at the two high-resolution stages, BCS-Mamba at the two low-resolution stages). HG-SAF then performs pixel-wise adaptive fusion, FRA applies a bounded residual frequency correction, SOE acts on the refined feature, and CATP produces the semantic prediction. Auxiliary heads are used only in training.

fine-tuning; the stage projections themselves are trained from scratch. Matched controls keep the gate architecture and capacity of g_θ and drop only $\mathcal{H}$ from its input in (3), so that pixel-wise routing and the heterogeneity statistic are not confounded.

D. Frequency-Residual Adapter

After hierarchical fusion, the representation already contains substantial spatial and semantic information. We therefore avoid replacing it with an unrestricted spectral transformation. Instead, FRA uses frequency modeling only to construct a bounded residual correction (Fig. 4(b)). With a channel bottleneck of width $r = \max(C_e/4, 16)$,

$$\hat{Z} = \text{rFFT}(\text{Conv}_{1 \times 1}^{C_e \rightarrow r}(F_{\text{fuse}})), \quad (5)$$

$$\tilde{Z} = \hat{Z} \odot g, \quad g = 1 + \alpha \tanh(g_\phi(G)), \quad (6)$$

$$F_{\text{ref}} = F_{\text{fuse}} + \gamma \text{Conv}_{1 \times 1}^{r \rightarrow C_e}(\text{irFFT}(\tilde{Z}; H_1, W_1)), \quad (7)$$

Both transforms are orthonormal and act on the half spectrum, so the output of (7) is real by construction. Here G concatenates the log-amplitude, the sine and cosine of the phase, and a radial prior over the half-spectrum grid, $\alpha=0.5$, and γ is initialized to 0. Three constraints follow from (5)–(7). The bottleneck of width r in (5) caps both the cost of the spectral branch and the channel rank of its residual mapping. The multiplier g of (6) obeys $|g - 1| < \alpha$, so the relative rescaling of any spectral coefficient stays inside $(1-\alpha, 1+\alpha)$ and no coefficient can be arbitrarily amplified; the overall correction is additionally scaled by γ . The residual scale of (7) makes the first forward pass an exact identity: the branch starts from zero contribution and is learned from data during fine-tuning. At

inference, FRA remains a single residual correction on F_{fuse} ; it does not introduce an auxiliary prediction head.

E. Confusion-Aware Tri-Prior Decoder

Even after hierarchical fusion and feature refinement, a single semantic head provides only region-level supervision. CATP therefore uses lightweight auxiliary signals to regularize boundaries, foreground structure, and inter-class relations during training.

A two-layer fuse block produces $\bar{F}$. A boundary branch of two 3×3 convolutions at width $C_e/4$ followed by a 1×1 projection predicts single-channel boundary logits P_{edge} , which are fed back as

$$\bar{F}' = \bar{F} + \beta_e \text{enhance}(\text{maxpool}_{3 \times 3}(\sigma(P_{\text{edge}})) \odot \bar{F}), \quad (8)$$

where σ is the logistic sigmoid, broadcast over channels, enhance is a 3×3 convolution with normalization and GELU, and β_e is initialized to 0.05. Three 1×1 heads applied to $\bar{F}'$ output semantic logits P_{sem} , objectness logits P_{obj} and a prototype embedding $Q \in \mathbb{R}^{64 \times H_1 \times W_1}$ (Fig. 5). The objectness target T_{obj} marks the valid pixels outside the two background-like classes of each legend (*Impervious surface* and *Clutter* on ISPRS, *Background* and *Barren* on LoveDA, *Developed* and *Bareland* on OpenEarthMap), so the head learns a coarse foreground prior; small-component IoU is reserved for evaluation. The edge target is the 4-neighbour class-difference mask. Both $\mathcal{L}_{\text{edge}}$ and $\mathcal{L}_{\text{obj}}$ are binary CE in which positive pixels carry the weight $\text{clip}(N_-/N_+, 1, 20)$, where N_+ and N_- count positive and negative pixels of the mini-batch.

Some class pairs exhibit substantially larger cross-class errors than others. We therefore define a small relation set $\mathcal{C}$

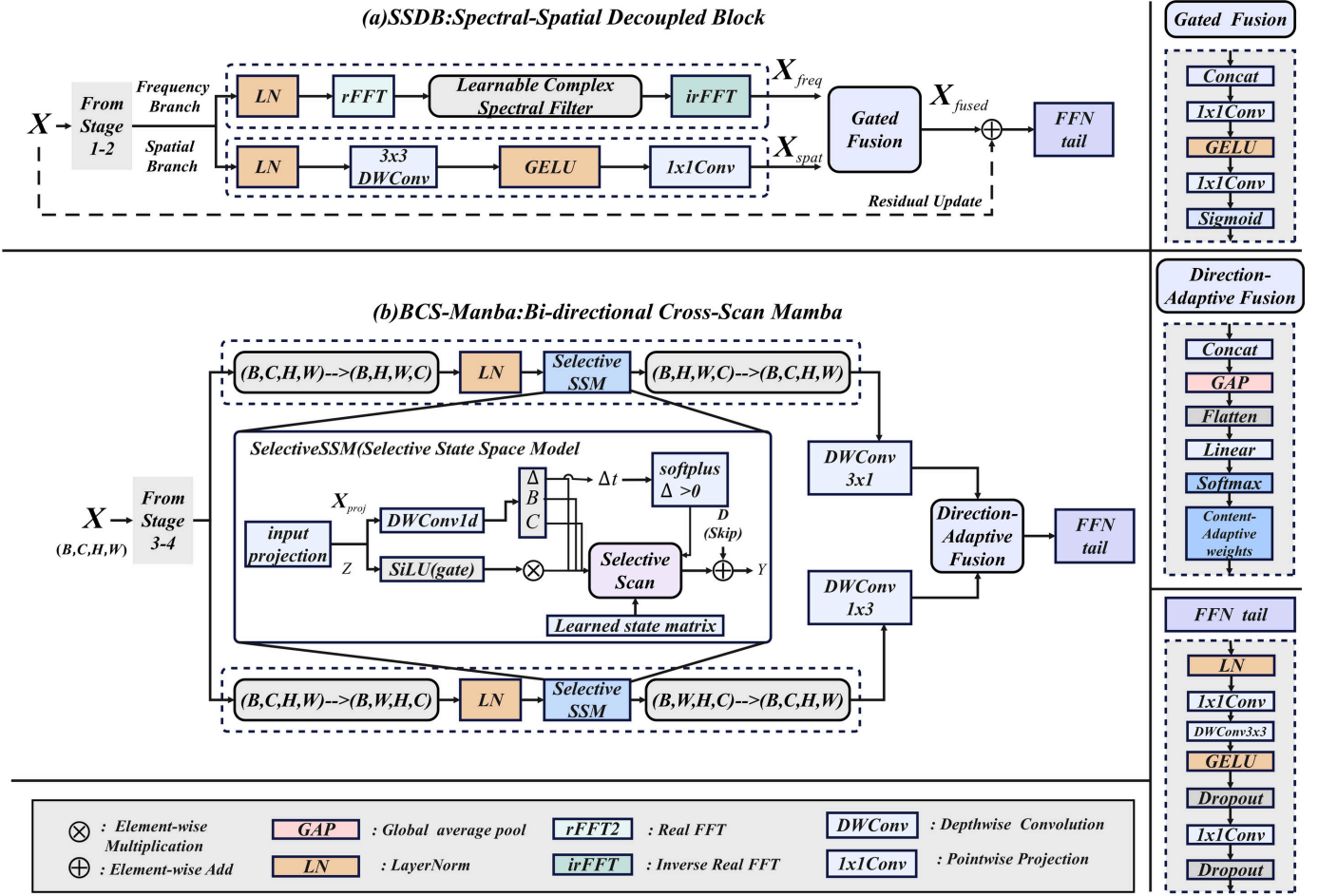


Fig. 3. Fixed stage preparation blocks, shown with the shared sub-blocks on the right. (a) SSDB decouples a normalized stage feature into a learnable complex spectral branch and a depthwise spatial branch, and merges them by a gated residual update. (b) BCS-Mamba scans rows and columns as independent selective state-space sequences, restores cross-axis context through two orthogonal depthwise bridges, and merges the two directions by content-adaptive weights. Both blocks are held identical across the module studies of Section VI; their measured effect is reported in Section VI-E.

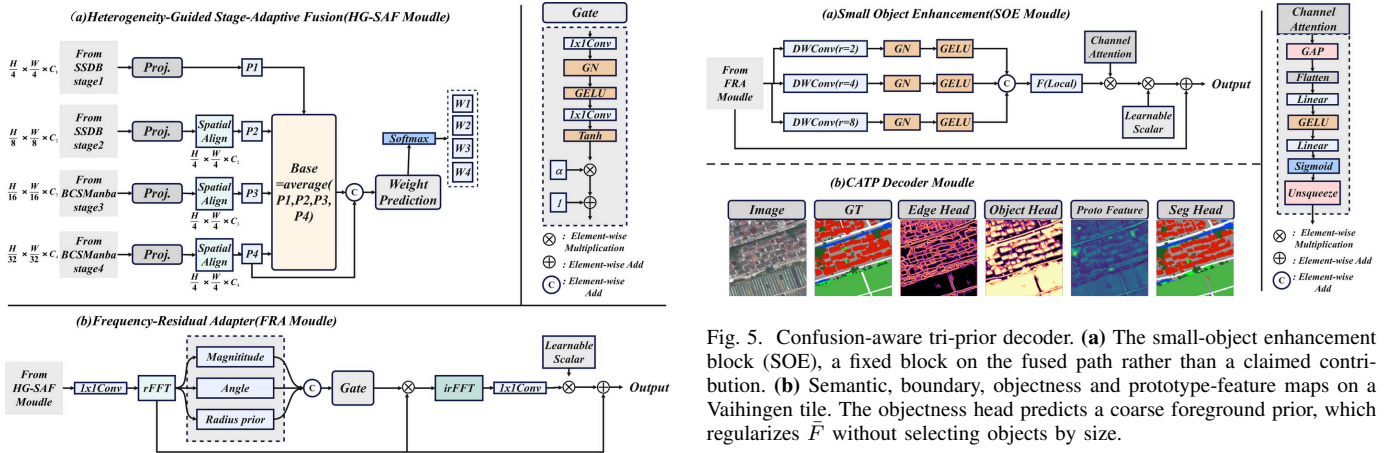


Fig. 4. The two refinement blocks after the encoder pyramid. (a) HG-SAF predicts per-pixel stage weights from B and $\mathcal{H}$ with a zero-initialized gate. (b) FRA adds a channel-bottlenecked residual rFFT branch with a tanh-bounded gate and $\gamma=0$.

from a pilot model using only an internal split of the training partition; the official evaluation split is never used for pair selection. Per-class prototype p_c is the mini-batch mean of Q

Fig. 5. Confusion-aware tri-prior decoder. (a) The small-object enhancement block (SOE), a fixed block on the fused path rather than a claimed contribution. (b) Semantic, boundary, objectness and prototype-feature maps on a Vaihingen tile. The objectness head predicts a coarse foreground prior, which regularizes $\bar{F}$ without selecting objects by size.

over the valid pixels of class c , re-normalized to unit ℓ_2 norm so that $\langle p_a, p_b \rangle$ is a cosine similarity; classes absent from a batch are skipped, and

$$\mathcal{L}_{\text{proto}} = \frac{1}{|C|} \sum_{(a,b) \in C} \max(0, m + \langle p_a, p_b \rangle), \quad (9)$$

with $m=0.2$, so a pair stops contributing once its prototypes

reach $\langle p_a, p_b \rangle \leq -m$. On the six-class ISPRS label set, $\mathcal{C}$ contains four pairs. The pair lists for all four datasets are given in the supplement. At inference, only P_{sem} is used.

