CRISP: Calibration-Aware Visual State Space Duality for Remote Sensing Semantic Segmentation

Kangning Wang[1,2], Haopeng Zhang[1,3,4⋆], and Zhiguo Jiang[1]

[1] Tianmushan Laboratory, Beihang University, Hangzhou 311115, China
[2] Hangzhou International Innovation Institute, Beihang University, Hangzhou 311115, China
[3] School of Astronautics, Beihang University, Beijing 102206, China
zhanghaopeng@buaa.edu.cn
[4] Key Laboratory of Spacecraft Design Optimization and Dynamic Simulation Technology,
Ministry of Education, Beijing 102206, China

Abstract. State space models, especially Visual State Space Duality (VSSD), have emerged as efficient linear-time alternatives to Transformers for dense visual tasks. However, we observe that VSSD compresses spatial context into a global aggregation that suppresses high-frequency responses, causing excessive boundary smoothing in remote sensing semantic segmentation. To address this, we propose CRISP, a calibration framework with two components. Its core, the Duality Calibration Operator (DCO), restores local contrast and boundary responses through residual injection and frequency calibration within the VSSD backbone, without altering its linear complexity. To retain the recovered detail, an Orthogonal Multi-Prototype (OMP) head assigns multiple orthogonally constrained prototypes per class to model large intra-class variance. Extensive experiments on Potsdam, Vaihingen, and LoveDA show that, with only ∼30M parameters, CRISP achieves consistent gains in mean F1 (mF) and mIoU while remaining competitive with state-of-the-art methods. Code is available at https://github.com/crazylifeha/crisp.

Keywords: Remote sensing semantic segmentation · state space models · frequency calibration · feature calibration · detail preservation

1 Introduction

High-resolution remote sensing image semantic segmentation is a fundamental task in Earth observation, vital for urban planning, disaster monitoring, and environmental assessment [5,47]. Unlike natural images, remote sensing imagery features exceedingly high spatial resolutions with dense, small-scale structures and complex category boundaries. Convolutional Neural Networks (CNNs) have

⋆ Corresponding author.

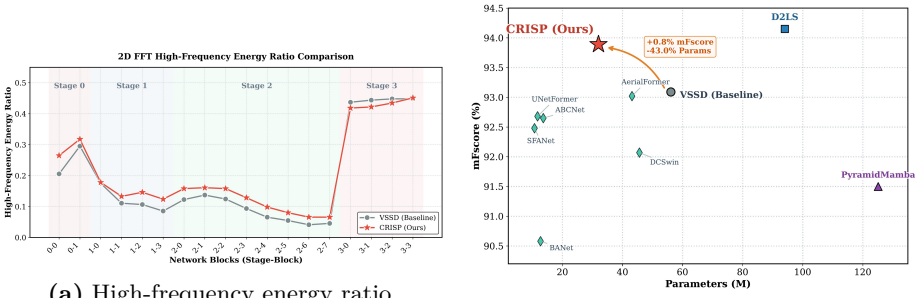
been widely adopted for their local representation capability [11,41], and Vision Transformers (ViTs) further advanced accuracy through long-range dependency modeling [8,36]. However, the quadratic complexity of self-attention in sequence length [27] causes severe memory and computation bottlenecks at the ultra-high resolutions typical of remote sensing.

Recently, State Space Models (SSMs), particularly the selective state space-based Mamba [13] and its upgraded version Mamba-2, which introduced State Space Duality (SSD) [7], have emerged as a promising alternative. SSD theoretically unifies sequence modeling and attention mechanisms; however, its causal scan assumes a single ordering of the input, which does not naturally fit 2D images. To alleviate this, recent studies proposed the Visual State Space Duality (VSSD) model [34]. Rather than a single causal pass, VSSD performs *multi-directional scanning* and fuses the results into a global, order-agnostic aggregation, transferring the linear-complexity advantage of SSD to visual feature extraction. Nevertheless, our empirical studies reveal that this global aggregation tends to be unstable when characterizing boundaries and small-scale objects in remote sensing images enriched with complex high-frequency details. To quantify this phenomenon, we evaluate the baseline from two perspectives: (i) boundary-sensitive metrics (Boundary F-score), and (ii) feature spectral energy ratio ($E_{\text{high}}/E_{\text{low}}$, executing 2D FFT on intermediate backbone layers). As shown in Fig. 1(a), the high-frequency energy ratio decays progressively across backbone layers, confirming the systematic suppression of high-frequency responses caused by the global normalized aggregation. Addressing this suppression, as we show later, lets CRISP reach a favorable performance–complexity trade-off (Fig. 1(b)).

To address this without compromising VSSD's linear complexity, we propose CRISP, which introduces an aggregation-aligned calibration branch for targeted compensation rather than disrupting the normalized aggregation structure. CRISP couples a Duality Calibration Operator (DCO) in the backbone with an Orthogonal Multi-Prototype (OMP) head in the decoder, which jointly preserve intra-class multi-modal structures from pixel-level features to the final classification.

In summary, the main contributions of this paper are as follows:

- We propose the Duality Calibration Operator (DCO), the first residual calibration for the VSSD aggregation, which reuses the backbone's aggregation weights to recover attenuated high-frequency responses while preserving linear complexity.
- We further introduce the Orthogonal Multi-Prototype (OMP) head, which learns orthogonally regularized sub-prototypes end-to-end to preserve the intra-class multi-modal structure recovered by DCO.
- We demonstrate that CRISP attains accuracy on par with state-of-the-art methods on Potsdam, Vaihingen, and LoveDA at only ∼30M parameters, achieving a markedly better accuracy–efficiency trade-off than prior methods.



(a) High-frequency energy ratio.

(b) Performance–complexity tradeoff on Potsdam.

Fig. 1: Feature smoothing and efficiency analysis. (a) High-frequency energy ratio across backbone layers. (b) Performance–complexity tradeoff on Potsdam.

2 Related Work

We organize prior work along four axes that together situate CRISP: (i) remote sensing semantic segmentation, (ii) visual state space models, (iii) frequency- and boundary-preserving modeling, and (iv) prototype-based decoders. The first two establish the backbone landscape CRISP builds on; the last two delineate the specific gap CRISP fills.

