HySeg: Learning Generative Priors for Structure-Aware Remote Sensing Segmentation
—Supplementary Material—

Jie Qiu[1], Xin Li[2], Fan Yang[2], Yan Wang[1], Dong Yu[3], Changying Wang[1],
Linwei Dai[4], Yongxiang Chen[1], Youqin Chen[5,*], Jianzhang Chen[1,*]
[1]Fujian Agriculture and Forestry University [2]Alpaca AI Lab
[3]Beijing Jiaotong University [4]iFLYTEK [5]Fujian University of Technology

This **Supplementary Material** provides additional evidence supporting the effectiveness of HySeg. We first present the detailed formulation of the Prior-to-Affinity Projection (P2A) module, including neighborhood unfolding, spatial index mapping, boundary handling, and affinity construction, which complements the simplified description in the main text. We then report detailed per-class results on Potsdam, Vaihingen, and UAVid [11, 15] using multiple backbones, including ResNet18, ResNet34, ResNet50 [5], Swin_T/S/B [9], LSKNet_T/S [7], and ConvNeXt_B [10]. Across all settings, HySeg consistently improves mIoU over purely discriminative heads, with especially strong gains on structure-sensitive and small-object categories. We next present extended qualitative comparisons on ISPRS and LoveDA scenes [15, 21]. HySeg produces more coherent low-vegetation regions and substantially more continuous thin-road structures, revealing its ability to preserve fine-scale topology. We also provide additional analyses using t-SNE feature visualizations [20] and training and validation curves on LoveDA, showing that MeanStruct priors lead to cleaner class separation and improved generalization while preserving stable optimization behavior.

Finally, we compare HySeg with dense CRF refinement [2, 14] under identical backbones to assess whether its gains can be attributed to simple post-hoc smoothing. The results show that CRF provides only modest refinement and has limited impact on overall accuracy, whereas HySeg delivers substantially larger and more stable improvements. Moreover, HySeg produces predictions with clearer boundaries and more reliable structural continuity in complex urban layouts, indicating that its benefits go well beyond conventional CRF regularization. We further evaluate HySeg with modern pretrained remote sensing backbones, where it remains consistently beneficial, demonstrating compatibility not only with conventional CNN and Transformer architectures but also with stronger multimodal pretrained representations. In addition, we compare HySeg against lightweight structural constraints such as boundary loss (*bl*) [1] and edge heads (*eh*) [19]. These comparisons show that simple local regularization yields only limited gains, whereas HySeg provides more consistent improvements, suggesting that its advantage stems from coupling generative structural priors with topology-aware posterior reasoning rather than from boundary smoothing alone.

*Corresponding authors: chenyouqin@fjut.edu.cn, jchen@fafu.edu.cn

1. Detailed Formulation of Prior-to-Affinity Projection (P2A)

This section provides the detailed mathematical formulation of P2A, including neighborhood unfolding, spatial index mapping, boundary masking, and affinity construction. These implementation-level details complement the simplified formulation in the main text and are omitted there for clarity.

Prior feature map and spatial indexing. For scale i, let the prior feature map be

$$s_i \in \mathbb{R}^{C\times h\times w}, \tag{1}$$

where C is the channel dimension and (h, w) is the spatial resolution. Let

$$L = h \times w \tag{2}$$

denote the number of spatial sites. We index spatial locations in row-major order. For each site index $j \in \{1, \dots, L\}$, let (τ_j, μ_j) denote its 1-based row and column coordinates, such that

$$\tau_j \in \{1, \dots, h\}, \qquad \mu_j \in \{1, \dots, w\}, \tag{3}$$

and

$$j = (\tau_j - 1)w + \mu_j. \tag{4}$$

Neighborhood unfolding. P2A extracts a local $K \times K$ neighborhood around each spatial site, where K is odd. Let

$$r = \lfloor K/2 \rfloor \tag{5}$$

be the half-window size. We first zero-pad the feature map to obtain

$$s_i^{\mathrm{pad}} \in \mathbb{R}^{C\times(h+2r)\times(w+2r)}, \tag{6}$$

defined element-wise as

$$s_i^{\mathrm{pad}}(c,u,v) = \begin{cases} s_i\big(c,\, u-r,\, v-r\big), & \text{if } 1 \le u-r \le h \\ & \wedge\ 1 \le v-r \le w, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

We enumerate the K^2 neighborhood offsets in row-major order:

$$\mathcal{N} = \{(\Delta h_k, \Delta w_k)\}_{k=1}^{K^2}, \tag{8}$$

where

$$\Delta h_k = \left\lfloor \frac{k-1}{K} \right\rfloor - r, \qquad \Delta w_k = ((k-1) \bmod K) - r. \tag{9}$$

The center index is

$$k_{\mathrm{ctr}} = \frac{K^2+1}{2}, \tag{10}$$

corresponding to offset $(0,0)$.

Using these offsets, the unfolded neighborhood tensor is defined as

$$P(c,j,k) = s_i^{\mathrm{pad}}\big(c, \tau_j + r + \Delta h_k, \mu_j + r + \Delta w_k\big), \tag{11}$$

with

$$P \in \mathbb{R}^{C\times L\times K^2}. \tag{12}$$

Boundary masking. For boundary locations, some offsets fall outside the valid spatial domain. We therefore define a binary validity mask

$$\Gamma(j,k) = \mathbf{1}[1 \le \tau_j + \Delta h_k \le h,\ 1 \le \mu_j + \Delta w_k \le w], \tag{13}$$

which is used to exclude padded entries from affinity normalization.