F. Conservative Initialization of Refinement Branches

HAFR-Net uses conservative initialization to avoid introducing large task-specific perturbations at the beginning of fine-tuning [47]–[49]. The three components do not share an identical mathematical behavior. As summarized in Table I, HG-SAF starts from baseline-equivalent uniform stage fusion, FRA is an exact residual identity when $\gamma=0$ in (7), and the structural feedback path in CATP starts from the small residual coefficient β_e of (8). We distinguish these cases rather than treating all components as exact identities.

These three reference behaviors are recorded so that later ablations can isolate initialization from architecture. The next subsection defines the training objective on the semantic head and the auxiliary structural signals.

G. Training Objective

Let $\Omega = \{i: y_i \neq 255\}$ be the valid pixels, so that ignore index 255 keeps unlabeled and eroded-boundary pixels out of $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{FocalDice}}$. Let $\text{DSC}_c = (2 \sum_{i \in \Omega} p_{i,c} y_{i,c} + \varepsilon_d) / (\sum_{i \in \Omega} p_{i,c} + \sum_{i \in \Omega} y_{i,c} + \varepsilon_d)$ with $\varepsilon_d=1$, one-hot targets $y_{i,c}$ and $p = \text{softmax}_c(P_{\text{sem}})$. The semantic loss is $\mathcal{L}_{\text{sem}} = \mathcal{L}_{\text{CE}} + \lambda_d \mathcal{L}_{\text{FocalDice}}$, where

$$\mathcal{L}_{\text{FocalDice}} = \frac{1}{N_{\text{cls}}} \sum_{c=1}^{N_{\text{cls}}} (1 - \text{DSC}_c)^\kappa, \quad (10)$$

N_{cls} is the number of scored classes, $\kappa=2.5$, and $\mathcal{L}_{\text{CE}}$ uses class-balanced weights (*Car* 3 $\times$, *Clutter* 5 $\times$ on ISPRS). The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{sem}} + \lambda_e \mathcal{L}_{\text{edge}} + \lambda_o \mathcal{L}_{\text{obj}} + \lambda_p \mathcal{L}_{\text{proto}} + \mathcal{L}_{\text{aux}}, \quad (11)$$

with $\lambda_d=1.0$, $\lambda_e=0.4$, $\lambda_o=0.2$ and $\lambda_p=0.1$. The auxiliary term $\mathcal{L}_{\text{aux}}$ attaches CE and FocalDice heads to the prepared stages $\{T_s(F_s)\}_{s=1}^4$ with weights (0.4, 0.3, 0.2, 0.1). Every term of (11) is evaluated at the label resolution: the auxiliary and structural logits are bilinearly upsampled to the input size, so the discrete label map is never interpolated and keeps its ignore index. Only the class map of $\mathcal{L}_{\text{proto}}$ is resampled, to (H_1, W_1) and by nearest neighbour. All auxiliary heads are disabled at inference. The next section specifies the datasets, matched protocol, and metrics used to evaluate this design.

TABLE I
CONSERVATIVE INITIALIZATION OF THE THREE REFINEMENT BRANCHES

Branch	Init.	First forward	Role
HG-SAF	last gate = 0	uniform 1/4 fusion	adaptive fusion
FRA	$\gamma = 0$	exact identity	residual refine.
CATP fb.	$\beta_e=0.05$	near-ref. residual	struct. feedback

The three branches do not share one mathematical identity: HG-SAF starts from mean fusion, FRA is an exact residual identity, and the CATP feedback path is a small residual.

IV. EXPERIMENTAL SETUP

A. Datasets and Exact Splits

ISPRS Vaihingen [50] contains 33 NIR-RG tiles at 9 cm GSD. The 16 official training tiles are 1, 3, 5, 7, 11, 13, 15, 17, 21, 23, 26, 28, 30, 32, 34 and 37; the remaining 17 tiles form the test set, stored as 113 non-overlapping 1024 $\times$ 1024 patches, which are the evaluation and bootstrap unit. Five foreground classes are scored; *Clutter* is excluded; official 3 px boundary erosion and ignore index 255 are used.

ISPRS Potsdam [51] contains 38 RGB tiles at 5 cm GSD and the same six-class legend. Train tiles: 2_10–2_12, 3_10–3_12, 4_10–4_12, 5_10–5_12, 6_7–6_12, 7_7–7_12. Test tiles: 2_13, 2_14, 3_13, 3_14, 4_13–4_15, 5_13–5_15, 6_13–6_15, 7_13. The same five-class eroded protocol is used.

LoveDA [52] contains 5987 RGB images at 30 cm GSD and seven classes. We report the official validation split. Several published Transformer and Mamba numbers were originally submitted to the online test server; they appear only as contextual literature values.

OpenEarthMap [53] contains 5000 images at 25–50 cm GSD. mIoU averages the eight official land-cover classes; unlabeled pixels are ignored. There is no separate Background class.

Every image enters the network at 512 $\times$ 512: the ISPRS tiles are held as the non-overlapping 1024 $\times$ 1024 patches listed above and each patch is resampled to that size, as are the LoveDA and OpenEarthMap images, which the two benchmarks already distribute as single tiles. Labels are resampled by nearest neighbour so that class indices stay valid, and the ignore index is preserved. Normalization uses ImageNet mean and standard deviation. No CRF, multi-crop merging, or other post-processing is applied.

B. Controlled Baselines and Published Comparisons

Conclusions are drawn from a controlled Swin-B comparison: UPerNet [13], FPN [12], DeepLabv3+ [7], SegFormer MLP [9], a UNetFormer-style decoder [14], SE-gated fusion [31] and an FFT mixer [16], all retrained with the single recipe of Section IV-C. Published CNN, Transformer and Mamba numbers are provided for contextual comparison and are not treated as strictly controlled comparisons, because training, backbone, metric, and inference configurations differ across sources. FT-UNetFormer is cited as the GeoSeg implementation [30]. A provenance table is given in the supplement.

C. Implementation Details

Training uses PyTorch 2.1 [54] and one RTX 4090. Every controlled model, including HAFA-Net, shares one recipe: 512 $\times$ 512 inputs, at most 80 epochs with batch size 8, AdamW [55] with head learning rate 1.2×10^{-4} and backbone learning rate 4×10^{-5} , a cosine schedule with a 5-epoch warmup [56], weight decay 10^{-4} , gradient clipping 1.0 and BF16 mixed precision. Augmentation is horizontal or vertical flip and discrete 90° rotation, each with probability 0.5. The auxiliary deep supervision of (11) belongs to this shared recipe: every controlled model carries CE and FocalDice heads on the corresponding encoder stages with the weights

(0.4, 0.3, 0.2, 0.1), so Table II compares decoding paths and not the amount of training signal. Seeds are 42, 43 and 44, and the reported checkpoint is the best-mIoU epoch under one monitoring rule that is identical for every controlled model.

Controlled inference is single-scale without test-time augmentation (TTA). For the two ISPRS sets we additionally report flip and $\{90^\circ, 180^\circ, 270^\circ\}$ TTA so that TTA gain can be separated from method gain. LoveDA is the only dataset with a stronger photometric recipe: color jitter 0.4, Gaussian blur 0.3, EMA with decay 0.999 [57] and early stopping with patience 12; ISPRS and OpenEarthMap use none of these. This difference is recorded rather than hidden, and no controlled baseline on LoveDA is trained under a weaker recipe than HAFR-Net.

D. Evaluation Metrics

We report mIoU and per-class IoU. HG-SAF is further examined through heterogeneity- and size-stratified IoU and stage-weight statistics; the strata are computed once from the reference model and reused for every variant, so all rows are scored on identical pixel sets. FRA is examined through Boundary IoU [58], boundary F-score [59], thin-structure IoU (local width below 8 px) and small-component IoU (area quartile Q1). CATP is examined through pairwise confusion mass $E_{a,b} = M_{ab} + M_{ba}$ on the pre-declared relation set, where M is the confusion matrix normalized by the number of valid pixels. Uncertainty is a paired bootstrap (1 000 resamples) of the mIoU difference versus UPerNet over the non-overlapping evaluation patches, so no pixel enters a replicate twice. Efficiency uses input 512×512 , batch size 1, BF16, 50 warm-up iterations and 500 timed iterations with `torch.cuda.synchronize()`; the timed loop excludes host-device copy and post-processing.

V. EXPERIMENTAL RESULTS

A. Controlled Comparison

Table II reports the controlled comparison under a matched training and inference protocol, and the conclusions of this paper are drawn from it. It holds the encoder, the schedule and the inference setting fixed, so that a difference between two rows is a difference in the decoding path rather than in training or inference. HAFR-Net carries the fixed blocks of Section III-B in addition to its three refinement stages, and Section VI-E measures their separate share of the margin.