2.1 Semantic Segmentation in Remote Sensing

Remote sensing semantic segmentation assigns a category to every pixel of high-resolution overhead imagery, where dense small objects and intricate boundaries make the preservation of fine spatial detail essential. Early CNN-based methods enlarge the effective receptive field while retaining locality: DANet [11] couples spatial and channel attention to model long-range context, OCR [48] aggregates object-contextual representations, and UNetFormer [41] attaches a lightweight Transformer decoder to a CNN encoder for real-time urban-scene parsing. With Vision Transformers, global dependency modeling improved accuracy—from plain ViT [8] and SegFormer [43] to hierarchical Swin [27] and mask-classification heads such as MaskFormer [4]—but at quadratic cost. To recover locality, hybrid CNN–Transformer designs [9] interleave convolution and self-attention. The shared limitation of these Transformer-based backbones is computational: the quadratic cost of self-attention in sequence length becomes memory-prohibitive at the ultra-high resolutions typical of remote sensing, which is precisely what motivates linear-complexity state-space alternatives.

2.2 Visual State Space Models

State Space Models (SSMs) originate in control theory: S4 [14] models long sequences with a structured linear-time recurrence, and Mamba [13] makes the recurrence input-dependent (selective) so it can route information content-adaptively.

Mamba-2 [7] establishes State Space Duality (SSD), showing this recurrence is equivalent to a masked linear-attention form; this connects SSMs to the broader linear-attention family [15,19]. Adapting SSMs to 2D images hinges on the *scanning strategy* used to impose an order on pixels. Vision Mamba [51] applies a bidirectional scan, VMamba [26] a four-directional cross-scan, PlainMamba [44] a continuous space-filling scan, and EfficientVMamba [31] an atrous selective scan for efficiency. Remote sensing backbones follow the same paradigm: RS-Mamba [2] and RS3Mamba [29] tailor the scan paths of selective Mamba to overhead imagery, while CVMH-UNet [1] even adds an explicit multi-frequency, multi-scale fusion module to counter detail loss—yet this remains an external add-on bolted onto an otherwise causal backbone. All of these variants still run *causal* recurrences along each direction and approximate global context by summing several scans. VSSD [34] departs from this line: instead of relying on a single per-direction causal ordering, it fuses multi-directional responses into a single global, position-agnostic aggregation, thereby reducing scan redundancy and roughly halving the scanning cost. We build on VSSD as our backbone for this reason. Crucially, however, neither line examines the *spectral* side effect of the global aggregation itself: the very step that yields linear cost acts as a low-pass operator, collapsing the directional scans into a near-uniform global summary and attenuating the high-frequency responses that encode boundaries and small objects. CRISP targets exactly this aggregation step, and is orthogonal to the choice of scan path or backbone topology.

2.3 Frequency- and Boundary-Preserving Modeling

To counter over-smoothing and recover lost high-frequency detail, prior methods can be grouped by *where* they reinject the high-frequency signal. (i) *Spatial post-processing* sharpens the prediction after the backbone: PointRend [20] adaptively re-samples uncertain boundary points to render crisper masks. Such methods correct the output mask but leave the already-smoothed backbone features untouched. (ii) *Frequency-domain operators* reinject high frequencies through a parallel spectral path: FreqMamba [53] couples Mamba with a frequency-domain branch for deraining, SpectralMamba [45] performs spectral-domain scanning for hyperspectral data, and generic spectral modules such as Fast Fourier Convolution [6] and Global Filter Networks [32] learn filters directly in the Fourier domain. Closer to dense prediction, FreqFusion [24] performs frequency-aware feature fusion on the decoder side to sharpen boundaries. (iii) *Locality restoration* reintroduces local detail in the spatial domain: LocalMamba [17] adds windowed selective scans and parallel convolutions to recover the fine structure suppressed by global mixing. (iv) *Feature calibration* corrects or re-weights intermediate features instead of adding a spectral path: SFC [50] shares feature calibration across classes for weakly-supervised segmentation, and CSFCal [21] jointly calibrates context and spatial features for real-time segmentation. CRISP belongs to this calibration family in spirit, but is the first to calibrate the SSM aggregation itself. Effective as they are, these methods share two costs: they attach a *separate* spectral branch, decoder-side fusion module, or convolutional path, adding

parameters and FLOPs, and they break the hardware-friendly continuity of the linear SSM scan. CRISP differs in both *where* and *what* it calibrates. Rather than appending a generic Fourier/DCT/wavelet module or a decoder-side fusion stage, our DCO calibrates the VSSD directional aggregation *in place*: it reuses the aggregation weights already computed by the backbone to inject a state-level residual that compensates the discretization-induced aliasing, restoring high-frequency responses while preserving $\mathcal{O}(L)$ complexity and the native scan structure.

2.4 Prototype-Based Decoders

The decoder determines whether recovered fine detail survives to the final logits. Conventional heads such as DeepLabv3+ [3] and UPerNet [42] are effectively *single-prototype* classifiers: each class is represented by one weight vector, so a pixel is labeled by its similarity to that single center. This collapses intra-class variation onto one direction and is ill-suited to remote sensing, where the same category (*e.g.*, buildings with different roof materials) exhibits very different spectral signatures. Multi-center decoders relax this assumption. Center-guided classification [49] maintains several class centers for remote sensing, NAPG [10] groups neighborhood-assisted multi-prototypes, and prototypical contrastive methods [46] learn multiple prototypes per class. These approaches, however, typically obtain their centers via offline clustering or auxiliary contrastive training, which complicates end-to-end optimization and couples the decoder to a separate clustering stage. Our Orthogonal Multi-Prototype (OMP) head instead learns K class-wise sub-prototypes *end-to-end*. Building on ideas from prototypical learning [35], orthogonal-prototype projection [25], and generalized few-shot remote sensing [23], OMP combines an intra-class orthogonality constraint with an inter-class angular margin and a mutual-information objective to keep the sub-prototypes diverse and non-redundant. This lets OMP preserve the fine-grained, multi-modal structure that DCO restores, without any offline clustering.