Center–neighbor discrepancy. For each site j, channel c, and neighborhood index k, we define the center feature and the k-th neighboring feature as

$$p_j^{(c)} = P(c,j,k_{\mathrm{ctr}}), \qquad x_j^{(c,k)} = P(c,j,k). \tag{14}$$

Their channel-wise discrepancy is

$$D_j^{(c,k)} = \big(p_j^{(c)} - x_j^{(c,k)}\big)^2. \tag{15}$$

Affinity construction. We aggregate the channel-wise discrepancies and convert them into normalized affinity weights using a learnable Gaussian kernel:

$$A_j^{(k)} = \mathrm{Softmax}_k\left(-\frac{\sum_{c=1}^{C} D_j^{(c,k)}}{2\sigma^2 + \xi}\right), \tag{16}$$

where the softmax is applied over valid neighbors only, as indicated by $\Gamma(j,k)$.

The resulting affinity field provides structurally informed weights for local message passing in DAS. Unlike self-attention, P2A does not compute affinities from pairwise similarities among segmentation-branch features; instead, it projects them from the generative structural prior learned by MeanStruct, yielding a topology-aware local weighting field for posterior inference.

2. Additional Quantitative Results

Tab. 7, 8 and 9 report detailed per-class results on the ISPRS Potsdam and Vaihingen benchmarks as well as the UAVid dataset using several representative backbones (ResNet18/34/50, Swin_T/S/B, LSKNet_T/S, and ConvNeXt_B). Across all settings, incorporating HySeg into the baseline architectures consistently yields higher mIoU than their purely discriminative counterparts. On Potsdam, the improvements are especially notable for structure-dominated categories such as *Impervious Surfaces*, *Building*, and *Low Vegetation*, where HySeg produces sharper boundaries and significantly fewer fragmented regions. Similar trends appear on Vaihingen and UAVid, indicating that the structural priors learned by HySeg transfer reliably across different altitudes and sensing conditions.

Across backbones, the best HySeg configurations improve mIoU by as much as +1.7 on Potsdam, +2.2 on Vaihingen, and +1.6 on UAVid. These gains are most pronounced for geometrically complex or small-object categories, including the *Tree* on Potsdam and Vaihingen, and the traffic-related classes *Moving Car* (MvC) and *Static Car* (StC) on UAVid, where HySeg reduces boundary breaks and inter-class confusions. This consistent behavior reflects the effectiveness of the MeanStruct prior and the P2A projection in converting generative structural information into topology-aware and class-aware affinities that steer the DAS head.

Taken together, the results demonstrate that HySeg introduces an explicit form of structural reasoning into otherwise purely discriminative segmentation networks. This leads to robust and architecture-agnostic improvements, particularly for land-cover categories whose semantics depend strongly on topology and spatial adjacency rather than local texture alone.

Table 7. **Comparison with SOTA methods** on the Potsdam dataset. △ indicates the performance change from the base model to its HySeg-enhanced variant. Best results are highlighted in **bold**.

Method	Backbone	Imp.surf.	Building	Low veg.	Tree	Car	mF1	OA	mIoU
Segmenter[18]	ViT-B	86.4	90.6	78.8	74.4	82.4	82.5	82.2	80.6
MANet[6]	ResNet50	92.8	96.5	87.8	89.3	96.2	92.5	91.1	86.3
DeepLabV3+[3]	ResNet50	89.2	92.9	83.1	81.9	92.2	97.9	86.2	78.7
CTMFNet[17]	ResNet50	93.6	96.7	88.0	89.5	97.0	93.0	91.5	87.1
DANet[4]	ResNet50	91.7	95.5	86.4	88.5	89.9	90.4	89.8	82.7
SAM_RS[12]	ResNet18	93.0	95.9	88.0	88.0	96.3	92.2	90.4	85.9
LOGCAN++[13]	ResNet50	93.4	97.1	88.1	89.7	96.4	92.9	91.5	87.0
In2NeCT[16]	ResNet18	/	/	/	/	/	93.2	91.8	87.9
UNetFormer[23]	ResNet18	92.8	96.3	87.3	88.4	96.1	92.2	90.7	85.7
	ResNet34	92.8	96.3	88.0	88.7	96.1	92.4	91.0	86.1
	ResNet50	93.0	96.5	87.7	89.0	96.6	92.6	91.0	86.4
Ours	ResNet18	93.4	96.6	87.8	89.0	96.8	92.7	91.3	86.5
	△	+0.6	+0.3	+0.5	+0.6	+0.7	+0.5	+0.6	+0.8
	ResNet34	93.5	96.4	88.3	89.1	96.8	92.8	91.3	86.7
	△	+0.7	+0.1	+0.3	+0.4	+0.7	+0.4	+0.3	+0.6
	ResNet50	93.6	96.9	87.6	89.7	96.9	93.0	91.5	86.9
	△	+0.6	+0.4	-0.1	+0.7	+0.3	+0.4	+0.5	+0.5
DCSwin[22]	Swin_T	92.2	95.9	87.8	88.6	95.5	92.0	90.1	85.3
	Swin_S	93.2	96.8	88.3	89.5	95.5	92.7	91.2	87.1
	Swin_B	93.4	96.8	88.2	89.5	96.5	92.9	91.3	86.9
Ours	Swin_T	93.6	97.1	88.9	90.5	96.6	93.3	91.9	87.7
	△	+1.4	+1.2	+1.1	+1.9	+1.1	+1.3	+1.8	+2.4
	Swin_S	94.0	97.2	89.3	90.7	96.7	93.6	92.2	88.1
	△	+0.8	+0.4	+1.0	+1.2	+1.2	+0.9	+1.0	+1.0
	Swin_B	94.1	97.4	89.1	90.5	96.7	93.6	92.2	88.1
	△	+0.7	+0.6	+0.9	+1.0	+0.2	+0.7	+0.9	+1.2
LSKNet[7]	LSKNet_T	93.1	96.5	88.0	88.3	95.3	92.3	91.1	86.1
	LSKNet_S	93.9	97.3	88.1	89.5	96.5	93.1	92.0	87.2
Ours	LSKNet_T	94.2	97.3	89.4	89.6	96.7	93.4	91.9	87.2
	△	+1.1	+0.8	+1.4	+1.3	+1.4	+1.1	+0.8	+1.1
	LSKNet_S	95.1	**97.9**	90.3	90.9	97.1	94.3	93.4	88.9
	△	+1.2	+0.6	+2.2	+1.4	+0.6	+1.2	+1.4	+1.7
D2LS[25]	ConvNeXt_B	/	/	/	/	/	94.7	/	/
Ours	ConvNeXt_B	**95.8**	97.7	**91.3**	**91.4**	**97.8**	**94.8**	**93.9**	**89.5**
	△	/	/	/	/	/	+0.1	/	/