Under the no-TTA protocol, HAFR-Net improves Swin-B+UPerNet by +0.55 pp on Vaihingen (95% CI [0.31, 0.79]), +0.95 pp on Potsdam ([0.66, 1.22]), +1.55 pp on LoveDA ([1.18, 1.91]), and +1.84 pp on OpenEarthMap ([1.47, 2.19]); all four intervals exclude zero, and the ordering is preserved across the three seeds. The gains are larger on the more heterogeneous LoveDA and OpenEarthMap sets than on the saturated ISPRS urban tiles. Among the controlled decoders, the UNetFormer-style decoder is the strongest alternative, and HAFR-Net exceeds it by +0.33 to +1.52 pp across the four sets. TTA adds a further +0.36 / +0.34 pp on the two ISPRS sets. The improvement holds on all four datasets under a single

TABLE II
CONTROLLED COMPARISON OF DECODERS UNDER A MATCHED PROTOCOL

Method (Swin-B)	Vaih.	Pots.	LoveDA	OEM	Avg.
UPerNet [13]	83.57	86.91	53.62	65.86	72.49
FPN [12]	83.21	86.54	53.18	65.41	72.09
DeepLabv3+ [7]	83.38	86.72	53.35	65.58	72.26
SegFormer MLP [9]	83.44	86.80	53.48	65.71	72.36
UNetFormer dec. [14]	<u>83.79</u>	<u>87.14</u>	<u>54.02</u>	<u>66.18</u>	<u>72.78</u>
SE-gated fusion [31]	83.68	87.05	53.88	66.02	72.66
FFT spectral mixer [16]	83.62	86.98	53.71	65.94	72.56
HAFR-Net (no TTA)	84.12	87.86	55.17	67.70	73.71
<i>TTA (flip+rotation), reported only for the two ISPRS sets</i>					
UPerNet [†]	<u>83.89</u>	<u>87.24</u>	–	–	–
HAFR-Net[†]	84.48	88.20	–	–	–

All models share the split, crop, optimizer, schedule, augmentation, and checkpoint rule of Section IV-C, and use single-scale inference without TTA unless marked [†]. Entries are the mean mIoU (%) over seeds 42, 43, and 44. The paired bootstrap of Section IV-D gives 95% confidence intervals of [0.31, 0.79], [0.66, 1.22], [1.18, 1.91] and [1.47, 2.19] pp for the no-TTA margin of HAFR-Net over UPerNet on the four datasets, so every margin excludes zero. The best and second-best entries in each column are **bold red** and underlined blue, respectively. The no-TTA and TTA blocks are ranked separately.

protocol, so it follows from the decoding design rather than from a dataset-specific trick.

B. Contextual Comparison with Published Results

For broader context, we additionally report representative published results from CNN-, Transformer-, and Mamba-based remote sensing models. Because these values are obtained under different backbones, training recipes, and inference settings, they are treated as contextual references rather than strict controlled comparisons. Under this reading, HAFR-Net attains the best or the second-best reported accuracy on each of the four benchmarks.

Table III summarizes the two ISPRS urban benchmarks. With TTA, HAFR-Net reaches 84.48% mIoU on Vaihingen and 88.20% on Potsdam. The per-class pattern is consistent with the later analyses: on Vaihingen HAFR-Net leads only on *Low Vegetation*, and on Potsdam its largest margins are on *Impervious surface* (+0.8 pp) and *Low Vegetation* (+0.7 pp). HAFR-Net holds the highest Potsdam mIoU of the table; on Vaihingen only PyramidMamba reports a higher mean, by 0.36 pp, with a different backbone and TTA setting.

Table IV reports LoveDA on the official validation split without TTA. HAFR-Net attains 55.17% mIoU. The largest per-class margin is on *Barren*, the class that the training-only relation set $\mathcal{C}$ of (9) pairs with *Background* on this dataset. Only D2LS reports a higher mean, by 0.13 pp, and part of the table was produced on the online test server rather than on the validation split, as recorded in the supplement.

Table V reports OpenEarthMap under the eight-class official protocol. The no-TTA model reaches 67.70% mIoU. Among the pre-declared relation pairs, the largest reduction of confusion mass is on *Rangeland* $\leftrightarrow$ *Agriculture* (Table VIII, Panel B). Across the three tables HAFR-Net therefore holds the best published mean on Potsdam and OpenEarthMap and trails the best entry by 0.36 and 0.13 pp on Vaihingen and

TABLE III
CONTEXTUAL COMPARISON WITH PUBLISHED RESULTS ON ISPRS VAIHINGEN AND POTSDAM

Method	ISPRS Vaihingen						ISPRS Potsdam					
	Imp.S.	Build.	L.Veg.	Tree	Car	mIoU	Imp.S.	Build.	L.Veg.	Tree	Car	mIoU
<i>CNN-based</i>												
UNet [5]	90.1	88.0	69.1	78.4	64.5	78.02	82.8	89.9	73.8	75.6	83.2	81.06
PSPNet [6]	91.0	88.7	69.5	79.1	66.8	79.02	84.2	91.0	74.8	76.5	85.2	82.34
DenseASPP [60]	90.3	87.6	68.4	77.9	62.5	77.34	83.3	90.3	73.5	75.3	83.8	81.24
MANet [28]	91.2	89.2	70.1	79.4	68.9	79.76	84.8	91.8	75.4	77.1	86.2	83.06
HRNet [25]	91.6	89.6	70.5	79.9	69.9	80.30	83.8	90.7	74.2	76.1	84.5	81.86
ABCNet [26]	90.7	88.5	69.3	78.9	67.6	79.00	84.3	91.2	75.0	76.7	85.5	82.54
BANet [27]	89.8	87.3	68.1	77.8	65.0	77.60	83.5	90.5	73.9	75.7	84.2	81.56
A ² -FPN [29]	91.4	89.4	70.4	79.5	69.5	80.04	85.2	92.1	75.9	77.5	86.6	83.46
CGNet [61]	90.5	88.1	68.9	78.6	67.2	78.66	84.0	90.8	74.5	76.4	85.0	82.14
<i>Transformer-based</i>												
FT-UNetFormer [†] [30]	91.9	89.9	71.1	80.3	73.3	81.30	86.2	92.9	76.8	78.3	88.0	84.44
DCSwin [†] [15]	92.7	90.5	71.7	80.8	75.1	82.16	88.0	94.2	78.6	80.8	91.9	86.70
Mask2Former [†] [42]	92.9	90.7	72.1	81.2	75.4	82.46	82.8	94.0	79.2	82.8	73.2	82.40
LOGCAN++ [62]	92.1	90.3	72.0	81.2	77.1	82.54	87.6	93.8	77.9	80.0	89.9	85.84
CG-Swin [†] [63]	93.1	90.8	72.4	81.3	77.6	83.04	88.7	<u>94.9</u>	79.4	81.3	93.4	87.54
UNetFormer [†] [14]	93.4	91.2	72.6	81.8	78.1	83.42	86.9	93.4	77.5	79.2	88.7	85.14
MMT [†] [64]	93.4	91.4	72.9	81.8	78.7	83.64	89.6	93.8	78.9	81.2	92.7	87.24
<i>Mamba-based</i>												
RS3Mamba [†] [43]	92.5	89.9	72.1	81.0	76.2	82.34	87.3	93.6	77.8	79.5	90.5	85.74
UNetMamba [†] [44]	92.9	90.8	72.7	81.2	76.8	82.88	88.3	94.4	78.7	80.7	92.1	86.84
D2LS [†] [65]	93.6	91.6	73.3	<u>82.0</u>	79.1	83.92	89.2	<u>94.9</u>	79.2	81.7	<u>93.1</u>	87.62
PyramidMamba [†] [11]	94.2	92.5	<u>74.7</u>	82.3	80.5	84.84	<u>89.9</u>	94.6	<u>79.5</u>	81.7	93.4	<u>87.82</u>
HAFR-Net[†] (Ours)	<u>93.8</u>	<u>92.1</u>	75.0	<u>82.0</u>	<u>79.5</u>	<u>84.48±0.11</u>	90.7	95.1	80.2	<u>82.1</u>	92.9	88.20±0.09

Per-class IoU and mIoU (%) on the five scored foreground classes (Imp.S.: impervious surface; Build.: building; L.Veg.: low vegetation); *Clutter* is excluded and the official 3-px boundary erosion is applied. Methods marked [†] use TTA as reported by their sources. Because backbones, splits, and inference settings differ across papers, these values are literature context rather than a matched benchmark; the matched Swin-B comparison is Table II. The best and second-best entries in each column are **bold red** and underlined blue, respectively. Best and second best are ranked separately within each dataset. The $\pm$ value is the three-seed standard deviation of HAFR-Net.

TABLE IV
CONTEXTUAL COMPARISON WITH PUBLISHED RESULTS ON THE LOVE-DA VALIDATION SPLIT

Method	Backgr.	Building	Road	Water	Barren	Forest	Agricult.	mIoU
<i>CNN-based</i>								
UNet [5]	50.24	59.77	52.79	57.02	31.08	41.73	51.81	49.20
PSPNet [6]	52.50	59.46	51.15	61.10	27.46	37.48	50.32	48.49
DenseASPP [60]	50.77	55.67	53.08	55.49	23.99	35.24	43.54	45.40
MANet [28]	49.82	58.08	52.98	54.30	25.21	37.65	51.87	47.13
HRNet [25]	49.90	57.33	55.94	52.34	31.64	41.94	48.50	48.23
A ² -FPN [29]	47.73	59.19	52.20	64.08	<u>32.09</u>	37.96	53.00	49.46
<i>Transformer-based</i>								
LOGCAN++ [62]	47.37	58.38	56.46	80.05	18.44	<u>47.91</u>	64.80	53.35
AerialFormer [66]	47.80	<u>60.70</u>	59.30	81.50	17.90	47.90	64.00	54.10
SFA-Net [67]	48.40	60.30	<u>59.10</u>	81.90	24.10	46.20	64.00	54.90
<i>Mamba-based</i>								
RS3Mamba [43]	41.60	58.23	54.03	77.34	17.97	43.81	61.37	50.62
UNetMamba [44]	47.08	59.16	56.74	81.37	18.15	46.61	<u>64.31</u>	53.35
PyramidMamba [11]	51.07	56.39	52.44	66.81	28.95	35.13	45.84	48.09
D2LS [65]	47.60	61.20	<u>59.10</u>	<u>81.60</u>	23.80	48.80	64.80	55.30
HAFR-Net (Ours)	<u>52.32</u>	60.16	58.80	70.38	39.80	42.67	62.05	<u>55.17±0.12</u>

Per-class IoU and mIoU (%) on the seven LoveDA classes. HAFR-Net is evaluated on the official validation split with single-scale inference and no TTA. Several published Transformer and Mamba entries were originally submitted to the online test server and are therefore not strictly comparable; they are shown only as literature context, and the matched Swin-B comparison is Table II. The best and second-best entries in each column are **bold red** and underlined blue, respectively. The $\pm$ value is the three-seed standard deviation of HAFR-Net; per-class standard deviations do not exceed 0.18 and are omitted for readability.

LoveDA, and Table II is the matched evidence behind that standing.

