remote sensing

MDPI

Article

UGDMoE: An Uncertainty-Guided Mixture-of-Experts Decoder for Open-Vocabulary Remote Sensing Segmentation

Wenqiu Qu [1], Guifei Jing [1,*], Qiang Yuan [2], Zhushenyu Guo [3] and Jianfeng Zhang [4]

[1] Hangzhou International Innovation Institute, Beihang University, Hangzhou 311115, China; qwq17@buaa.edu.cn
[2] Hangzhou Institute of Communications Planning Design & Research Co., Ltd., Hangzhou 311121, China; yuanqiang@hzjt.com
[3] Hangzhou Low-Altitude Economy Industry Development Co., Ltd., Hangzhou Industrial Investment Group Co., Ltd., Hangzhou 311121, China; guozsy@hz-dk.cn
[4] Wasu Media Network Co., Ltd., Hangzhou 311121, China; zhangjianfeng@wasu.com
* Correspondence: guifeijing@buaa.edu.cn

Highlights

What are the main findings?

- We propose UGDMoE, an uncertainty-guided mixture-of-experts framework for open-vocabulary remote sensing segmentation, which jointly uses an annotation-free alignment-risk cue for expert routing, logit calibration, and structure-aware refinement.
- Extensive cross-dataset experiments on eight OVRSIS benchmarks show that UGDMoE achieves state-of-the-art or highly competitive performance and remains robust under different training sources.

What are the implications of the main findings?

- The results indicate that dense open-vocabulary remote sensing segmentation benefits from geometry-specialized decoding, especially in scale-mixed scenes containing thin structures, medium-scale objects, and large homogeneous regions.
- The proposed alignment-risk cue provides an effective way to stabilize pixel–text matching without extra annotations, offering a practical direction for more reliable open-vocabulary geospatial perception.

Abstract

Rapid urbanization and the rapid accumulation of multi-source and multi-temporal Earth observation data are creating an increasing demand for remote sensing models that can flexibly support fine-grained monitoring beyond fixed label taxonomies. Open-vocabulary remote sensing image semantic segmentation (OVRSIS) aims to segment text-specified categories beyond a fixed label space with vision–language foundation models. However, dense remote sensing scenes make pixel–text matching highly vulnerable to semantic confusion and misalignment, owing to extreme scale variation, thin structures, repetitive textures, and prompt sensitivity. To address these challenges, we propose UGDMoE, an uncertainty-guided mixture-of-experts framework for OVRSIS. First, we design a domain-specific MoE decoder with three geometrically specialized experts—for slender structures, mid-scale objects, and large-region context—routed by the alignment-risk cue U_0. Second, we introduce a lightweight prompt–response estimation strategy that quantifies prediction dispersion across semantically equivalent prompts to derive U_0 in an annotation-free manner. Third, we develop prompt ensemble-based likelihood calibration (PELC), which takes the shared alignment-risk cue U_0 as input to calibrate prompt-specific logits before refinement. Finally, we design a lightweight uncertainty-aware structure refinement module that,

Check for updates

Academic Editors: Tao Hu, Yan Zhang and Peixiao Wang

Received: 19 March 2026
Revised: 15 April 2026
Accepted: 21 April 2026
Published: 28 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

guided by U_0, selectively fuses early visual features with segmentation logits to restore boundary continuity and connectivity of thin structures. We conduct extensive experiments on eight OVRSIS benchmarks under cross-dataset evaluation protocols. Trained on DLRSD, it achieves 46.97 m-mIoU and 63.31 m-mACC, surpassing the strongest baseline by 0.76 and 0.62 points; trained on iSAID, it reaches 37.47 m-mIoU and 58.52 m-mACC, improving over the strongest competitor by 0.71 and 0.61 points. UGDMoE consistently achieves state-of-the-art performance and remains robust under training-source changes.

Keywords: open-vocabulary semantic segmentation; mixture-of-experts; vision–language model; multimodal remote sensing

1. Introduction

The world is witnessing an unprecedented rate of urbanization, which intensifies the demand for timely and fine-grained urban environment monitoring from remote sensing imagery, such as urban land use and land-cover mapping and infrastructure assessment [1–3]. Meanwhile, the rapid growth of Earth observation platforms and the availability of multi-source, multi-temporal data have made remote sensing imagery widely accessible, yet large-scale interpretation remains constrained by the cost and inflexibility of pixel-level annotation. In real applications, the semantic inventory is rarely fixed: operational needs evolve, new land-cover types and man-made objects emerge, and analysts often require concept-driven queries that go beyond a pre-defined taxonomy. These realities motivate open-vocabulary remote sensing semantic segmentation (OV-RSS), which aims to segment image regions according to free-form textual descriptions while generalizing to categories unseen during training [4–6].