3. In-Depth Qualitative Comparisons

Fig. 6 presents two groups of qualitative comparisons. The first group (top) shows segmentation outputs on an ISPRS input image, with emphasis on the red-marked regions corresponding to *Low Vegetation*. For the same input, the baseline predictions of UNetFormer_R34, UNetFormer_R50, DCSwin_T, DCSwin_B, and LSKNet_T frequently misclassify these areas or merge them into surrounding categories. After integrating HySeg (HySeg_R34, HySeg_R50, HySeg_Swin_T, HySeg_Swin_B, HySeg_LSK_T), the same red-marked areas are more consistently recognized as *Low Vegetation*, forming coherent and contiguous patches that better match the underlying scene layout. This behavior indicates that the generative structural priors injected by HySeg help the network better separate vegetation from adjacent land-cover types and reduce ambiguity in structurally complex regions.

The second group (bottom) provides qualitative results on a LoveDA input image, focusing on the curved road segment highlighted by the red box. With identical inputs, baseline predictions produced by UNetFormer_R34, UNetFormer_R50, DCSwin_T, DCSwin_B, and LSKNet_T often yield roads that appear broken, thinned, or partially missing, with substantial confusion between roads and neighboring forest regions. After introducing HySeg (HySeg_R34, HySeg_R50, HySeg_Swin_T, HySeg_Swin_B, HySeg_LSK_T), the curved road is generally more continuous and smoother in overall shape. This illustrates that the combination of the MeanStruct prior, the P2A projection, and the DAS structural constraints improves the topological consistency of thin structures such as roads. Although small gaps or missed segments may still occur in extremely narrow seg-

Table 8. **Comparison with SOTA methods** on the Vaihingen dataset. △ indicates the performance change from the base model to its HySeg-enhanced variant. Best results are highlighted in **bold**.

Method	Backbone	Imp.surf.	Building	Lowveg.	Tree	Car	mF1	OA	mIoU
Segmenter[18]	ViT-B	86.9	88.8	79.2	87.0	67.6	81.9	85.3	76.1
MANet[6]	ResNet50	93.1	95.8	84.7	90.6	89.7	90.8	91.3	83.3
DeepLabV3+[3]	ResNet50	88.1	89.8	79.4	87.6	61.3	81.2	86.2	79.6
CTMFNet[17]	ResNet50	93.5	96.2	85.1	90.7	90.0	91.1	91.6	83.9
DANet[4]	ResNet50	90.6	94.3	82.8	88.8	72.9	85.9	89.2	76.0
SAM_RS[12]	ResNet18	92.6	96.5	81.0	90.1	90.3	90.1	91.2	82.9
LOGCAN++[13]	ResNet50	93.7	96.2	85.6	90.7	90.6	91.4	91.8	84.3
In2NeCT[16]	ResNet18	/	/	/	/	/	91.3	91.5	84.5
UNetFormer[23]	ResNet18	92.7	95.4	84.4	90.1	87.2	90.0	90.9	82.0
	ResNet34	92.8	95.5	84.8	90.2	88.0	90.3	91.1	82.5
	ResNet50	93.1	95.8	84.8	90.2	89.1	90.6	91.3	83.1
Ours	ResNet18	92.9	96.6	86.4	90.5	87.3	90.7	91.5	82.8
	△	+0.2	+1.2	+2.0	+0.4	+0.1	+0.7	+0.6	+0.8
	ResNet34	94.2	96.2	85.5	91.0	88.3	91.0	91.8	83.2
	△	+1.4	+0.7	+0.7	+0.8	+0.3	+0.7	+0.7	+0.7
	ResNet50	94.7	96.1	85.7	91.5	89.9	91.6	92.1	84.4
	△	+0.6	+0.3	+0.9	+1.3	+0.8	+1.0	+0.8	+1.3
DCSwin[22]	Swin_T	93.0	95.7	84.8	90.7	86.0	90.0	91.2	82.1
	Swin_S	93.6	96.2	85.2	90.4	87.9	90.6	91.6	83.1
	Swin_B	93.6	96.2	85.3	90.4	87.1	90.5	91.7	83.4
Ours	Swin_T	94.1	96.5	86.3	90.8	84.7	90.5	92.2	82.9
	△	+1.1	+0.8	+1.5	+0.1	-1.3	+0.5	+1.0	+0.8
	Swin_S	94.9	96.3	86.4	91.4	87.6	91.3	93.1	84.3
	△	+1.3	+0.1	+1.2	+1.0	-0.3	+0.7	+1.5	+1.2
	Swin_B	95.1	97.6	**88.1**	**91.7**	87.2	92.0	93.5	85.4
	△	+1.6	+1.4	+2.8	+1.3	+0.1	+1.5	+1.8	+2.0
LSKNet[7]	LSKNet_T	93.4	96.2	84.9	90.4	90.0	90.9	91.5	83.6
	LSKNet_S	93.6	96.2	85.2	90.6	89.5	91.0	91.7	83.7
Ours	LSKNet_T	93.6	96.8	84.9	91.3	**90.9**	91.5	92.9	84.6
	△	+0.2	+0.6	+0.0	+0.9	+0.9	+0.6	+1.4	+1.0
	LSKNet_S	**95.4**	**97.8**	87.2	91.1	89.7	**92.2**	**93.8**	**85.9**
	△	+1.8	+1.6	+2.0	+0.5	+0.2	+1.2	+2.1	+2.2
D2LS[25]	ConvNeXt_B	/	/	/	/	/	91.9	/	/
Ours	ConvNeXt_B	95.3	97.2	87.4	90.9	89.7	92.1	93.6	85.7
	△	/	/	/	/	/	+0.2	/	/