VI. ANALYSIS AND ABLATION

The previous section reports overall accuracy. This section asks a narrower question: whether each refinement stage behaves in the way it was designed to behave. We keep the

TABLE V
CONTEXTUAL COMPARISON WITH PUBLISHED RESULTS ON OPENEARTHMAP

Method	Bareland	Range.	Develop.	Road	Tree	Water	Agricult.	Building	mIoU
<i>CNN-based</i>									
UNet [5]	33.6	49.2	49.8	56.1	64.7	71.4	70.9	72.3	58.50
PSPNet [6]	35.2	50.1	50.6	57.4	65.8	72.6	71.8	73.0	59.56
DenseASPP [60]	32.9	48.4	49.1	55.6	63.9	70.7	70.1	71.6	57.79
MANet [28]	36.0	50.8	51.2	58.0	66.5	73.4	72.5	73.8	60.27
HRNet [25]	35.6	50.5	50.9	57.7	66.1	73.0	72.1	73.4	59.91
ABCNet [26]	34.4	49.7	50.2	56.9	65.3	72.1	71.4	72.7	59.09
BANet [27]	33.0	48.6	49.4	55.8	64.2	70.9	70.4	71.9	58.03
A ² -FPN [29]	36.4	51.0	51.5	58.3	66.8	73.7	72.8	74.1	60.57
CGNet [61]	34.9	49.9	50.4	57.1	65.6	72.4	71.6	72.9	59.35
<i>Transformer-based</i>									
SegFormer [†] [9]	41.0	56.4	53.2	58.7	70.9	77.7	76.7	76.2	63.85
FT-UNetFormer [†] [30]	41.8	55.0	52.6	57.6	69.4	77.2	75.5	75.1	63.02
DCSwin [†] [15]	42.5	55.6	53.2	58.1	69.9	77.8	76.0	75.7	63.60
Mask2Former [†] [42]	42.6	56.0	53.9	59.8	70.0	77.5	76.3	76.4	64.06
UNetFormer [†] [14]	42.8	56.2	53.5	60.9	70.2	77.4	76.6	76.9	64.31
LOGCAN++ [62]	43.6	57.8	55.4	62.1	70.8	78.9	77.6	77.0	65.40
<i>Mamba-based</i>									
RS3Mamba [†] [43]	39.9	51.0	48.7	56.9	66.8	74.4	75.0	71.4	60.51
UNetMamba [†] [44]	43.2	57.1	54.8	61.4	70.5	78.3	77.1	77.2	64.95
PyramidMamba [†] [11]	<u>45.0</u>	<u>59.4</u>	<u>57.9</u>	<u>64.9</u>	72.1	81.3	<u>79.4</u>	<u>79.6</u>	<u>67.45</u>
HAFR-Net (Ours)	45.6	59.8	58.4	65.3	<u>71.9</u>	<u>80.8</u>	79.7	80.1	67.70$\pm$0.10

Per-class IoU and mIoU (%) over the eight official land-cover classes; unlabeled pixels are ignored and are not counted as a Background class. Methods marked [†] use TTA as reported by their sources, whereas HAFR-Net uses single-scale inference. These literature numbers are not a matched protocol; see Table II for the controlled comparison. The best and second-best entries in each column are **bold red** and underlined blue, respectively. The $\pm$ value is the three-seed standard deviation of HAFR-Net; per-class standard deviations do not exceed 0.13 and are omitted for readability.

Swin-B encoder fixed and swap only the component under test, and we close with visual examples of the strata that the analyses identify. Unless a table note states otherwise, ISPRS scores in this section follow the flip-and-rotation protocol of Table III, so the full model is the 84.48% entry of Table II.

A. Stage-Adaptive Fusion

Table VI isolates the contribution of HG-SAF. Every row of Panel A replaces only the fusion stage, so the ladder decomposes the +0.20 pp that HG-SAF gains over mean fusion: a globally pooled gate recovers +0.04, a pixel-wise gate that sees only the mean context B recovers +0.06, and the heterogeneity statistic of (2) supplies the remaining +0.14. The cue is therefore the larger part of the effect, and it is the statistic rather than the gate that carries it: with the same gate and capacity, the input-gradient and prediction-entropy proxies recover only +0.06 and +0.03 of that +0.14.

Panel B of Table VI shows that stratified IoU increases most in the highest quartile of (2) (+1.57 pp) and on small objects (+1.52 pp), whereas large objects change by only +0.19 pp. Panel C reports the corresponding routing statistics, and Fig. 6 shows them directly on the predicted maps. The heterogeneity statistic tracks object outlines and small structures (Fig. 6(c)), and the predicted weight field varies spatially at a comparable scale (Fig. 6(d)). The shift is consistent in sign across strata: shallow stages take a higher share on boundaries, small objects and high- $\mathcal{H}$ pixels than in large interiors, the deep stages keep the larger share of the total weight everywhere, and the pixel-wise trend of $w_1 + w_2$ against $\mathcal{H}$ is positive (Fig. 6(e),(f)). HG-SAF therefore redistributes stage emphasis toward detail

exactly in the strata where Panel B reports the largest gains, while the deep semantic context of the pretrained pyramid remains dominant. Fig. 7 provides a qualitative counterpart on eight tiles.

The fused maps in Fig. 7 retain shallow edge responses in heterogeneous regions and deeper homogeneous responses in large interiors. This visualization is the qualitative counterpart of the routing statistics in Panel C. The next subsection asks whether a subsequent residual correction can further improve structural detail.

B. Frequency-Residual Adaptation

Table VII compares FRA with matched spatial and spectral alternatives at the same fused-feature location. The goal is to separate three possible sources of gain: the use of a residual branch, the use of a frequency operator, and the bounded low-rank parameterization.

A parameter-matched spatial residual and a standard FFT filter change mIoU only slightly, whereas unrestricted full-rank mixing is the weakest variant. A GFNet mixer with the same zero-init residual parameterization already recovers part of the gain and is the second-best variant on four of the five accuracy metrics, which supports residual parameterization itself; the additional bounded low-rank design of FRA further improves Boundary IoU and thin-structure IoU. Inspecting the learned gate confirms that it operates well inside the envelope that (6) imposes: averaged over the 113 Vaihingen test patches and over all coefficients of the half spectrum, its distance from identity is $|g-1| = 0.28$, that is, 56% of the $\alpha=0.5$ bound, so the branch is neither idle nor saturated at

TABLE VI
ANALYSIS OF HETEROGENEITY-GUIDED STAGE-ADAPTIVE FUSION ON ISPRS VAIHINGEN

Setting		Pixel-wise	Extra cue	mIoU	Δ
<i>Panel A: fusion-gate controls</i>					
H1	Mean fusion	–	–	84.28	↓0.20
H2	Global-pool gate	–	GAP	84.32	↓0.16
H3	Pixel gate, B only	✓	–	84.34	↓0.14
H4	Pixel gate + input gradient	✓	$ \nabla I $	84.40	↓0.08
H5	Pixel gate + entropy	✓	$\mathbb{H}(p)$	84.37	↓0.11
H6	HG-SAF ($B+\mathcal{H}$)	✓	$\mathcal{H}$	84.48	–

Stratum	UPerNet	+HG-SAF	Δ	Region	Shallow	Deep
<i>Panel B: stratified IoU</i>				<i>Panel C: mean stage weights</i>		
Heterogeneity Q1 (lowest)	86.12	86.28	+0.16	Boundary	0.37	0.63
Heterogeneity Q2	84.05	84.41	+0.36	Small object	0.38	0.62
Heterogeneity Q3	81.74	82.38	+0.64	High $\mathcal{H}$	0.37	0.63
Heterogeneity Q4 (highest)	76.84	78.41	+1.57	Low $\mathcal{H}$	0.36	0.64
Small objects (area Q1)	74.52	76.04	+1.52	Large interior	0.35	0.65
Medium objects	83.56	83.94	+0.38			
Large objects	87.12	87.31	+0.19			

Panel A isolates the heterogeneity statistic $\mathcal{H}$ from pixel-wise routing. Every row replaces only the fusion stage and keeps the preparation blocks, FRA and CATP, so H1 coincides with the fusion-off row of Table IX; H3–H6 additionally share the gate capacity of (3) and differ only in the extra cue. *Panel B* reports IoU by heterogeneity quartile and by object size. *Panel C* lists mean stage weights (shallow = w_1+w_2 , deep = w_3+w_4), with Spearman $\rho(\mathcal{H}, w_1+w_2) = 0.10$ over the 113 test patches; Fig. 6 shows the underlying maps and the pixel-wise distribution behind these two panels. The deep pair holds the larger share in every stratum, so the effect is a shift in share rather than a change of the dominant stage. As in Table III, Vaihingen scores include flip and rotation TTA, so the full model corresponds to 84.48%. The best and second-best entries in each column are **bold red** and underlined blue, respectively. Ranking applies to the mIoU column of *Panel A*, where Δ is the drop with respect to H6; *Panels B* and *C* are paired measurements rather than a competition.