Recent vision–language foundation models, notably CLIP [7–9], have established a compelling paradigm for open-vocabulary recognition by aligning visual and textual representations within a unified embedding space. A prevailing strategy for OV-RSS is to transfer this image-level alignment to dense prediction tasks by matching pixel-wise or region-wise embeddings against text prototypes. However, pipelines optimized for natural images often yield suboptimal results in the geospatial domain. Unlike natural scenes, aerial and satellite imagery are characterized by extreme scale variation, where tiny instances (e.g., vehicles) and slender topological structures (e.g., roads and rivers) co-exist with vast homogeneous regions (e.g., water bodies and cropland). These properties, coupled with repetitive textures and high inter-class similarity, exacerbate semantic ambiguity and structural fragmentation during dense inference [10,11].

Existing OV-RSS methodologies generally fall into two paradigms. Training-free approaches attempt to unlock the intrinsic localization capabilities of frozen vision–language models through zero-shot prompt engineering or self-attention refinement [12–15]. While computationally efficient, these methods often suffer from coarse boundaries due to the patch-level nature of CLIP attention maps, limiting their precision on fine-grained geospatial targets. Training-based approaches instead introduce learnable adapters, auxiliary decoders, or prompt tuning mechanisms to bridge the domain gap. Representative methods such as SAN [16] and Cat-Seg [17] refine visual–semantic alignment using side networks or cost aggregation, while domain-specific extensions [18,19] further incorporate remote sensing priors or domain generalization strategies [20]. Despite these advances, most existing OV-RSS methods still rely on decoder designs largely inherited from natural-image segmentation, typically using a single decoding pathway for all pixels. This design is not well matched to scale-mixed remote sensing scenes, where thin and elongated struc-

tures, medium-scale objects, and large homogeneous regions place conflicting demands on dense feature decoding. As a result, models often face a trade-off between preserving high-frequency structural details and maintaining low-frequency semantic consistency over large areas.

To address these issues, we propose UGDMoE, an uncertainty-guided mixture-of-experts framework for open-vocabulary remote sensing segmentation. Rather than relying on a single decoding pathway, UGDMoE introduces an explicit alignment-risk cue to coordinate geometry-specialized expert decoding, prompt-level calibration, and structure-aware refinement, thereby improving dense prediction in scale-mixed and semantically ambiguous remote sensing scenes.

The rest of this paper is organized as follows. Section 2 reviews the related literature. Section 3 details the proposed UGDMoE framework. Section 4 presents the experimental settings, quantitative comparisons, and ablation analyses. Section 5 concludes the paper.

Our contributions are summarized as follows:

- **An uncertainty-guided expert decoding framework for OVRSIS.** We propose UGDMoE, an uncertainty-guided mixture-of-experts framework for open-vocabulary remote sensing image semantic segmentation. The framework introduces geometry-specialized expert branches to handle scale-mixed remote sensing scenes, where thin structures, medium-scale objects, and large homogeneous regions impose different decoding requirements.
- **Annotation-free alignment-risk estimation via synonym-consistent prompting.** We introduce a lightweight prompt–response estimation strategy that measures prediction dispersion across semantically equivalent prompts and derives an annotation-free alignment-risk cue U_0. This cue provides an explicit signal for identifying spatial regions with unstable pixel–text alignment.
- **Unified use of alignment risk for decoding, calibration, and refinement.** Based on the shared cue U_0, we develop a risk-guided decoding and refinement pipeline, including expert routing, prompt-ensemble likelihood calibration, and a lightweight structure-aware refinement module. This design improves prediction stability in ambiguous regions and helps recover the continuity of thin and elongated targets.
- **Extensive cross-dataset evaluation on OVRSIS benchmarks.** We conduct comprehensive experiments on eight open-vocabulary remote sensing image segmentation benchmarks under cross-dataset evaluation settings. The results show that UGDMoE achieves competitive and often state-of-the-art performance, while maintaining good robustness across different training sources.

2. Related Work

2.1. Remote Sensing Image Semantic Segmentation

Remote sensing image semantic segmentation (RSISS) has been predominantly studied under a closed-set setting with fixed semantic taxonomies. Representative benchmarks such as LoveDA [21] and iSAID [22] have supported steady progress in this area, while methods such as FarSeg [23] and AerialFormer [24] have improved multi-scale modeling and global contextual reasoning for dense prediction. To alleviate annotation scarcity and domain shift, prior studies have also explored semi-supervised learning, few-shot learning, weak supervision, and unsupervised domain adaptation [25–28]. However, these efforts still operate within closed vocabularies and therefore do not directly address the requirement of segmenting arbitrary categories specified at test time.