ments or in regions where the road visually resembles the background, HySeg demonstrates a consistently stronger ability to preserve road geometry, maintain spatial continuity, and reduce fragmentation compared with purely discriminative baselines.

4. Qualitative Representation Analysis

Fig. 7 visualizes segmentation features on LoveDA using t-SNE with different priors. Without any prior (Baseline), the feature clusters of *Building*, *Road*, *Water*, *Barren*, *Forest*, and *Agriculture* exhibit substantial overlap, especially along the boundaries between Building and Road and between Barren and Agriculture. This overlap suggests that the network relies mainly on local appearance cues and encounters difficulty in separating structurally related categories. Introducing discriminative priors from common backbones (ResNet_Prior, SwinTransformer_Prior, LSKNet_Prior, and ConvNeXt_Prior) gradually tightens intra-class clusters and modestly enlarges inter-class margins, although noticeable mixing among vegetation-related classes still remains. In contrast, when HySeg adopts the MeanStruct generative prior (MeanStruct_Prior), the t-SNE embeddings form more compact and clearly separated clusters, with more distinct separation between man-made and natural classes. This observation supports our claim that transforming the MeanStruct field into topology-aware and class-aware affinities through P2A, and propagating them within DAS, encourages structure-consistent representations and produces cleaner feature organization than using discriminative priors alone.

Fig. 8 plots training and validation curves (F1, OA,

Table 9. **Comparison with SOTA methods** on the UAVid dataset. △ indicates the performance change from the base model to its HySeg-enhanced variant. Best results are highlighted in **bold**.

Method	Backbone	Clt	Bld	Rd	Tre	Vgt	MvC	StC	Hum	mIoU
Segmenter[18]	ViT-B	66.6	86.3	80.1	79.6	62.3	72.5	52.5	28.5	66.0
MANet[6]	ResNet50	64.5	85.4	77.8	77.0	60.3	67.2	53.6	14.9	62.6
DeepLabV3+[3]	ResNet50	58.7	82.3	76.4	75.2	57.8	60.7	52.9	20.7	60.6
CTMFNet[17]	ResNet50	67.2	85.8	81.2	80.4	63.7	74.2	56.2	30.4	67.4
DANet[4]	ResNet50	64.9	85.9	77.9	78.3	61.5	59.6	47.4	29.1	60.6
SAM_RS[12]	ResNet18	62.7	84.9	81.0	80.1	59.2	68.8	50.1	23.5	63.8
LOGCAN++[13]	ResNet50	68.2	87.0	80.9	81.3	65.2	74.2	58.2	35.3	68.8
In2NeCT[16]	ResNet18	/	/	/	/	/	/	/	/	/
UNetFormer[23]	ResNet18	68.4	87.4	81.5	80.2	63.5	73.6	56.4	31.0	67.8
	ResNet34	68.7	88.0	81.3	80.3	63.7	73.9	56.8	31.3	68.0
	ResNet50	68.8	88.2	81.6	80.7	63.9	73.7	57.4	33.5	68.5
Ours	ResNet18	68.8	88.5	81.7	80.6	64.2	74.2	57.3	34.1	68.7
	△	+0.4	+1.1	+0.2	+0.4	+0.7	+0.6	+0.9	+3.1	+0.9
	ResNet34	69.0	88.4	81.9	80.6	64.5	74.4	57.4	33.8	68.8
	△	+0.3	+0.4	+0.6	+0.3	+0.8	+0.5	+0.6	+2.5	+0.8
	ResNet50	69.5	89.5	82.2	81.7	64.9	74.7	58.0	35.3	69.5
	△	+0.7	+1.3	+0.6	+1.0	+1.0	+1.0	+0.6	+1.8	+1.0
DCSwin[22]	Swin_T	68.7	87.2	81.8	80.6	63.8	73.8	56.8	31.5	68.0
	Swin_S	68.9	87.5	82.0	80.8	64.3	74.1	57.2	32.1	68.4
	Swin_B	68.5	88.1	81.6	81.2	63.5	73.7	56.5	31.9	68.1
Ours	Swin_T	69.1	88.3	82.0	81.0	64.6	74.5	57.9	33.5	68.9
	△	+0.4	+1.1	+0.2	+0.4	+0.8	+0.7	+1.1	+2.0	+0.9
	Swin_S	69.3	88.6	82.4	81.4	64.8	75.1	58.5	35.0	69.4
	△	+0.4	+1.1	+0.4	+0.6	+0.5	+1.0	+1.3	+2.9	+1.0
	Swin_B	70.3	88.9	83.2	81.8	65.0	75.0	58.7	34.3	69.7
	△	+1.8	+0.8	+1.6	+0.6	+1.5	+1.3	+2.2	+2.4	+1.6
LSKNet[7]	LSKNet_T	69.6	87.9	82.8	80.6	64.8	77.3	60.2	31.3	69.3
	LSKNet_S	69.6	84.8	82.9	80.9	65.5	76.8	64.9	31.8	70.0
Ours	LSKNet_T	70.6	88.3	82.7	81.8	64.9	77.5	60.2	32.3	69.8
	△	+1.0	+0.4	-0.1	+1.2	+0.1	+0.2	+0.0	+1.0	+0.5
	LSKNet_S	71.2	86.5	83.3	82.0	**66.9**	**78.1**	**65.6**	33.3	70.9
	△	+1.6	+1.7	+0.4	+1.1	+1.4	+1.3	+0.7	+1.5	+0.9
D2LS[25]	ConvNeXt_B	71.0	89.7	83.2	82.1	66.1	75.0	59.0	**41.4**	70.9
Ours	ConvNeXt_B	**71.6**	**90.1**	**83.5**	**82.5**	66.7	76.1	60.2	40.7	**71.4**
	△	+0.6	+0.4	+0.3	+0.4	+0.6	+1.1	+1.2	-0.7	+0.5