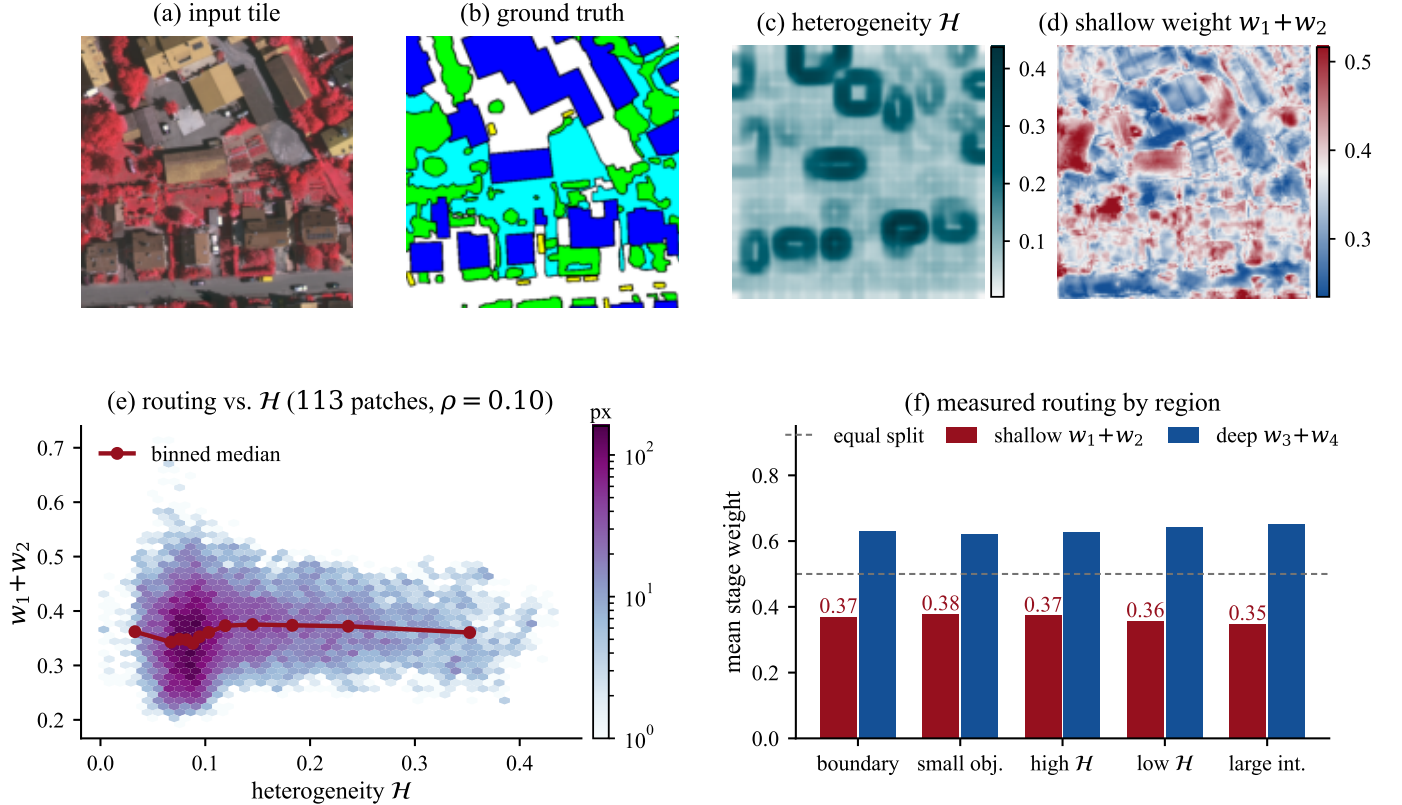


Fig. 6. Measured HG-SAF routing on ISPRS Vaihingen. (a) Input tile and (b) ground truth. (c) The heterogeneity statistic of (2) responds to object outlines and small structures. (d) Predicted shallow-stage weight w_1+w_2 of (3), with the colour range set to the 1–99 percentile of the tile so that the actual spatial variation is visible. (e) Pixel density of w_1+w_2 against $\mathcal{H}$ over the 113 test patches, with the binned median; the rank correlation over all pixels is $\rho=0.10$. (f) Mean stage weights per region, i.e. the quantities in Panel C of Table VI. The routing shift is consistent in sign across strata while deep stages keep the larger share everywhere.

its limit. The constraints of (5)–(7) apply to every coefficient alike, so the margin in Table VII follows from the bounded

low-rank parameterization rather than from a preference for any part of the spectrum. FRA is therefore effective as a

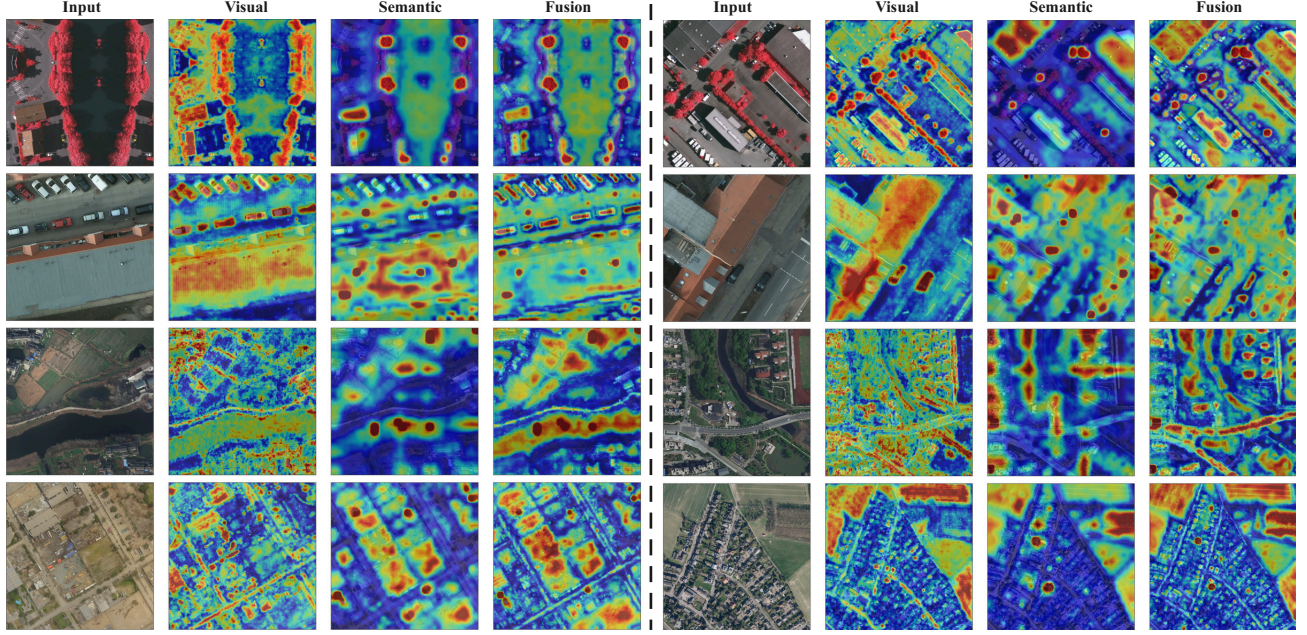


Fig. 7. Illustrative HG-SAF fusion on eight tiles (*Input*, shallow visual cues, deep semantic context, fused map). Quantitative routing statistics are in Table VI.

TABLE VII
MATCHED FREQUENCY-MODULE ALTERNATIVES ON ISPRS VAIHINGEN

Variant	Par.	mIoU	BloU	BF	Thin	Small
None (spatial only)	—	<u>84.31</u>	71.42	74.86	68.26	76.61
Spatial residual (matched)	1.1	84.22	71.78	75.14	68.58	76.84
Standard FFT filter	1.2	84.19	72.06	75.41	68.94	76.72
FFC-like branch [17]	3.4	84.08	71.64	75.02	68.41	76.38
GFNet mixer [16]	4.8	83.96	71.28	74.71	68.05	76.12
GFNet + zero-init res.	4.8	84.28	<u>72.54</u>	<u>75.88</u>	<u>69.42</u>	<u>77.05</u>
Full-rank spectral mixer	5.6	83.65	70.84	74.18	67.52	75.66
Unbounded gate	1.2	83.88	71.16	74.52	67.88	75.94
FRA	1.2	84.48	73.58	76.91	70.84	77.86

Every variant replaces FRA at the same fused-feature location, and Vaihingen scores include flip and rotation TTA as in Table III, so the first row coincides with the frequency-off row of Table IX. Par.: added parameters (M); BloU: Boundary IoU; BF: boundary F -score; Thin: thin-structure IoU; Small: small-component IoU. The best and second-best entries in each column are **bold red** and underlined blue, respectively. The parameter column reports cost and is not ranked.

conservative spectral residual, and not as a generic frequency-module replacement.

C. Tri-Prior Decoding

Table VIII compares the structural terms of (11) with conventional auxiliary losses at matched loss weights, and reports pairwise confusion mass on the frozen relation set. The question is whether the tri-prior combination improves on adding the same families of loss as independent multi-task terms.

The evidence that separates CATP from a generic multi-task head is Panel B: pairwise confusion mass decreases by 23–28% on every pre-declared pair of (9) that the panel reports, the largest reduction being OpenEarthMap *Rangeland* $\leftrightarrow$ *Agriculture*, which is exactly where the hinge applies pressure. Aggregate mIoU agrees: adding boundary, objectness or pixel-contrast losses individually improves the

TABLE VIII
CONFUSION-AWARE TRI-PRIOR DECODER ANALYSIS

Variant		mIoU	Δ
Panel A: matched auxiliary-loss controls (Vaihingen)			
S1	CE+FocalDice only	84.26	$\downarrow$ 0.22
S2	+ boundary loss [22]	84.28	$\downarrow$ 0.20
S3	+ objectness loss	84.29	$\downarrow$ 0.19
S4	+ pixel contrast [23]	84.31	$\downarrow$ 0.17
S5	all three conventional losses	<u>84.32</u>	$\downarrow$ 0.16
S6	CATP (feedback + pair hinge)	84.48	–

Pre-declared pair	w/o CATP	CATP	Rel.
Panel B: pairwise confusion mass $E_{a,b}$			
Vaih. Imp. Surf. $\leftrightarrow$ Building	0.142	0.108	–24%
Vaih. Low Veg. $\leftrightarrow$ Tree	0.118	0.091	–23%
Pots. Imp. Surf. $\leftrightarrow$ Building	0.096	0.074	–23%
LoveDA Background $\leftrightarrow$ Barren	0.186	0.141	–24%
OEM Rangeland $\leftrightarrow$ Agriculture	0.094	0.068	–28%

Panel A compares CATP with conventional auxiliary losses on ISPRS Vaihingen under matched loss weights and epochs, so row S1 coincides with the decoder-off row of Table IX; Vaihingen scores include flip and rotation TTA, as in Table III. Panel B reports the pairwise confusion mass $E_{a,b} = M_{ab} + M_{ba}$ on the training-only pair set; *Rel.* is the relative reduction against the matched model without CATP. OEM: OpenEarthMap. The best and second-best entries in each column are **bold red** and underlined blue, respectively. Ranking applies to the mIoU column of Panel A, where $\downarrow$ is the drop with respect to S6; a lower confusion mass is better throughout Panel B.

CE/FocalDice reference of (10) by at most 0.05 pp, and the three of them together remain 0.16 pp below CATP, so the gain comes from how the structural signals are coupled and not from the additional supervision alone. The next subsection tests whether the three stages remain useful when they are combined.