3 Methodology

In this section, we formulate the proposed CRISP (Calibration-aware Visual State Space Duality) framework for remote sensing semantic segmentation. First, in Section 3.1, we briefly review the State Space Duality (SSD) proposed in Mamba-2 and the mechanism causing feature information attenuation. Using its global-aggregation approximation as a starting point, we subsequently discuss the Duality Calibration Operator (DCO) in Section 3.2, analyzing how it resolves spectral bias through a three-level hierarchical structure. Finally, in Section 3.3, we present the Orthogonal Multi-Prototype (OMP) head, demonstrating how it adapts to dispersed high-dimensional local characteristics via penalty regularizations.

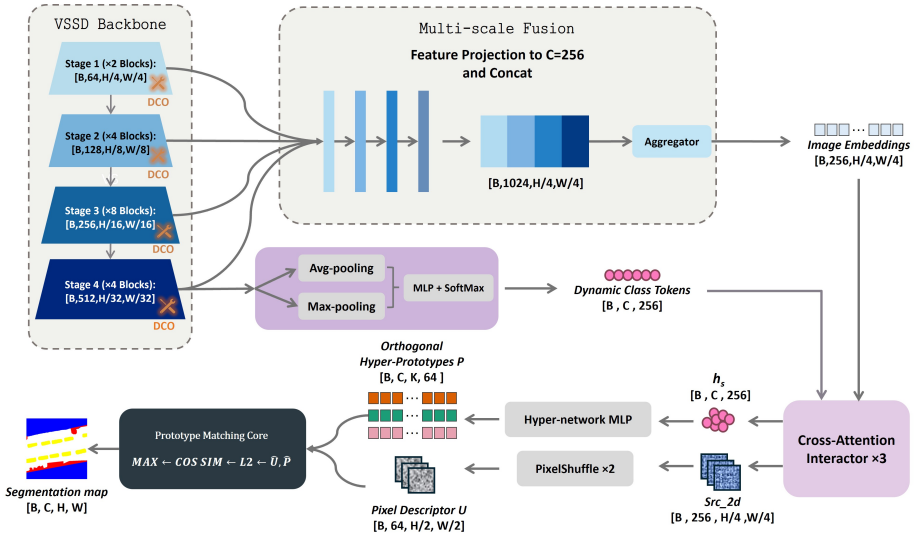


Fig. 2: Overall CRISP architecture. DCO calibrates high-frequency responses inside the VSSD backbone, while OMP performs dynamic token generation and multi-prototype matching.

3.1 Mamba-2, VSSD, and Feature Smoothing Analysis

Because state space models for vision are still relatively recent, we first give an intuitive picture before any formalism. A state space model (SSM) processes a sequence of tokens one position at a time, carrying a small running summary (a *hidden state*) that is updated as it goes; reading this state out gives the output at each position. This makes the cost grow *linearly* with the number of tokens, in contrast to self-attention, whose all-pairs comparison grows quadratically—which is why SSMs are attractive for the long token sequences produced by high-resolution remote sensing images. The Mamba family makes this update content-aware (so the model can decide what to keep or forget), and Mamba-2 shows that the whole process can be rewritten as a form of attention with a distance-based decay, which is both faster to compute and easier to reason about. The catch when moving to images is that pixels have no natural reading order, so visual variants such as VSSD scan the image along several directions and then *fuse* the results into a single position-agnostic summary. As we make precise below, this fusion behaves like a large smoothing window: it captures global context well but averages away the fine high-frequency detail (edges, thin structures) that dense remote sensing segmentation depends on. The rest of this subsection formalizes this mechanism and pinpoints where the high-frequency loss originates, directly motivating our calibration design.

Background: from selective SSMs to SSD. A state space model maps an input token sequence $x_1, \ldots, x_L$ to outputs through a latent state h_i that is updated recurrently, $h_i = \bar{A}h_{i-1} + \bar{B}x_i$, $y_i = C^\top h_i$, where $(\bar{A}, \bar{B})$ are the dis-

cretized state-transition and input matrices and C reads out the state. Selective SSMs (Mamba) [13] make B, C, and the discretization step *input-dependent*, so that the model can route information selectively rather than with fixed dynamics. Mamba-2 [7] introduces State Space Duality (SSD), showing that this recurrence is mathematically equivalent to a masked linear-attention form, which makes the connection to attention explicit and enables efficient parallel computation. We use this attention view below because it exposes *how* the spatial mixing weights are formed, which is where over-smoothing originates.

In this dual form, the parameters $A, B, C \in \mathbb{R}^{L\times N}$ are produced from the input token features $X \in \mathbb{R}^{L\times D}$ by three separate linear projection layers, $B = XW_B$, $C = XW_C$, and the (input-dependent part of the) state decay $A = XW_A$, with learnable W_A, W_B, W_C; the values $V = XW_V \in \mathbb{R}^{L\times D}$ are likewise a linear projection of the input. Thus B/C play the role of keys/queries and A controls the distance decay. Given these, the sequential interactions in SSD can be uniformly expressed as a weighted sum in a prefix-scan kernel format. Specifically, the output feature at the i-th time step can be represented as:

$$y_i = \sum_{j\le i} \kappa_{i,j}\langle c_i, b_j\rangle v_j \tag{1}$$

where the kernel function $\kappa_{i,j}$ is implicitly determined by the state transition matrix and the discretization step size, reflecting a distance-dependent kernel that decays with sequence distance. The inner product $\langle c_i, b_j\rangle$ captures content-based interactions. The summation range $j \le i$ ensures strict causality, while the kernel $\kappa_{i,j}$ imparts a decay property that smoothly decreases as sequence distance increases.