Table 10. **Effect of CRF and HySeg** on LoveDA performance with ResNet18/50, Swin_B, LSKNet_S, and ConvNeXt_B backbones.

Method/Backbone	**ResNet18**	**ResNet50**	**Swin_B**	**LSKNet_S**	**ConvNeXt_B**
Baseline	52.42	52.52	52.03	54.01	52.77
+ CRF	52.81	53.34	52.81	53.75	54.47
+ HySeg (Ours)	54.48	54.74	54.46	55.37	55.63

and mIoU) on LoveDA for four representative backbones (ResNet18, Swin_S, LSKNet_S, and ConvNeXt_B), both with and without HySeg. During training, HySeg closely tracks the baseline and sometimes yields slightly lower F1 or mIoU, indicating that structure-consistent generative priors do not impede convergence and function instead as a mild regularizer that prevents overfitting. In validation, however, HySeg consistently achieves higher or more rapidly stabilizing F1, OA, and mIoU across all backbones, with the gaps becoming most prominent in later epochs, particularly for LSKNet_S and ConvNeXt_B. The validation curves with HySeg are also smoother and show reduced oscillation compared with their purely discriminative counterparts. These results jointly confirm that injecting structure-consistent generative priors based on MeanStruct, and propagating them via P2A and DAS, improves generalization. HySeg maintains competitive optimization behavior during training while delivering consistently better performance at test time.

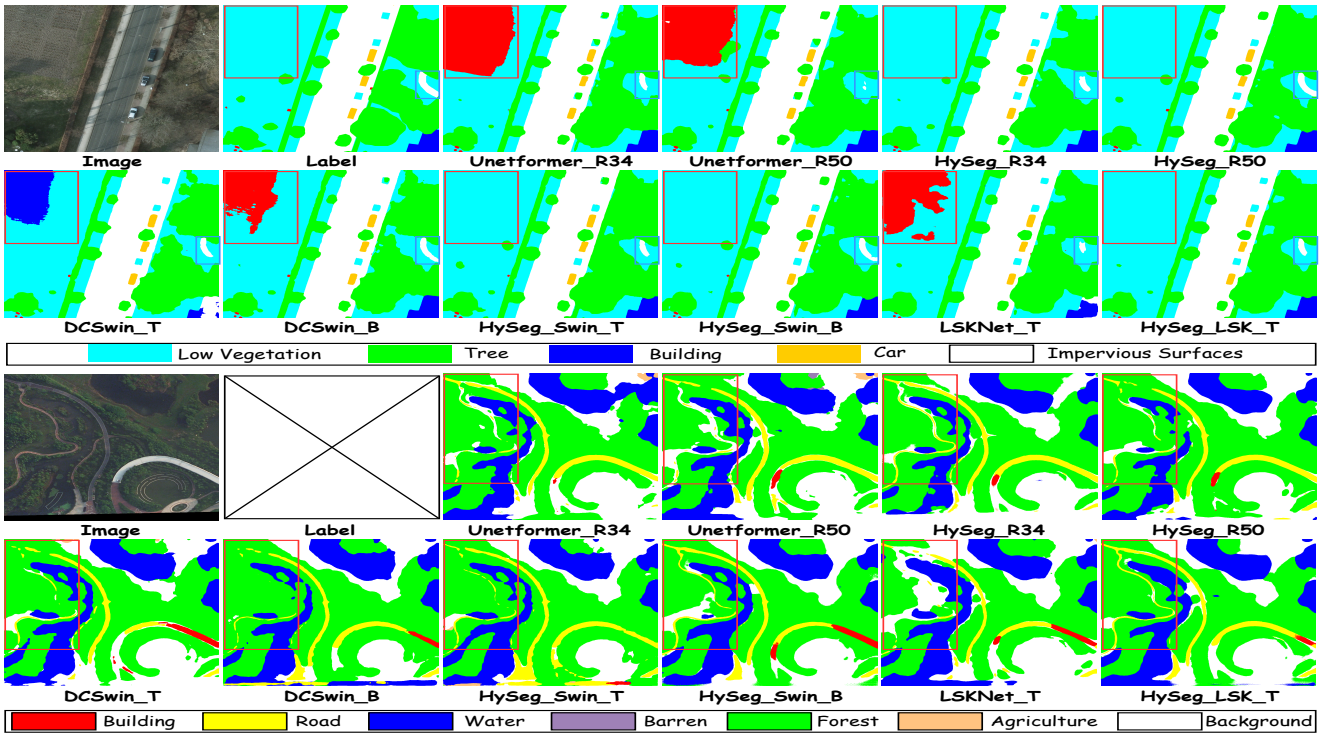

Figure 6. **Visual comparison** on ISPRS and LoveDA scenes using UNetFormer_R34/50, DCSwin_T/B, and LSKNet_T, with and without HySeg. Highlighted regions show that HySeg recovers more continuous low-vegetation strips between roads and background on ISPRS, and more complete, smooth curved roads on LoveDA, reducing confusions with surrounding vegetation and improving the topology of thin, elongated structures. The incomplete recognition of the overpass in the bottom right corner of the second panel represents a failure case, though such instances remain localized rather than dominant.