D. Component Complementarity and Initialization

Table IX trains all eight combinations of HG-SAF, FRA and CATP and compares conservative initialization with random and fully active starts. Combining the two analyses in one

TABLE IX
COMPONENT COMPLEMENTARITY AND INITIALIZATION STUDY

HG-SAF	FRA	CATP	Vaih.	Pots.	LoveDA	OEM	Avg.
<i>Panel A: factorial combinations</i>							
–	–	–	83.81	87.30	53.91	66.32	72.84
✓	–	–	84.04	87.65	54.46	66.85	73.25
–	✓	–	83.99	87.52	54.18	66.61	73.08
–	–	✓	84.19	87.78	54.58	66.94	73.37
✓	✓	–	84.26	87.93	54.81	67.27	73.57
✓	–	✓	<u>84.31</u>	<u>88.01</u>	<u>54.92</u>	<u>67.38</u>	<u>73.66</u>
–	✓	✓	84.28	87.94	54.71	67.12	73.51
✓	✓	✓	84.48	88.20	55.17	67.70	73.89
<i>Panel B: alternative initializations (Vaihingen)</i>							
Initialization	mIoU	std ₁₀	Best ep.	Div.			
Reference (ours)	84.48	0.42	67	0/3			
Small random	<u>84.19</u>	0.71	71	0/3			
Standard random	84.12	0.91	74	0/3			
Fully active	83.88	1.24	76	1/3			

Panel A lists the 2^3 independently trained combinations of HG-SAF, FRA, and CATP; its first row switches all three off, leaving mean fusion and a single semantic head on the prepared stages. mIoU increases along every path of the lattice on all four datasets, and pairwise interactions on Vaihingen are $I_{HF} = +0.04$, $I_{HC} = -0.11$ and $I_{FC} = -0.09$ pp, i.e., the modules are complementary and largely additive. Panel B reports initialization stability over three seeds, where std_{10} is the training-loss standard deviation over the first 10 epochs, *Best ep.* is the best epoch, and *Div.* counts diverged runs. ISPRS columns include flip and rotation TTA as in Table III; LoveDA and OpenEarthMap (OEM) use single-scale inference, so Avg. here mixes the two inference settings and is not the matched no-TTA mean of Table II; it is used only to compare rows within this table. The best and second-best entries in each column are **bold red** and underlined blue, respectively. Only the mIoU column is ranked in Panel B, because the remaining three columns are stability diagnostics for which a lower value is better.

table avoids separating architecture from optimization.

Each module is beneficial on its own, every pair improves on both of its members, and the full combination is best on all four datasets. Restoring the missing module to a pair adds +0.20 for HG-SAF, +0.17 for FRA and +0.22 pp for CATP on Vaihingen, and the pairwise interactions ($I_{HF} = +0.04$, $I_{HC} = -0.11$, $I_{FC} = -0.09$ pp, each formed as $m_{11} - m_{10} - m_{01} + m_{00}$) stay well below the main effects, so the three stages are complementary and largely additive. Reference initialization attains the highest mIoU, the lowest early-loss variance, and no diverged run. Starting from a known reference behavior therefore stabilizes fine-tuning and also improves its endpoint.

E. Fixed Preparation Blocks

Table X measures the fixed blocks of Section III-B on the reference decoder, with HG-SAF, FRA and CATP switched off.

SSDB, BCS-Mamba and SOE contribute +0.09, +0.14 and +0.07 pp individually and +0.22 pp jointly, so row T5 is the preparation stage adopted inside HAFR-Net. Two consequences are recorded explicitly. First, of the +0.55 pp that HAFR-Net gains over the reference decoder on Vaihingen in Table II, +0.22 pp comes from this fixed stage and +0.33 pp from the three refinement stages. Second, all module analyses in Tables VI–IX keep the preparation blocks fixed, so those comparisons are unaffected by this share.

TABLE X
INDIVIDUAL EFFECT OF THE FIXED PREPARATION BLOCKS

Preparation	mIoU	Δ
T1 none (reference decoder)	83.57	–
T2 SSDB only	83.66	+0.09
T3 BCS-Mamba only	<u>83.71</u>	+0.14
T4 SOE only	83.64	+0.07
T5 SSDB + BCS-Mamba + SOE	83.79	+0.22

ISPRS Vaihingen mIoU (%), single-scale and no TTA, so row T1 is exactly the Swin-B+UPerNet reference of Table II and the +0.22 pp of row T5 is the share of the no-TTA margin of Table II that the preparation blocks contribute. Only the preparation blocks change; HG-SAF, FRA and CATP are absent in all five rows. The best and second-best entries in each column are **bold red** and underlined blue, respectively.

F. Qualitative Comparison

Figs. 8–10 close the analysis with representative predictions on the four benchmarks. Every map shown, including those of the comparison methods, is produced by a model we trained ourselves and decoded with the same inference code; no prediction is reproduced from a published figure. Each observation below belongs to a stratum that the analyses above already quantify, so the figures show where those measured effects become visible. The same three tendencies recur across the four datasets. Thin elongated structures stay connected over longer spans, adjacent instances of one class are merged into a single region less often, and the two classes of a pre-declared relation pair leak into each other less. The regions we discuss are marked in each figure, and the errors that survive are collected in Section VII-B.

The ISPRS examples in Fig. 8 concern vehicles adjacent to impervious surfaces and building outlines, which are the two regimes that the HG-SAF and CATP analyses measure separately. Vehicles are the smallest annotated class in these benchmarks and fall almost entirely inside the small-object stratum of Table VI, where stratified IoU improves by +1.52 pp in Panel B and Panel C shows the largest shift toward the shallow stages; in the marked parking rows the neighbouring vehicles stay separated rather than merging into one region. Building outlines are the opposite case: long boundaries between two classes with overlapping colour statistics, i.e., the pair whose confusion mass CATP reduces in Panel B of Table VIII. In the marked regions the predicted outline follows the annotated corner more closely. Neither observation is specific to these tiles: both are consistent with the strata of Table VI and the pair panel of Table VIII, and the boundary and small-component metrics of Table VII improve alongside mIoU rather than in place of it.

The LoveDA examples in Fig. 9 highlight thin rural roads and low-texture *Barren* patches. Thin roads are exactly the regime that thin-structure IoU in Table VII quantifies, and in the marked regions the road stays continuous where its surface is partly occluded instead of breaking into fragments. *Barren* is the harder case, because it is characterized less by a texture than by the absence of one, so a decoder that leans on local appearance tends to absorb it into *Background*; that is the pair whose confusion mass CATP reduces in Panel B of Table VIII, and the marked patches keep one label over a larger area.

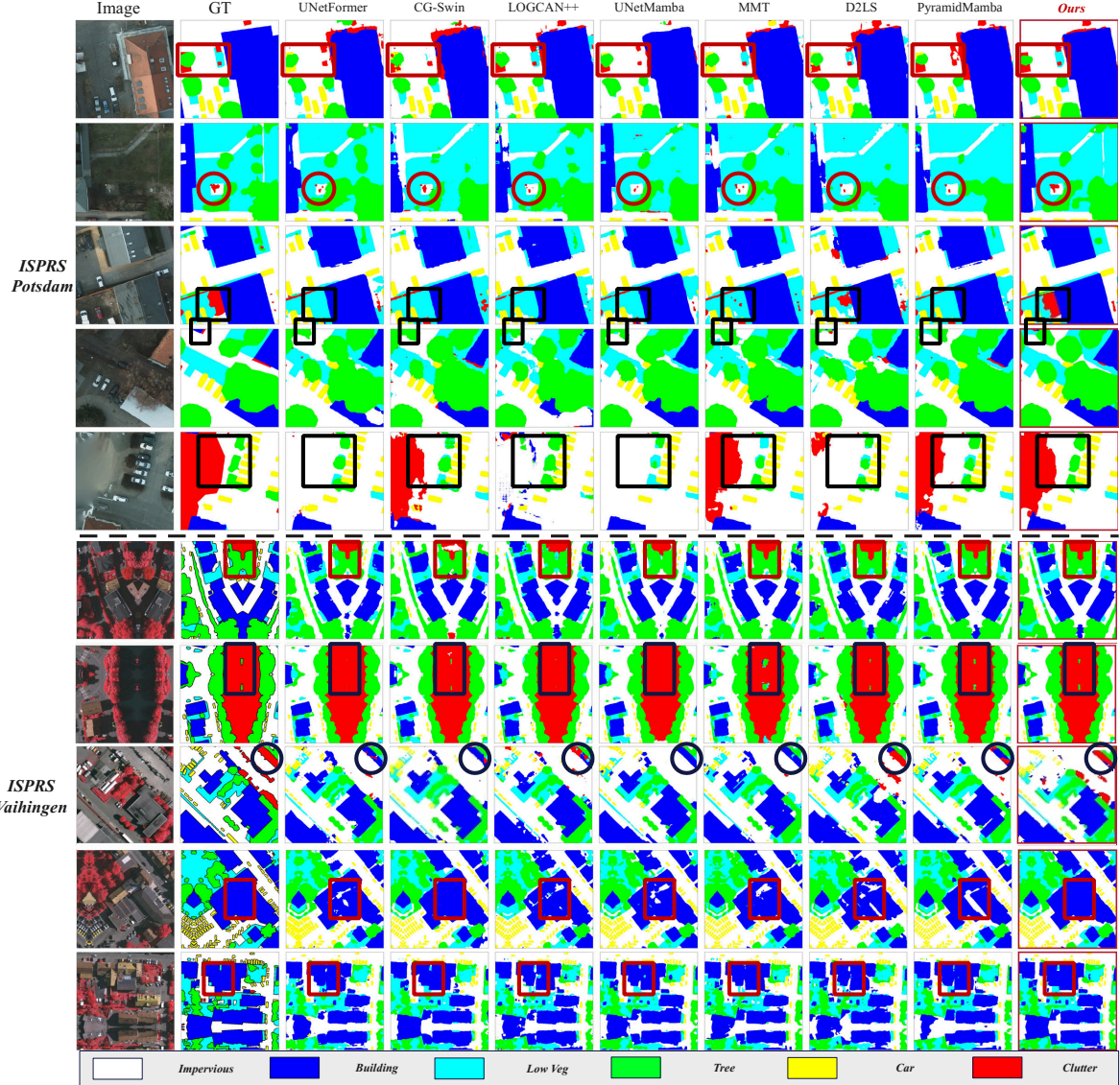


Fig. 8. Representative ISPRS examples. Boxes mark vehicles and adjacent class boundaries discussed with the HG-SAF and CATP analyses.

LoveDA also accounts for part of the dataset-level pattern in Table II: its tiles mix urban and rural domains, so a larger share of pixels falls in the upper heterogeneity quartiles, which is where Panel B of Table VI reports the largest stratified gains.

The OpenEarthMap examples in Fig. 10 show *Range-land* $\leftrightarrow$ *Agriculture* transitions and thin roads. That pair is the pre-declared relation of this dataset and the one with the largest relative reduction of confusion mass in Panel B of Table VIII. Its transition is gradual rather than sharp, so the decision is where to cut a broad gradient rather than where to find a visible edge, and in the marked regions the predicted boundary follows the field outline instead of oscillating inside the gradient. OpenEarthMap carries the largest number of classes and the widest geographic spread of the four benchmarks, which is consistent with it also showing the largest matched-protocol margin in Table II. Taken together, the four figures locate the measured gains in small objects, thin structures

and pre-declared class pairs, and they do not show a uniform improvement over large homogeneous interiors, which is what Panel B of Table VI already reports numerically.

The next section reports the compute cost of the full model and the remaining limitations.

VII. EFFICIENCY AND LIMITATIONS

A. Efficiency

Accuracy gains are useful only if their cost is stated under the same protocol as the controlled comparison. Table XI reports parameters, GFLOPs, peak memory, latency and FPS for the re-implemented Swin-B family. Published mixed-protocol models are omitted from that table.