However, 2D images inherently lack a unique and strict temporal orientation in spatial distribution. Therefore, architectures like Visual State Space Duality (VSSD) replace the single causal scan with a multi-directional sweeping and feature-fusion strategy. By relaxing the strictly causal lower-triangular mask while maintaining the $\mathcal{O}(L)$ computational bound, such models capture positional correlations through multi-directional scanning and subsequent aggregation. To facilitate a frequency-domain analysis of the impact of this aggregation, we abstract the multi-directional fusion operator into a global normalized aggregation approximation:

$$\hat{y}_i \approx \frac{\sum_{j=1}^{L} w_{i,j} v'_j}{\sum_{j=1}^{L} w_{i,j}}, \quad w_{i,j} = \phi(\langle c_i, b_j\rangle),\ \phi(\cdot) \ge 0. \tag{2}$$

In the equation above, $\phi(\cdot)$ is a non-negative monotonic mapping function (such as ReLU or exponential) and the resulting $w_{i,j}$ constitute the spatial fusion weights. The term $v'_j = \rho_j \odot v_j$ denotes the value after the per-token, input-dependent scaling that VSSD applies before aggregation, where ρ_j is the gain (induced by discretization and decay) absorbed into the value path. We make this scaled value explicit here because the difference $v_j - v'_j$ is precisely the signal that our calibration branch recovers in Section 3.2.

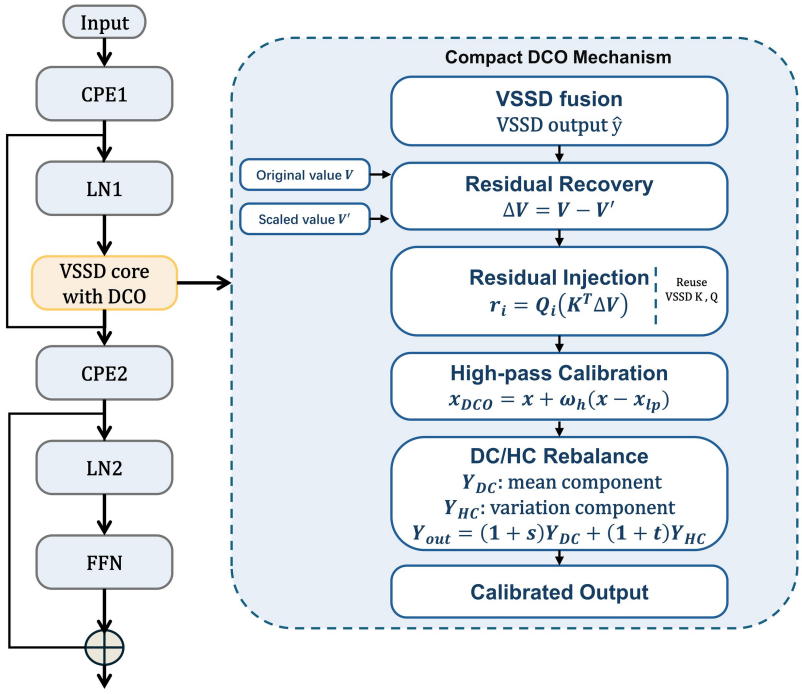


Fig. 3: Compact illustration of DCO inside a VSSD block, including residual recovery, state-level injection, token-level high-pass calibration, and DC/HC rebalancing.

Frequency-domain analysis of feature over-smoothing. We decompose each value into a slow-varying term and a local high-frequency perturbation, $v'_j = v_j^{\text{low}} + \epsilon_j$, where ϵ_j is approximately zero-mean within a local window. When the fusion weights $w_{i,j}$ vary smoothly over a large receptive field, the weighted summation in Equation (2) averages out the oscillatory ϵ_j, so the high-frequency energy in the output decays—manifesting as smoothed boundaries and details. Equivalently, with smooth weights the normalized aggregation $\alpha_{i,j} = w_{i,j} / \sum_j w_{i,j}$ approximates a large-window (in the uniform limit, $\alpha_{i,j} \approx 1/L$, global) averaging operator, which is inherently low-pass and collapses the L spatial tokens toward a low-rank summary. Crucially, once this aggregation has attenuated the high-frequency directions, position-dependent post-modulations alone cannot restore the lost amplitudes: an explicit high-frequency compensation branch is required. This provides a mechanistic explanation for the over-smoothing observed in visual Mamba baselines during dense remote sensing boundary delineation, and directly motivates DCO.

3.2 Duality Calibration Operator

DCO is inserted after VSSD's fusion operation, as illustrated in Fig. 3. It preserves VSSD's linear complexity by reusing the original aggregation weights,

while compensating for the local contrast and boundary details weakened by global fusion. DCO has three near-identity calibration steps: **(i)** state-level residual injection for recovering contrast removed by value rescaling (Eqs. (3)–(4)); **(ii)** token-level high-pass modulation for strengthening de-meaned responses (Eq. (5)); and **(iii)** channel-level DC/HC rebalancing between low-frequency semantics and high-frequency details (Eq. (6)).

Residual extraction and linear-time injection. Following Section 3.1, the difference between the original and scaled values captures local detail weakened by VSSD fusion. We extract this residual variation as $\Delta v_j = v_j - v'_j$ and reuse the same kernelized key/query statistics already computed by VSSD. With $k_j = \psi_v(b_j)$, the residual statistics are accumulated as

$$\boldsymbol{S}_\Delta = \sum_{j=1}^{L} k_j \Delta v_j^\top. \tag{3}$$

Since $\boldsymbol{S}_\Delta$ is shared across all query positions, it is computed once in linear time. The residual for token i is then read out by the corresponding query feature $q_i = \psi_v(c_i)$:

$$r_i \approx q_i^\top \boldsymbol{S}_\Delta, \tag{4}$$

which uses the existing VSSD aggregation statistics and adds no quadratic interaction. The residual is injected through a small-gain near-identity gate, $\tilde{y}_i = \hat{y}_i + (\varepsilon + \lambda_h) \odot r_i$, with λ_h initialized to 0.01. We use $\psi_v(\boldsymbol{x}) = \mathrm{ELU}(\boldsymbol{x}) + 1$ [19] and insert one DCO after the fusion operator in every VSSD block. Pseudocode, kernel choice, and complexity analysis are detailed in the supplementary material.