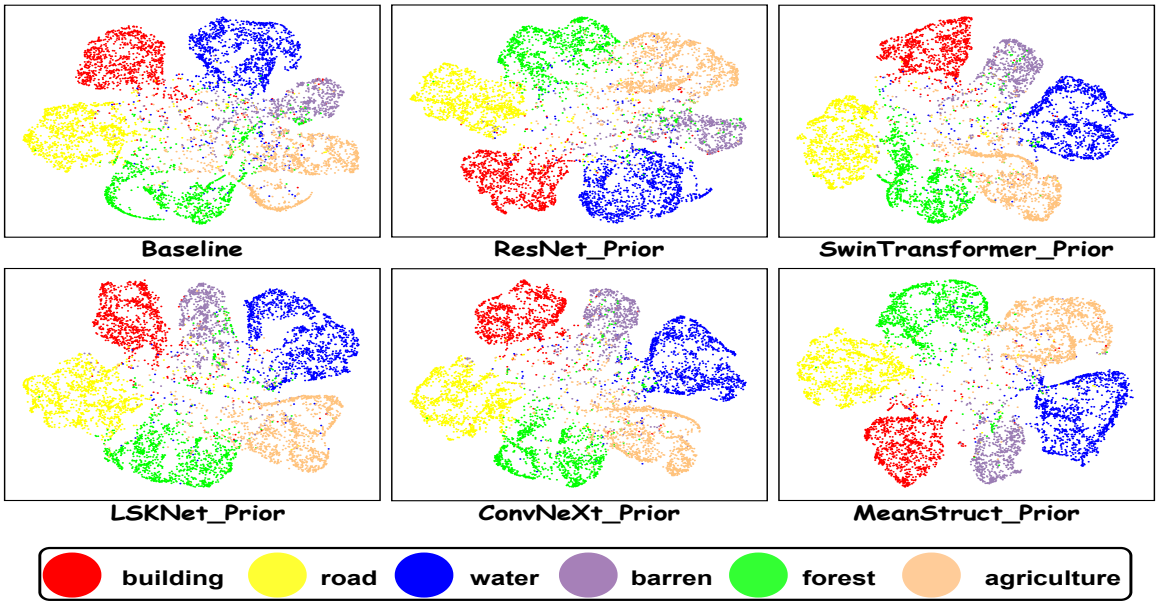

Figure 7. **Visual comparison** of t-SNE visualizations of segmentation features on LoveDA: The figure compares the segmentation performance with different priors (ResNet, SwinTransformer, LSKNet, ConvNeXt, and MeanStruct). The t-SNE plots highlight the impact of these priors on improving feature clustering across classes.

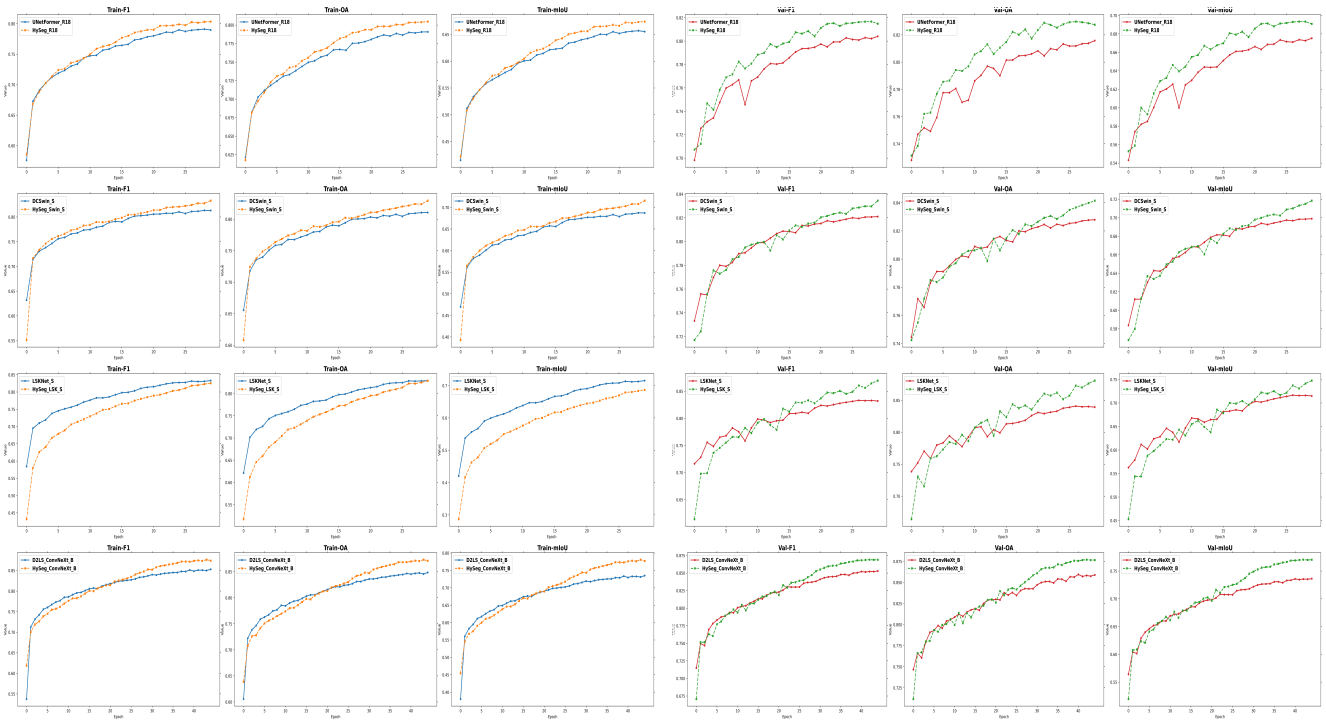

Figure 8. **Impact of HySeg on training and validation dynamics.** Across all backbones, HySeg tracks the baseline closely on the training curves while producing markedly smoother validation trajectories and consistently higher final F1/OA/mIoU. This reflects stable optimization on the training set and improved generalization on the validation split.