HAFR-Net adds 8.5 M parameters over Swin-B+UPerNet. The FP32 FFT of FRA costs 2.9 ms per 512×512 tile, i.e., about 12% of the 24.8 ms forward pass of Table XI. The accuracy gain in Table II is therefore obtained at a moderate,

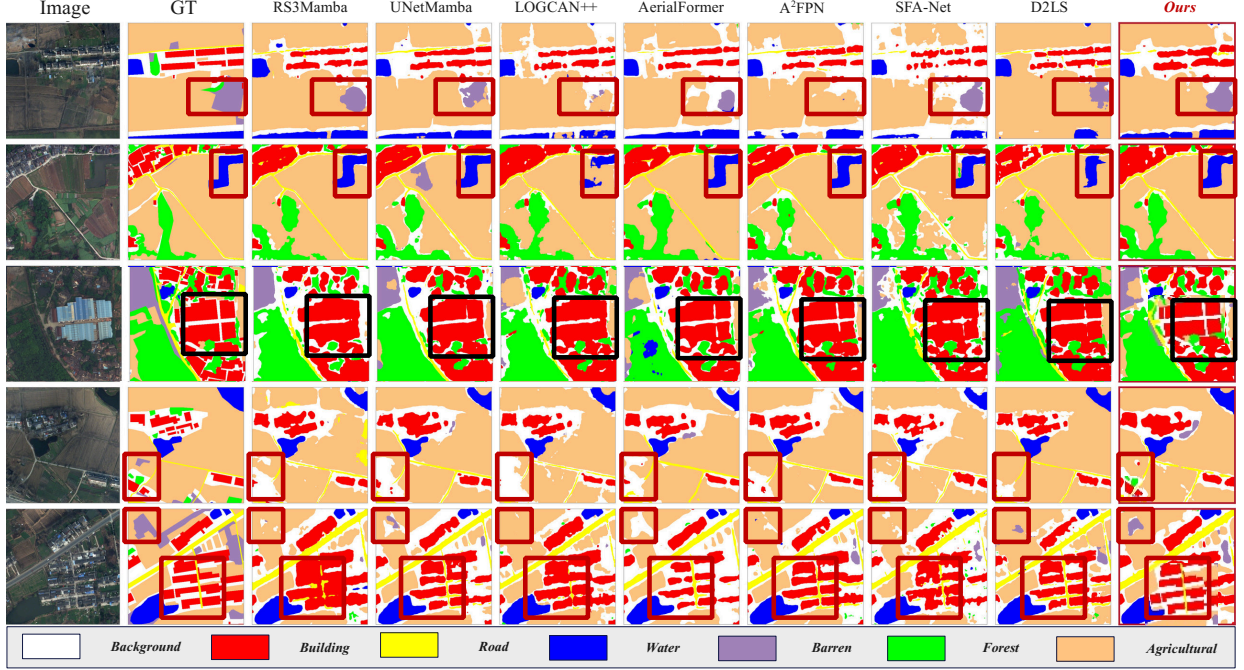


Fig. 9. Representative LoveDA examples. Boxes mark thin rural roads and low-texture *Barren* regions.

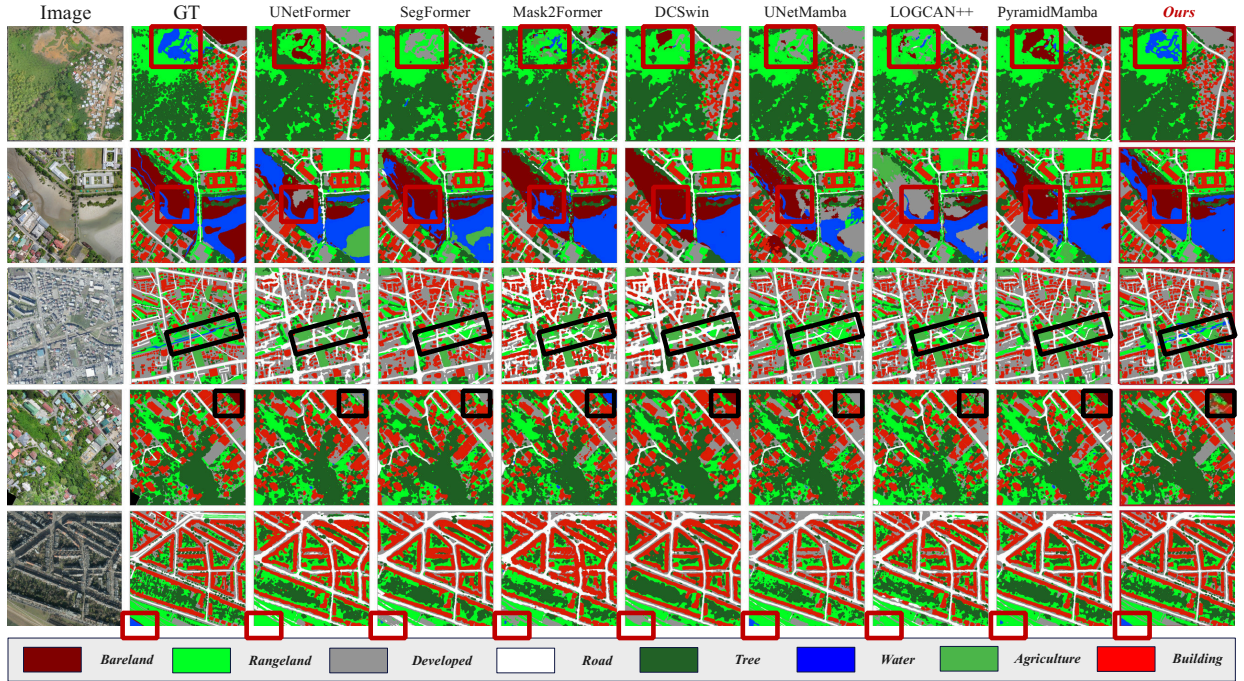


Fig. 10. Representative OpenEarthMap examples. Boxes mark *Rangeland* ↔ *Agriculture* transitions and thin roads.

fully specified compute overhead. Fig. 11 places the same mean next to published literature values.

Stating the same cost as a ratio against the reference makes the trade-off easier to compare with other decoders. Relative to Swin-B+UPerNet, HAFR-Net adds 9.5% parameters, 10.8% GFLOPs and 15.2% peak memory, and it lowers throughput

from 46.7 to 40.3FPS, i.e., by 13.7%, in exchange for +1.22pp of four-dataset mean mIoU without TTA. Latency grows somewhat faster than GFLOPs (15.9% against 10.8%), which is consistent with the FP32 FFT of FRA being bound by memory traffic rather than by arithmetic; the ordering of the five methods in Table XI is the same under either measure. We

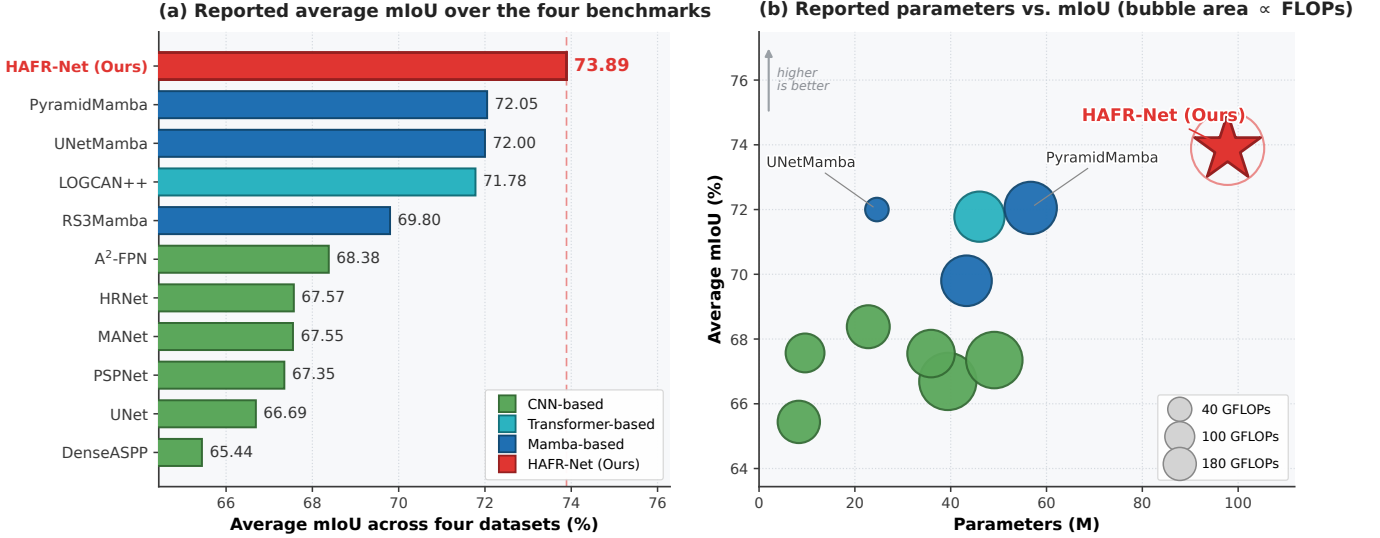


Fig. 11. Published accuracy and cost, shown as context. Panel (a) gives the four-dataset mean mIoU of the methods present in all three contextual tables; for HAFR-Net this is the 73.89% mean, which uses TTA on the two ISPRS sets. Panel (b) plots the reported parameter count against that mean, with bubble area proportional to the reported GFLOPs. The entries come from different backbones and inference settings, so no frontier or optimality is drawn: the matched comparison is Table II and its measured cost Table XI.

TABLE XI
EFFICIENCY OF FIVE RE-IMPLEMENTED SWIN-B DECODERS

Method	Par.	GFLOPs	Mem.	Lat.	FPS	mIoU
Swin-B+FPN	90.1	102.6	1.72	18.9	52.9	72.09
Swin-B+UPerNet	89.3	118.4	1.84	21.4	46.7	72.49
Swin-B+UNetFormer dec.	92.6	124.1	1.91	22.6	44.2	<u>72.78</u>
Swin-B+FFT mixer	93.8	129.7	1.98	23.5	42.6	72.56
HAFR-Net	97.8	131.2	2.12	24.8	40.3	73.71

Cost is measured on one RTX 4090 under the timing protocol of Section IV-D, with TTA off. Par.: parameters (M); Mem.: peak memory (GB); Lat.: latency (ms); mIoU is the four-dataset mean of Table II without TTA. The best and second-best entries in each column are **bold red** and underlined blue, respectively. Only mIoU is ranked, because HAFR-Net is the most expensive entry in every cost column.

report FPS at batch size 1, the setting in which that overhead is least amortized. At 512×512 the full model still needs 24.8 ms per tile on one RTX 4090, so tiled inference over a large scene stays bounded by tiling and mosaicking rather than by the decoder.