Token-level frequency calibration. We further elevate fine-grained responses using a de-meaned high-pass modulation:

$$y'_i = \tilde{y}_i + \omega_h \odot (\tilde{y}_i - \mu), \quad \text{with} \quad \mu = \frac{1}{L} \sum_{i=1}^{L} \tilde{y}_i, \tag{5}$$

This is followed by a mean/variation decomposition $Y' = Y_{\mathrm{dc}} + Y_{\mathrm{hc}}$. Finally, DCO reallocates channel weights via

$$Y_{\mathrm{out}} = (1 + s) \odot Y_{\mathrm{dc}} + (1 + t) \odot Y_{\mathrm{hc}}, \tag{6}$$

where ω_h, s, and t are bounded and initialized small (to 0.01). Thus, DCO starts near identity and then learns to rebalance low-frequency semantics and high-frequency structures during optimization.

3.3 Orthogonal Multi-Prototype Decoder

While DCO improves local separability, a single-prototype classifier may collapse rich intra-class modalities (e.g., "same object, different spectra") into a single direction. We propose an Orthogonal Multi-Prototype (OMP) decoder to preserve multi-modal structures in the logit space.

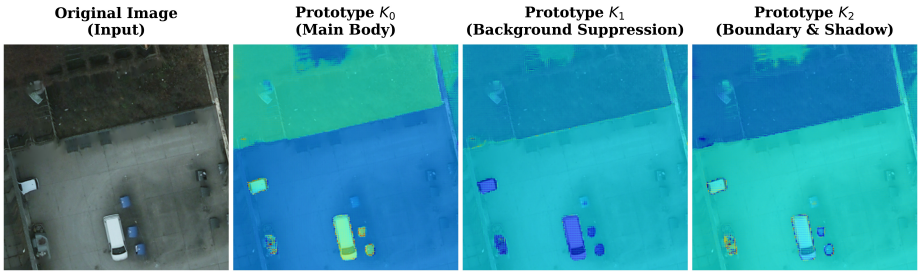


Fig. 4: Sub-prototype spatial allocation on Potsdam (K=3). Different sub-prototypes capture distinct intra-class spectral modes.

Dynamic tokens and bidirectional interaction. As illustrated in the right inset of Fig. 2, OMP generates content-adaptive class tokens in two steps: (1) a dual-path global pooling branch (max and average) extracts per-channel statistics from the feature map and fuses them through a shared MLP to produce content-adaptive query weights; (2) these weights are used to compose dynamic class tokens from a learned prototype embedding bank via a weighted Einsum. The resulting tokens are then refined with spatial features through stacked bidirectional cross-attention (details in supplementary).

Hyper-prototype matching. A hyper-network maps refined class tokens to K prototypes per class $P \in \mathbb{R}^{B\times|\mathcal{C}|\times K\times d_h}$. Here *pixel descriptors* $\hat{U} \in \mathbb{R}^{(H'W')\times d_h}$ denote the per-location feature vectors of the decoder feature map (one d_h-dimensional vector per spatial position), after L_2 normalization; they are the entities matched against the prototypes. As detailed in the left inset of Fig. 2, given pixel descriptors $\hat{U}$ and normalized prototypes $\hat{P}$, the per-class confidence is computed by max-response matching across the K sub-prototypes:

$$z_{c,i} = \max_{1\le k\le K} \langle \hat{U}_i, \hat{P}_{c,k} \rangle, \tag{7}$$

so each pixel selects the closest sub-prototype of class c while different sub-prototypes can specialize to different intra-class modes. Figure 4 visualizes the spatial activation of each sub-prototype, confirming that different sub-prototypes specialize in distinct intra-class modes rather than collapsing to the same region.

Dual prototype regularization. We impose intra-class orthogonality and inter-class angular margin constraints:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}} + \lambda_{\text{margin}}\mathcal{L}_{\text{margin}}, \tag{8}$$

where $\mathcal{L}_{\text{seg}}$ is the standard segmentation loss, $\mathcal{L}_{\text{orth}}$ penalizes redundant sub-prototypes, and $\mathcal{L}_{\text{margin}}$ enforces an angular margin δ between class centers $\hat{\bar{P}}_{b,c}$:

$$\mathcal{L}_{\text{orth}} = \frac{1}{B|\mathcal{C}|K(K-1)} \sum_{b,c} \sum_{k\neq l} \left| \langle \hat{P}_{b,c,k}, \hat{P}_{b,c,l} \rangle \right|, \tag{9}$$

$$\mathcal{L}_{\text{margin}} = \frac{1}{B} \sum_{b} \sum_{c\neq c'} \text{ReLU}\big(\langle \hat{\bar{P}}_{b,c}, \hat{\bar{P}}_{b,c'} \rangle - (1-\delta) \big). \tag{10}$$

Sensitivity of mIoU to $K \in \{2, 3, 4, 5\}$ on Potsdam is reported in Table S3; performance peaks at $K = 3$ and remains stable for larger K, confirming that the gains come from multi-modal prototype coverage rather than increased model capacity. The margin δ, loss weights $\lambda_{\text{orth}}, \lambda_{\text{margin}}$, and calibration gains $\lambda_h, \omega_h, s, t$ are listed in the supplementary material.

4 Experiments

4.1 Experimental Setup

We evaluate on ISPRS Potsdam and Vaihingen (5 classes, clutter ignored) and LoveDA (7 classes) [33, 37]. We report mIoU and mean F1-score (mF), and provide class-wise IoU for representative categories. All models are re-trained under a unified MMSegmentation [30] protocol with 512×512 random crops, AdamW (lr 6×10^{-5}, wd 0.05) [28], linear warmup and Poly decay ($p = 0.9$) [12]. We optimize the overall objective in Eq. (8). Additional training details (AMP, gradient clipping, and augmentation recipe) are deferred to the supplementary material.