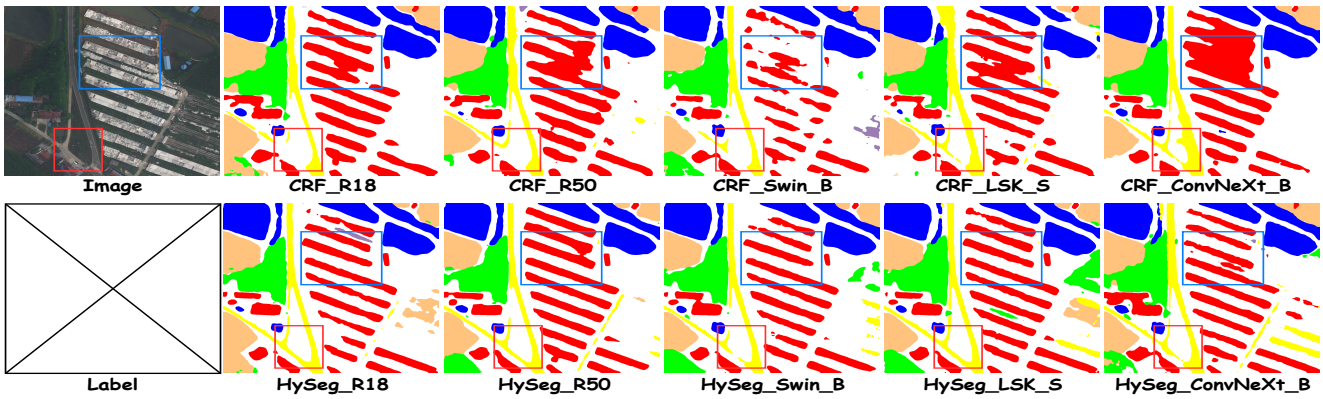

Figure 9. **CRF vs. HySeg under unified backbones.** Top: Predictions refined with CRF; bottom: Outputs from the HySeg head using the same set of backbones (ResNet18/50, Swin_B, LSKNet_S, ConvNeXt_B). HySeg yields structurally cleaner results, producing more compact building blocks and sharper narrow corridors than CRF, particularly in the blue- and red-highlighted regions.

For the LSKNet_S backbone in particular, we observe that HySeg attains slightly lower training F1/OA/mIoU than the standard head, whereas its validation scores are consistently higher. We attribute this to the stronger regularizing effect of the MeanStruct–P2A–DAS pipeline on the rich, highly expressive feature representations produced by LSKNet_S: while the baseline head tends to overfit the training set, HySeg accepts a small loss in training accuracy in exchange for more topology-consistent predictions and improved generalization on the validation set.

5. Evaluation Against CRF Post-processing

Tab. 10 presents LoveDA results obtained by augmenting five representative backbones (ResNet18, ResNet50, Swin_B, LSKNet_S, and ConvNeXt_B) with either a classical CRF post-processing module or our HySeg head. CRF is included as a reference baseline because it remains one of the most widely used structural priors in semantic segmentation. HySeg, in contrast, introduces structural information directly during end-to-end training through the MeanStruct

generative prior and the P2A–DAS segmentation head.

Across all five backbones, adding CRF yields only modest and sometimes inconsistent improvements. ResNet18, ResNet50, and Swin_B gain roughly +0.4 to +0.8 mIoU, while LSKNet_S shows a slight performance drop. ConvNeXt_B benefits somewhat more, although the improvement remains limited. Replacing the standard head with HySeg produces clearly larger and stable gains across all configurations, improving mIoU by +1.4 to +2.9 and surpassing the CRF variants by more than +1.1 mIoU on every backbone. These results indicate that transforming generative structural priors into topology-aware and class-aware affinity fields through P2A, and propagating them within DAS, provides a more effective form of structural reasoning than CRF.

Fig. 9 offers qualitative comparisons between CRF-enhanced models (top) and HySeg-enhanced models (bottom) using the same five backbones on LoveDA. In the blue-marked regions containing densely arranged building strips, CRF still generates fragmented and irregular patterns, with frequent label bleeding into adjacent categories such as roads or background. HySeg reconstructs more compact and homogeneous buildings with much cleaner boundaries. In the red-marked regions, CRF often misclassifies thin corridor-like structures, while HySeg preserves these narrow elements more reliably and produces smoother, more continuous shapes. These visual observations align with the quantitative findings in Tab. 10: CRF offers mainly limited smoothing effects, whereas HySeg provides substantial and structure-consistent improvements, especially in scenes dominated by fine-scale man-made structures and intricate boundaries.

6. Effectiveness with Modern Pretrained RS Foundation Backbones

HySeg also remains effective when paired with modern pretrained remote sensing backbones. As shown in Tab. 11, integrating HySeg improves LoveDA mIoU from 53.32 to 54.85 on RemoteCLIP [8] and from 54.25 to 55.50 on SkySense++ [24], corresponding to gains of +1.53 and +1.25, respectively. These results further indicate that the proposed structural prior mechanism is compatible not only with conventional CNN and Transformer backbones, but also with stronger multimodal pretrained remote sensing representations.

Table 11. **LoveDA mIoU** (↑) with modern pretrained backbones.