Within Table XI the closest cost neighbour is the matched FFT mixer at 23.5 ms and 93.8M parameters: HAFR-Net spends 1.3 ms and 4.0M more and returns +1.15 pp of four-dataset mean mIoU. The next subsection lists the cases that this cost does not resolve.

B. Limitations

Four limitations remain. First, $\mathcal{H}$ is a fixed functional of the learned feature $\hat{F}_4$, namely the mean per-channel local standard deviation of (2). It adds no parameters of its own, but it inherits whatever $\hat{F}_4$ encodes, so a semantically homogeneous yet feature-variable region such as tree canopy can produce a boundary-like response. The measured association between $\mathcal{H}$ and the shallow-stage weight is correspondingly weak (Fig. 6(e)), and a learned heterogeneity measure is left to future work. Second, the spectral branches of SSDB and

FRA evaluate their FFTs in FP32 for numerical stability, which is the latency reported in Section VII-A. A half-precision or windowed spatial approximation would remove that cost, but we have not verified that it preserves the boundary metrics of Table VII, so we report the conservative variant. Third, the prototype relation set depends on a training-only pilot split and does not adapt online, so a pair that becomes confusable only in a deployment domain receives no hinge; the number of candidate pairs also grows quadratically with the number of classes, which is why the set is declared in advance rather than searched. Fourth, all experiments use RGB or NIR-RG VHR imagery and have not been extended systematically to SAR or hyperspectral data, where the spectral assumptions behind SSDB and FRA would have to be re-examined rather than transferred. Residual errors that are not captured by hierarchical fusion, bounded refinement, or structural regularization—including cast shadows, mixed vegetation, and the catch-all *Clutter* class—also remain difficult, and they dominate the marked regions of Figs. 8–10 that HAFR-Net does not resolve.

VIII. CONCLUSION

This work studies VHR remote sensing image segmentation from the perspective of how pretrained hierarchical representations are decoded and refined. Rather than adding unconstrained decoder capacity, we propose HAFR-Net, which organizes the capacity it does add as a progressive refinement of the encoder hierarchy in three stages. HG-SAF performs heterogeneity-conditioned pixel-wise fusion of multi-stage features, FRA introduces a bounded residual frequency correction inside a channel bottleneck, and CATP regularizes the final prediction with structural and class-relation cues. Controlled experiments across four VHR benchmarks show gains of +0.33 to +1.52 pp mIoU over the strongest matched baseline at a moderate and fully reported compute overhead, while module-specific analyses demonstrate spatially varying stage

routing, improved boundary and thin-structure prediction, and reduced structural and class-confusion errors. Adaptive and conservative refinement of strong pretrained representations is therefore an effective design principle for accurate VHR segmentation. Future work will investigate learned heterogeneity measures, adaptive relation discovery, and other remote sensing modalities.

ACKNOWLEDGMENT

The authors would like to thank the ISPRS, LoveDA, and OpenEarthMap organizers for releasing their benchmark datasets.

REFERENCES

- [1] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Syst. Appl.*, vol. 169, p. 114417, 2021.
- [2] L. Huang, B. Jiang, S. Lv, Y. Liu, and Y. Fu, "Deep-learning-based semantic segmentation of remote sensing images: A survey," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 8370–8396, 2024.
- [3] R. Gupta, B. Goodman, N. Patel, R. Hosfelt, S. Sajeev, E. Heim, J. Doshi, K. Lucas, H. Choset, and M. Gaston, "Creating xBD: A dataset for assessing building damage from satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2019, pp. 10–17.
- [4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3431–3440.
- [5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.
- [6] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2881–2890.
- [7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
- [8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10 012–10 022.
- [9] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 12 077–12 090.
- [10] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *Proc. Conf. Lang. Model. (COLM)*, 2024.
- [11] L. Wang, D. Li, S. Dong, X. Meng, X. Zhang, and D. Hong, "Pyramid-Mamba: Rethinking pyramid feature fusion with selective space state model for semantic segmentation of remote sensing imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 144, p. 104884, 2025.
- [12] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2117–2125.
- [13] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 418–434.
- [14] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 196–214, 2022.
- [15] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, art. no. 6506105.
- [16] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 980–993.
- [17] L. Chi, B. Jiang, and Y. Mu, "Fast Fourier convolution," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 4479–4488.
- [18] Y. Yang, G. Yuan, and J. Li, "SFFNet: A wavelet-based spatial and frequency domain fusion network for remote sensing segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–17, 2024, art. no. 3000617.
- [19] X. Li, F. Xu, A. Yu, X. Lyu, H. Gao, and J. Zhou, "A frequency decoupling network for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–21, 2025, art. no. 5607921.
- [20] F. Gao, M. Fu, J. Cao, J. Dong, and Q. Du, "Adaptive frequency enhancement network for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–15, 2025, art. no. 5619415.
- [21] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape CNNs for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 5229–5238.
- [22] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *Proc. Int. Conf. on Medical Imaging with Deep Learning (MIDL)*, 2019, pp. 285–296.
- [23] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, "Exploring cross-image pixel contrast for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 7303–7313.
- [24] T. Zhou, W. Wang, E. Konukoglu, and L. Van Gool, "Rethinking semantic segmentation: A prototype view," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 2582–2593.
- [25] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, 2021.
- [26] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 181, pp. 84–98, 2021.
- [27] L. Wang, R. Li, D. Wang, C. Duan, T. Wang, and X. Meng, "Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images," *Remote Sens.*, vol. 13, no. 16, p. 3065, 2021.
- [28] R. Li, S. Zheng, C. Zhang, C. Duan, J. Su, L. Wang, and P. M. Atkinson, "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022, art. no. 5607713.
- [29] R. Li, L. Wang, C. Zhang, C. Duan, and S. Zheng, "A²-FPN for semantic segmentation of fine-resolution remotely sensed images," *Int. J. Remote Sens.*, vol. 43, no. 3, pp. 1131–1155, 2022.
- [30] L. Wang and R. Li, "FT-UNetFormer implementation in the GeoSeg toolkit," [Online]. Available: <https://github.com/WangLibo1995/GeoSeg>, 2022, accessed: Jun. 13, 2026.
- [31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141.
- [32] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
- [33] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontañón, "FNet: Mixing tokens with Fourier transforms," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL)*, 2022, pp. 4296–4313.
- [34] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, "Fourier neural operator for parametric partial differential equations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
- [35] H. Zhang, G. Xie, L. Li, X. Xie, and J. Ren, "Frequency-domain guided Swin Transformer and global-local feature integration for remote sensing images semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–11, 2025, art. no. 5612611.
- [36] X. Li, F. Xu, J. Zhang, H. Zhang, X. Lyu, F. Liu, H. Gao, and A. Kaup, "Frequency-guided denoising network for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 64, pp. 1–17, 2026, art. no. 5400217.
- [37] Y. Wang, L. Cao, and H. Deng, "WgANet: A wavelet-guided attention network for remote sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 19, pp. 702–716, 2026.
- [38] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 5872–5881.
- [39] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2980–2988.

- [40] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. Int. Conf. 3D Vision (3DV)*, 2016, pp. 565–571.
- [41] M. Berman, A. R. Triki, and M. B. Blaschko, "The lovasz-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4413–4421.
- [42] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 1290–1299.
- [43] X. Ma, X. Zhang, and M.-O. Pun, "RS3Mamba: Visual state space model for remote sensing image semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024, art. no. 6011405.
- [44] E. Zhu, Z. Chen, D. Wang, H. Shi, X. Liu, and L. Wang, "UNetMamba: An efficient UNet-like Mamba for semantic segmentation of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 22, pp. 1–5, 2025, art. no. 6001205.
- [45] X. Ma, X. Zhang, M.-O. Pun, and B. Huang, "A unified framework with multimodal fine-tuning for remote sensing semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–15, 2025.
- [46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255.
- [47] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 506–516.
- [48] N. Houlsby, A. Giurghi, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 2790–2799.
- [49] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 32–42.
- [50] ISPRS, "ISPRS 2D semantic labeling contest — Vaihingen," [Online]. Available: <https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx>, accessed: Jun. 13, 2026.
- [51] —, "ISPRS 2D semantic labeling contest — Potsdam," [Online]. Available: <https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx>, accessed: Jun. 13, 2026.
- [52] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in *Proc. NeurIPS Datasets and Benchmarks Track*, 2021.
- [53] J. Xia, N. Yokoya, B. Adriano, and C. Broni-Bediako, "OpenEarthMap: A benchmark dataset for global high-resolution land cover mapping," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2023, pp. 6254–6264.
- [54] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 8024–8035.
- [55] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.
- [56] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch SGD: Training ImageNet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.
- [57] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1195–1204.
- [58] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, "Boundary IoU: Improving object-centric image segmentation evaluation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 15 334–15 342.
- [59] G. Csúrká, D. Larlus, and F. Perronnin, "What is a good evaluation measure for semantic segmentation?" in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2013, pp. 32.1–32.11.
- [60] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 3684–3692.
- [61] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "CGNet: A light-weight context guided network for semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 1169–1179, 2021.
- [62] X. Ma, R. Lian, Z. Wu, H. Guo, F. Yang, M. Ma, S. Wu, Z. Du, W. Zhang, and S. Song, "LOGCAN++: Adaptive local-global class-aware network for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–16, 2025.
- [63] X. Meng, Y. Yang, L. Wang, T. Wang, R. Li, and C. Zhang, "Class-guided Swin Transformer for semantic segmentation of remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, art. no. 6517505.
- [64] Z. Xu, J. Geng, and W. Jiang, "MMT: Mixed-mask transformer for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023, art. no. 5613415.
- [65] X. Zou, Y. Li, S. Zhang, K. Li, S. Wang, P. Tao, J. Xing, and C. Lang, "Dynamic dictionary learning for remote sensing image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025, pp. 22 457–22 466.
- [66] T. Hanyu, K. Yamazaki, M. Tran, R. A. McCann, H. Liao, C. Rainwater, M. Adkins, J. Cothren, and N. Le, "AerialFormer: Multi-resolution transformer for aerial image segmentation," *Remote Sens.*, vol. 16, p. 2930, 2024.
- [67] G. Hwang, J. Jeong, and S. J. Lee, "SFA-Net: Semantic feature adjustment network for remote sensing image segmentation," *Remote Sens.*, vol. 16, no. 17, p. 3278, 2024.