4.2 Comparison with State-of-the-Arts

Quantitative and qualitative results. Table 1 reports LoveDA results, and Tables 2 and 3 report Potsdam and Vaihingen respectively. Across all benchmarks, CRISP achieves consistently strong mF and mIoU with a lightweight 32.32M parameter count. Gains are most pronounced on categories with high intra-class variance and irregular contours (*e.g.*, Low-veg and Tree on Potsdam/Vaihingen, Forest on LoveDA), which aligns directly with the design motivation of restoring attenuated high-frequency dynamics via DCO and preserving multi-modal structures via OMP. Figure 5 visualizes predictions on all three datasets: compared with the VSSD baseline, CRISP better preserves thin structures and object boundaries and reduces local shape distortion on small objects. Error map comparisons are in Fig. S1.

Efficiency. A full efficiency breakdown (parameters, FLOPs, peak memory, latency, and FPS) is provided in Table S1. Compared with the VSSD base model, CRISP reduces parameters and computation substantially while improving accuracy, indicating a favorable accuracy–efficiency trade-off rather than a purely capacity-driven gain. The reduction is mainly attributed to replacing the heavy UPerHead with the lightweight MultiProtoDecoder, while DCO introduces negligible overhead ($< 10^{-3}$M parameters and only 3.19G FLOPs); a component-wise breakdown is given in Table S2.

4.3 Ablation Studies

We conduct ablations on the Potsdam validation set under single-scale inference. Table 4 ablates the three DCO sub-components (residual injection, high-pass modulation, and DC/HC rebalancing), and Table 5 evaluates end-to-end component synergy. Activation-map visualizations are provided in Fig. S2.

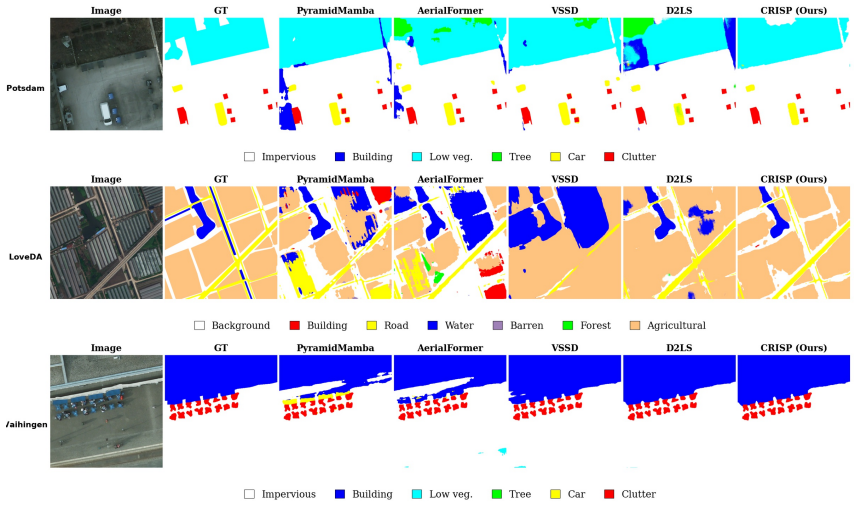


Fig. 5: Qualitative comparison with VSSD on Potsdam, LoveDA, and Vaihingen. CRISP improves boundary delineation and local shape fidelity.

Table 1: Quantitative comparison on LoveDA. Class-wise IoU (%) and mIoU are reported; best and second-best results are **bolded** and underlined.

Method	Params (M)	Bg	Bu	Rd	Wa	Ba	Fo	Ag	mIoU
ABCNet [22]	13.6	49.78	46.56	48.08	63.83	20.03	38.33	47.04	44.81
VSSD [34]	56.1	50.41	50.96	51.86	64.94	24.00	40.57	47.35	47.15
DCSwin [40]	45.6	49.70	56.95	47.92	66.14	27.53	42.04	43.56	47.69
PyramidMamba [39]	125.1	50.73	51.06	51.47	67.81	26.78	40.08	49.17	48.16
BANet [38]	12.7	53.10	59.52	48.83	65.06	23.09	41.12	49.00	48.53
SFA-Net [18]	10.7	52.58	58.73	53.53	66.84	27.31	39.16	48.85	49.57
UNetFormer [41]	11.7	51.29	58.89	53.76	66.92	27.68	41.49	49.11	49.88
AerialFormer [16]	43.2	52.09	64.05	55.21	64.31	27.40	41.78	48.07	50.41
CRISP (Ours)	32.32	**54.06**	**65.16**	54.60	67.63	24.63	**43.92**	50.90	51.56
D2LS [52]	94.0	52.60	63.27	**57.76**	**72.06**	**33.35**	41.10	**54.53**	**53.53**

DCO sub-components. Table 4 isolates the three calibration steps inside DCO. Each step, when activated alone, yields a consistent improvement over the 87.32 baseline (residual injection +0.33, high-pass modulation +0.36, DC/HC rebalancing +0.32), indicating that they address complementary facets of the over-smoothing problem rather than overlapping in effect. Residual injection restores the contrast removed by the per-token value rescaling, the high-pass term re-amplifies de-meaned boundary responses, and DC/HC rebalancing redistributes channel energy between low- and high-frequency components; none of them dominates in isolation. Combining the two frequency-domain steps (Hi-Pass + DC/HC) reaches 87.83, and adding residual injection on top recovers the remaining gap to the full DCO at 88.10. The fact that the full configuration exceeds the best two-way combination by a further +0.27 confirms that the state-level residual and the token-level frequency calibration are not redundant:

Table 2: Quantitative comparison on ISPRS Potsdam. Best/second-best results are **bolded**/underlined.

Method	Params (M)	Imp.	Build.	Low-veg.	Tree	Car	mFscore	mIoU
BANet	12.7	92.31	94.42	86.27	84.93	94.95	90.58	83.04
PyramidMamba	125.1	93.41	95.93	86.77	85.86	95.54	91.50	84.62
DCSwin	45.6	93.57	96.13	87.51	87.03	96.12	92.07	85.56
VSSD	56.1	88.28	92.81	77.19	77.30	92.25	92.07	85.57
SFA-Net	10.7	94.06	96.63	87.96	87.28	96.46	92.48	86.28
ABCNet	13.6	94.14	96.88	88.07	88.04	96.13	92.65	86.55
UNetFormer	11.7	94.15	96.83	88.21	88.07	96.14	92.68	86.59
AerialFormer	43.2	94.36	96.74	88.51	88.77	96.71	93.02	87.17
CRISP (Ours)	32.32	95.24	97.78	89.87	89.47	97.31	93.93	88.77
D2LS	94.0	**95.38**	**97.99**	**89.89**	**89.71**	**97.78**	**94.15**	**89.17**

Table 3: Quantitative comparison on ISPRS Vaihingen. Best/second-best results are **bolded**/underlined.