Method/Backbone	**RemoteCLIP** [8]	**SkySense++** [24]
Baseline (Backbone + Standard Head)	53.32	54.25
+HySeg (Ours)	**54.85**	**55.50**

7. Comparison with Lightweight Structural Constraints

HySeg introduces only modest additional complexity while consistently improving segmentation performance across diverse backbones and pretrained remote sensing backbones. We also find that the proposed two-stage training strategy is important in practice, as it helps avoid prior collapse and preserves useful structural information during optimization. To further examine whether the gains of HySeg could be reproduced by simpler structural regularizers, we compare against boundary loss (*bl*) [1] and edge heads (*eh*) [19], both of which mainly encourage local smoothness or boundary refinement. As shown in Tab. 12, these lightweight alternatives provide only limited improvements, whereas HySeg yields consistently larger gains across ResNet18, ResNet50, Swin_B, LSKNet_S, and ConvNeXt_B. This result suggests that the advantage of HySeg does not come from simple local regularization alone, but from coupling generative structural priors with topology-aware posterior reasoning.

Table 12. HySeg vs. lightweight constraints (LoveDA, mIoU↑).

Method/Backbone	**ResNet18**	**ResNet50**	**Swin_B**	**LSKNet_S**	**ConvNeXt_B**
Baseline (Backbone+Standard Head)	52.42	52.52	52.03	54.01	52.77
+ *boundary loss (bl)*	52.80	52.66	52.74	54.23	53.26
+ *edge heads (eh)*	52.67	52.70	52.80	54.30	53.51
+ *bl* & *eh*	53.02	52.94	53.24	54.57	54.00
HySeg (Ours)	**54.48**	**54.74**	**54.46**	**55.37**	**55.63**

References

[1] Shubhankar Borse, Ying Wang, Yizhe Zhang, and Fatih Porikli. InverseForm: A loss function for structured boundary-aware segmentation. In *CVPR*, 2021. 1, 8

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834– 848, 2018. 1

[3] Liang-Chieh Chen, Yukuan Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder–decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 3, 4, 5

[4] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 3, 4, 5

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[6] Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Jianlin Su, Libo Wang, and Peter M. Atkinson. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:5607713, 2021. 3, 4, 5

[7] Yuxuan Li, Xiang Li, Yimain Dai, Qibin Hou, Yongxiang Liu Li Liu, Ming-Ming Cheng, and Jian Yang. LSKNet: A foundation lightweight backbone for remote sensing. *IJCV*, 133:1410–1431, 2025. 1, 3, 4, 5

[8] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:5622216, 2024. 8

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

[10] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *CVPR*, 2022. 1

[11] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Miicheal Ying Yang. UAVid: A semantic segmentation dataset for uav imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165:108–119, 2020. 1

[12] Xianping Ma, Qianqian Wu, Xingyu Zhao, Xiaokang Zhang, Man-On Pun, and Bo Huang. SAM-assisted remote sensing imagery semantic segmentation with object and boundary constraints. *IEEE Transactions on Geoscience and Remote Sensing*, 62:5636916, 2024. 3, 4, 5

[13] Xiaowen Ma, Rongrong Lian, Zhenkai Wu, Hongbo Guo, Fan Yang, Mengting Ma, Sensen Wu, Zhenhong Du, Wei Zhang, and Siyang Song. LOGCAN++: Adaptive local–global class-aware network for semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 63:4404216, 2025. 3, 4, 5

[14] Martina Pastorino, Giovanni Poggi, Giuseppe Scarpa, and Luisa Verdoliva. CRFNet: A deep convolutional network to learn the potentials of a CRF for the semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1

[15] Franz Rottensteiner, Günter Sohn, Jürgen Jung, Markus Gerke, Christian Baillard, Sergio Benitez, and Uwe Breitkopf. International Society for Photogrammetry and Remote Sensing, 2D semantic labeling contest. Accessed: Mar. 1, 2024. 1

[16] Junao Shen, Qiyun Hu, Tian Feng, Xinyu Wang, Hui Cui, Sensen Wu, and Wei Zhang. In2nect: Inter-class and intra-class neural collapse tuning for semantic segmentation of imbalanced remote sensing images. In *AAAI*, 2025. 3, 4, 5

[17] Pengfei Song, Jinjiang Li, Zhiyong An, Hui Fan, and Limwei Fan. CTMFNet: Cnn and transformer multiscale fusion network of remote sensing urban scene imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:5900314, 2022. 3, 4, 5

[18] Romain Strudel, Ruben Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 3, 4, 5

[19] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *ICCV*, 2019. 1, 8

[20] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 1

[21] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. LoveDA: A remote sensing land-cover dataset for domain-adaptive semantic segmentation. In *NeurIPS*, 2021. 1

[22] Libo Wang, Rui Li, Chenxi Duan, Ce Zhang, Xiaoliang Meng, and Shenghui Fang. A novel transformer-based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19:6506105, 2022. 3, 4, 5

[23] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M. Atkinson. 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. 3, 4, 5

[24] Kang Wu, Yingying Zhang, Lixiang Ru, Bo Dang, Jiangwei Lao, Lei Yu, Junwei Luo, Zifan Zhu, Yue Sun, Jiahao Zhang, Qi Zhu, Jian Wang, Ming Yang, Jingdong Chen, Yongjun Zhang, and Yansheng Li. A semantic-enhanced multi-modal remote sensing foundation model for earth observation. *Nature Machine Intelligence*, 7:1235–1249, 2025. 8

[25] Xuechao Zou, Yue Li, Shun Zhang, Kai Li, Shiying Wang, Pin Tao, Junliang Xing, and Congyan Lang. Dynamic dictionary learning for remote sensing image segmentation. In *ICCV*, 2025. 3, 4, 5