This limitation is also reflected at the dataset level. Although FLAIR [29] and SAMRS [30] enlarge the scale of supervision, they are not explicitly designed for open-vocabulary evaluation. Their label spaces remain constrained, and they provide limited

support for studying robust pixel–text alignment on unseen concepts. These gaps motivate open-vocabulary remote sensing segmentation, where the model must generalize beyond pre-defined categories while remaining reliable in structurally complex scenes.

2.2. Open-Vocabulary Remote Sensing Image Semantic Segmentation

Open-vocabulary semantic segmentation (OVSS) aims to segment arbitrary categories without requiring explicit training samples for every class. Recent progress in this direction has been largely driven by vision–language models, especially CLIP [31], which enable zero-shot transfer through aligned visual and textual representations. Existing CLIP-based OVSS methods can be broadly divided into training-required and training-free paradigms. Training-required methods adapt frozen vision–language backbones with additional modules or partial fine-tuning [32–34]; representative examples include LSeg [35] and SAN [16]. By contrast, training-free methods attempt to exploit the intrinsic localization ability of CLIP through lightweight architectural adjustment or self-attention refinement, as explored in MaskCLIP [36], TagCLIP [15], and related works [37]. Another important line of research adopts a two-stage pipeline that first generates category-agnostic masks and then classifies them with text embeddings [38–40], often with additional semantic priors or contextual calibration [41].

In remote sensing, recent methods such as TPOV-Seg [18] and FreeMix [19] have begun to transfer open-vocabulary segmentation to geospatial imagery. Nevertheless, most existing approaches still rely on a single monolithic decoder inherited from natural-image segmentation. This design is not well matched to remote sensing scenes, where small objects, thin structures, and large homogeneous regions coexist within the same image. As a result, high-frequency structural details and large-scale semantic consistency are often difficult to preserve simultaneously. Our method is built on this observation and addresses the problem through structurally specialized decoding.

2.3. Mixture of Experts for Remote Sensing

The mixture-of-experts (MoE) framework, originally proposed by Jacobs et al. [42], enables dynamic computation by routing inputs to specialized subnetworks (“experts”) via trainable gating mechanisms. Over the past decade, MoE has achieved remarkable success across diverse domains [43]. In multi-task learning, MMoE [44] introduces task-specific gates over shared experts to model inter-task relationships. In natural language processing, ST-MoE [45] demonstrates that sparse expert models can be both scalable and robust, while OpenMoE [46] represents an early attempt toward open-source large language models based on the MoE paradigm. In computer vision, V-MoE [47] scales vision transformers to billions of parameters through conditional token routing, while in multimodal learning, LIMoE [48] and MoE-LLaVA [49] leverage modality-aware expert selection to improve alignment and efficiency in vision-language tasks.

Despite these advances, the application of MoE in remote sensing remains nascent. Early efforts such as MV-MoE [50] and SW-MoE [51] explore MoE for single-modality remote sensing tasks—MV-MoE employs attention-based expert routing for multi-view image representation, while SW-MoE activates scale-specific experts to handle scene complexity and data imbalance. More recently, RS-MoE [52] adopts an MoE architecture to fuse heterogeneous data sources (e.g., optical, SAR, LiDAR), and RSUniVLM [53] utilizes distinct experts to specialize in different downstream tasks such as captioning and visual question answering. These works highlight the versatility of MoE in the remote sensing context.

However, existing MoE adaptations overlook the intra-image geometric variance inherent in segmentation tasks. A single satellite image is often a mosaic of conflicting geometric regimes, where fine-grained infrastructure co-exists with coarse land-cover

patterns. Conventional global or task-specific routing fails to adaptively process this structural diversity within a single scene. We are the first to introduce a Geometric-Granularity MoE paradigm to remote sensing segmentation. Instead of specializing by modality or downstream task, our experts specialize by geometric topology—differentiating between high-frequency boundary preservation and low-frequency contextual smoothing. This allows us to resolve the scale-mixed dilemma by dynamically routing pixels to experts that match their local structural properties, a capability fundamentally absent in prior monolithic or task-level MoE frameworks.