Method	Params (M)	Imp.	Build.	Low-veg.	Tree	Car	mFscore	mIoU
SFA-Net	10.7	91.76	94.35	82.42	88.43	75.73	86.54	76.87
UNetFormer	11.7	92.23	94.65	83.75	89.04	79.16	87.77	78.64
DCSwin	45.6	91.85	94.30	83.96	89.18	80.68	87.99	78.92
PyramidMamba	125.1	92.38	94.99	83.49	88.88	81.41	88.23	79.32
BANet	12.7	91.79	94.41	83.21	89.03	83.70	88.43	79.54
VSSD	56.1	85.88	90.37	70.73	79.43	72.35	88.54	79.75
ABCNet	13.6	92.82	95.44	83.29	89.06	84.86	89.09	80.64
AerialFormer	43.2	92.80	95.75	83.88	89.00	88.46	89.98	82.03
CRISP (Ours)	32.32	93.73	96.62	84.55	89.46	88.37	90.55	83.00
D2LS	94.0	**93.74**	**96.95**	**84.73**	**89.71**	**90.79**	**91.18**	**84.06**

the former supplies the high-frequency signal that the latter then re-weights. Because every gate is initialized near zero, DCO begins as an identity map and the improvements above are obtained without destabilizing the pretrained backbone.

Component synergy. Table 5 shows that DCO alone improves the VSSD+UPerHead baseline from 87.32 to 88.10 mIoU, confirming that the gain is not merely due to decoder capacity. OMP alone underperforms the baseline (85.63), which is expected: without the high-frequency detail restored by DCO, the multiple sub-prototypes have little distinct intra-class structure to latch onto and the orthogonality constraint mainly disperses redundant directions. Once DCO is present, the picture reverses—OMP with both regularizers reaches the best 88.25, and the two regularizers contribute almost equally (orthogonality 88.06, margin 88.02 when applied singly), with their combination adding a further margin. This ordering (DCO $\rightarrow$ OMP $\rightarrow$ regularizers) indicates that multi-prototype matching is a beneficiary of, rather than a substitute for, the separability introduced by DCO: the calibrated features expose the intra-class modes, and the prototype regularizers keep those modes from collapsing back onto a single direction.

Table 4: DCO sub-component ablation on Potsdam (UPerHead, single-scale).

Config.	Res. Inj.	Hi-Pass	DC/HC	mIoU
Baseline				87.32
Res. Inj. only	✓			87.65
Hi-Pass only		✓		87.68
DC/HC only			✓	87.64
Hi-Pass + DC/HC		✓	✓	87.83
Full DCO	✓	✓	✓	**88.10**

Table 5: End-to-end component ablation on Potsdam (VSSD backbone).

Dec.	DCO	Orth	Margin	mIoU
UPer				87.32
UPer	✓			88.10
OMP		✓	✓	85.63
OMP	✓			86.85
OMP	✓	✓		88.06
OMP	✓		✓	88.02
OMP	✓	✓	✓	**88.25**

5 Conclusion

This paper proposes CRISP to mitigate high-frequency attenuation and boundary blurring in visual state-space backbones. DCO treats the duality error of the state-space dual form as a measurable and correctable signal, restoring attenuated high-frequency responses through residual injection, local duality correction, and frequency-aware gating, while OMP preserves the recovered intra-class diversity with orthogonally regularized sub-prototypes. Experiments on remote-sensing semantic segmentation benchmarks demonstrate strong accuracy–efficiency trade-offs. Future work will explore spatially adaptive calibration conditioned on local frequency content.

Acknowledgements

This work was supported by the Key Research Program of Hangzhou (No. 2025SZD2B02). It was also supported by the Major Science and Technology Program of Zhejiang Province (No. 2026LDC01007(JT)).

References

1. Cao, Y., Liu, C., Wu, Z., Zhang, L., Yang, L.: Remote sensing image segmentation using vision Mamba and multi-scale multi-frequency feature fusion. Remote Sensing **17**(8), 1390 (2025)
2. Chen, K., Chen, B., Liu, C., Li, W., Zou, Z., Shi, Z.: RSMamba: Remote sensing image classification with state space model. IEEE Geoscience and Remote Sensing Letters **21**, 1–5 (2024)

3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. pp. 833–851 (2018)
4. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. In: NeurIPS. pp. 17864–17875 (2021)
5. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE **105**(10), 1865–1883 (2017)
6. Chi, L., Jiang, B., Mu, Y.: Fast fourier convolution. In: NeurIPS. pp. 4479–4488 (2020)
7. Dao, T., Gu, A.: Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In: ICML. pp. 10041–10071 (2024)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR. pp. 1–14 (2021)
9. Fan, L., Zhou, Y., Liu, H., Li, Y., Cao, D.: Combining Swin transformer with UNet for remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–11 (2023)
10. Fan, R., Xie, J., Liu, J., Zhang, J., Zhang, Y., Hou, H., Yang, J.: NAPG: Neighborhood-assisted multiprototype group model for cross-domain semantic segmentation of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–19 (2025)
11. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: CVPR. pp. 3146–3154 (2019)
12. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch SGD: Training ImageNet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
13. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: COLM. pp. 1–37 (2024)
14. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. In: ICLR. pp. 1–24 (2022)
15. Han, D., Pan, X., Han, Y., Song, S., Huang, G.: FLatten Transformer: Vision transformer using focused linear attention. In: ICCV. pp. 5961–5971 (2023)
16. Hanyu, T., Yamazaki, K., Tran, M., McCann, R.A., Liao, H., Rainwater, C., Adkins, M., Cothren, J., Le, N.: AerialFormer: Multi-resolution transformer for aerial image segmentation. Remote Sensing **16**(16), 2930 (2024)
17. Huang, T., Pei, X., You, S., Wang, F., Qian, C., Xu, C.: LocalMamba: Visual state space model with windowed selective scan. In: ECCVW. pp. 12–22 (2024)
18. Hwang, G., Jeong, J., Lee, S.J.: SFA-Net: Semantic feature adjustment network for remote sensing image segmentation. Remote Sensing **16**(17), 3278 (2024)
19. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are RNNs: Fast autoregressive transformers with linear attention. In: ICML. pp. 5156–5165 (2020)
20. Kirillov, A., Wu, Y., He, K., Girshick, R.: PointRend: Image segmentation as rendering. In: CVPR. pp. 9796–9805 (2020)
21. Li, K., Geng, Q., Wan, M., Cao, X., Zhou, Z.: Context and spatial feature calibration for real-time semantic segmentation. IEEE Transactions on Image Processing **32**, 5465–5477 (2023)
22. Li, R., Zheng, S., Zhang, C., Duan, C., Wang, L., Atkinson, P.M.: ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing **181**, 84–98 (2021)