3. Method

3.1. Overview

Open-vocabulary remote sensing semantic segmentation (OV-RSS) aims to predict pixel-wise masks for categories specified by text, including categories beyond the supervised training label space. This setting is challenging in remote sensing imagery because objects of the same category may appear at very different scales, thin and elongated structures are easily fragmented during dense prediction, and repetitive textures often reduce the discriminability of pixel–text matching. In addition, semantically equivalent prompts can still produce different prediction maps, indicating that pixel–text alignment is spatially non-uniform and sensitive to prompt formulation.

As shown in Figure 1, UGDMoE uses cross-prompt prediction consistency to derive an explicit alignment-risk signal for subsequent decoding and refinement. We first extract high-resolution multi-scale visual features with a CLIP-based backbone and a lightweight neck (Section 3.2), and inject textual priors through query-guided fusion (QGF) to obtain query-conditioned features (Section 3.3). A lightweight prompt–response estimation branch is then applied to the query-conditioned features to estimate a prompt-dispersion risk map U_0. This risk map is incorporated into three subsequent stages: risk-aware expert routing in the mixture-of-experts decoder (Section 3.4), spatially adaptive logit calibration in PELC (Section 3.5), and gated structure refinement for thin and fragile regions (Section 3.6). The prompt–response estimation branch is a lightweight auxiliary branch attached to the query-conditioned features, and its sole purpose is to provide the shared alignment-risk cue U_0 for subsequent decoding, calibration, and refinement. The resulting pipeline couples prompt-level alignment uncertainty with pixel-level decoding and structural correction.

3.2. VLM Encoders and Multi-Scale Neck

We build UGDMoE on top of a pretrained vision–language model (VLM) to preserve open-vocabulary transferability while adapting feature extraction to dense prediction in remote sensing imagery. Specifically, we use CLIP with a ViT-B/16 image encoder E_{img} and a text encoder E_{text} as the semantic backbone, and augment the visual branch with high-resolution dense feature extraction and multi-scale aggregation.

Remote sensing images are typically high-resolution and contain abundant fine-grained geometric patterns. To preserve such details, we interpolate the positional embeddings of the ViT-based image encoder to match the input resolution and extract intermediate representations from multiple transformer stages. In our implementation, we use features from three representative stages (early/mid/late), i.e., the outputs of the 4th, 8th, and 12th transformer blocks of ViT-B/16, and project them to a unified channel dimension.