23. Li, Z., Lu, F., Zou, J., Hu, L., Zhang, H.: Generalized few-shot meets remote sensing: Discovering novel classes. In: CVPRW. pp. 2744–2754 (2024)
24. Lin, C., Li, J., Che, Y., Yao, Y., Cao, X., Zheng, Z., Luo, W.: FreqFusion: Frequency-aware feature fusion for dense image prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 10763–10780 (2024)
25. Liu, S.A., Zhang, Y., Qiu, Z., Xie, H., Zhang, Y., Yao, T.: Learning orthogonal prototypes for generalized few-shot semantic segmentation. In: CVPR. pp. 11319–11328 (2023)
26. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: VMamba: Visual state space model. In: NeurIPS. pp. 1–33 (2024)
27. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 9992–10002 (2021)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR. pp. 1–18 (2019)
29. Ma, X., Zhang, X., Pun, M.O.: RS^{3}Mamba: Visual state space model for remote sensing image semantic segmentation. IEEE Geoscience and Remote Sensing Letters **21**, 1–5 (2024)
30. MMSegmentation Contributors: MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark. arXiv preprint arXiv:2108.06423 (2021)
31. Pei, X., Huang, T., Xu, C.: EfficientVMamba: Atrous selective scan for light weight visual Mamba. In: AAAI. pp. 6443–6451 (2025)
32. Rao, Y., Zhao, W., Zhu, Z., Lu, J., Zhou, J.: Global filter networks for image classification. In: NeurIPS. pp. 980–993 (2021)
33. Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., Breitkopf, U.: The ISPRS benchmark on urban object classification and 3d building reconstruction. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **I-3**, 293–298 (2012)
34. Shi, Y., Li, M., Dong, M., Xu, C.: VSSD: Vision Mamba with non-causal state space duality. In: ICCV. pp. 10819–10829 (2025)
35. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NeurIPS. pp. 4077–4087 (2017)
36. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: ICCV. pp. 7262–7272 (2021)
37. Wang, J., Zheng, Z., Ma, A., Lu, X., Zhong, Y.: LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In: NeurIPS. pp. 1–10 (2021)
38. Wang, L., Li, R., Wang, D., Duan, C., Wang, T., Meng, X.: Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. Remote Sensing **13**(16), 3065 (2021)
39. Wang, L., Li, D., Dong, S., Meng, X., Zhang, X., Hong, D.: PyramidMamba: Rethinking pyramid feature fusion with selective space state model for semantic segmentation of remote sensing imagery. International Journal of Applied Earth Observation and Geoinformation **144**, 104884 (2025)
40. Wang, L., Li, R., Duan, C., Zhang, C., Meng, X., Fang, S.: A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2022)
41. Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M.: UNetformer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS Journal of Photogrammetry and Remote Sensing **190**, 196–214 (2022)

42. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: ECCV. pp. 418–434 (2018)
43. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and efficient design for semantic segmentation with transformers. In: NeurIPS. pp. 12077–12090 (2021)
44. Yang, C., Chen, Z., Espinosa, M., Ericsson, L., Wang, Z., Liu, J., Crowley, E.J.: PlainMamba: Improving non-hierarchical Mamba in visual recognition. In: BMVC. pp. 1–10 (2024)
45. Yao, J., Hong, D., Li, C., Chanussot, J.: SpectralMamba: Efficient Mamba for hyperspectral image classification. arXiv preprint arXiv:2404.08489 (2024)
46. Yu, J., Xie, G., Liu, Q., Kan, Z., Wang, L., Liu, T., Ling, Q., Xu, W., Gao, F.: Contrastive learning with multiple prototypes for unsupervised domain adaptive semantic segmentation. IEEE Transactions on Multimedia **27**, 5267–5282 (2025)
47. Yuan, X., Shi, J., Gu, L.: A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Systems with Applications **169**, 114417 (2021)
48. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: ECCV. pp. 173–190 (2020)
49. Zhang, W., Ma, M., Jiang, Y., Lian, R., Wu, Z., Cui, K., Ma, X.: Center-guided classifier for semantic segmentation of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **64**, 1–16 (2026)
50. Zhao, X., Tang, F., Wang, X., Xiao, J.: SFC: Shared feature calibration in weakly supervised semantic segmentation. In: AAAI. pp. 7525–7533 (2024)
51. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision Mamba: Efficient visual representation learning with bidirectional state space model. In: ICML. pp. 62429–62442 (2024)
52. Zou, X., Li, Y., Zhang, S., Li, K., Wang, S., Tao, P., Xing, J., Lang, C.: Dynamic dictionary learning for remote sensing image segmentation. In: ICCV. pp. 22457–22466 (2025)
53. Zou, Z., Yu, H., Huang, J., Zhao, F.: FreqMamba: Viewing Mamba from a frequency perspective for image deraining. In: ACM MM. pp. 1905–1914 (2024)