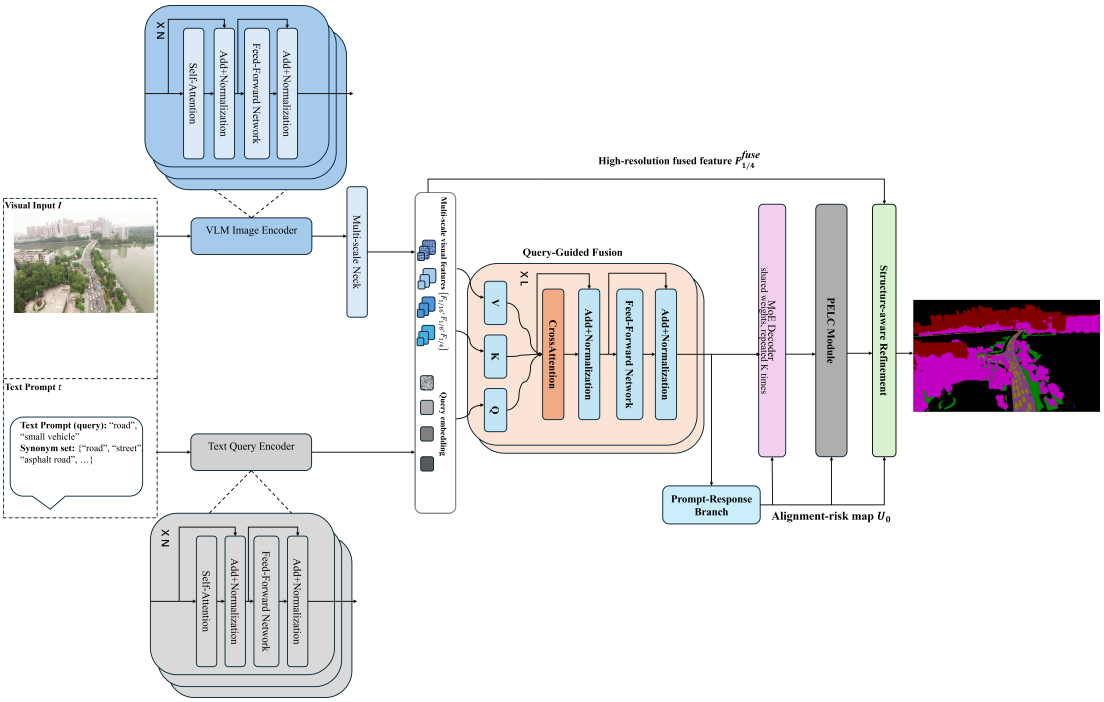


Figure 1. Architecture of UGDMoE. Given an input image and text prompts, the framework first extracts multi-scale visual features and query-conditioned features through the VLM encoders, multi-scale neck, and query-guided fusion. A lightweight prompt–response branch estimates the shared alignment-risk map U_0, which is then reused by the risk-aware MoE decoder, prompt-ensemble logit calibration (PELC), and structure-aware refinement module to produce the final segmentation prediction.

These multi-level features are aggregated by a lightweight FPN-style neck with lateral projections and top-down fusion. Each selected stage feature is first projected to a common channel dimension $C = 256$ using a 1×1 projection layer (or an equivalent token-space projection before reshaping), and then fused to produce two resolution-aligned feature maps:

$$F_s^{\text{fuse}} = \text{Neck}\big(E_{\text{img}}(I)\big), \qquad F_s^{\text{fuse}} \in \mathbb{R}^{H_s \times W_s \times C}, \quad s \in \left\{\tfrac{1}{8}, \tfrac{1}{4}\right\}, \quad C = 256. \tag{1}$$

The coarser feature $F_{1/8}^{\text{fuse}}$ provides stronger semantic context, whereas the higher-resolution feature $F_{1/4}^{\text{fuse}}$ preserves structural detail for thin or elongated targets such as roads, rivers, and field boundaries.

On the text side, we use a synonym-consistent prompt ensemble to support open-vocabulary inference and to estimate cross-prompt prediction variability. For each category $c \in \mathcal{V}$, we construct a prompt set $\{g_k(c)\}_{k=1}^{K}$ consisting of standard templates and curated synonym variants describing the same semantic concept. In the default setting, $K = 4$, including two template-based prompts and two synonym-based prompts after template normalization. For the category *road*, the four prompts are "a remote sensing image of a road", "a satellite image containing a road", "a remote sensing image of a highway", and "a remote sensing image of a street". The same template construction is applied to all categories using their corresponding category names and curated synonym lists.

The text encoder produces normalized prompt embeddings:

$$q_{k,c} = \text{Norm}\big(E_{\text{text}}(g_k(c))\big), \qquad q_{k,c} \in \mathbb{R}^{d_t}, \tag{2}$$

where d_t denotes the CLIP text embedding dimension. The multi-scale visual features $\{F_{1/8}^{\text{fuse}}, F_{1/4}^{\text{fuse}}\}$ and prompt embeddings $\{q_{k,c}\}$ are then used by the query-guided fusion module to produce prompt-conditioned features for prompt-specific decoding.

3.3. Query-Guided Fusion via Cross-Attention

Pixel–text association in OV-RSS is often unstable in remote sensing imagery because repetitive textures, subtle inter-class differences, and strong scale variation make naive visual–text matching unreliable. To condition visual features on the queried concept before decoding, we introduce query-guided fusion (QGF), which injects textual priors into the multi-scale visual stream through a lightweight cross-attention operation.

Given a scale $s \in \left\{\frac{1}{8}, \frac{1}{4}\right\}$, we flatten $F_s^{\text{fuse}} \in \mathbb{R}^{H_s \times W_s \times C}$ into a token sequence $V_s \in \mathbb{R}^{N_s \times C}$, where $N_s = H_s W_s$. For each prompt embedding $q_{k,c} \in \mathbb{R}^{d_t}$, QGF first projects the text embedding to the visual feature dimension:

$$\bar{q}_{k,c} = W_q q_{k,c} \in \mathbb{R}^C. \tag{3}$$

We then compute a lightweight cross-attention response from the visual tokens:

$$K_s = V_s W_k, \qquad V_s' = V_s W_v, \tag{4}$$

$$a_{s,k,c} = \text{Softmax}\left(\frac{\bar{q}_{k,c} K_s^\top}{\sqrt{C}}\right) \in \mathbb{R}^{1 \times N_s}, \tag{5}$$

$$\Delta v_{s,k,c} = a_{s,k,c} V_s' \in \mathbb{R}^{1 \times C}. \tag{6}$$

The response is broadcast to the token grid and fused with the original tokens through a residual projection:

$$\tilde{V}_{s,k,c} = V_s + \Pi_s(\Delta v_{s,k,c}), \tag{7}$$

where $\Pi_s(\cdot)$ denotes a learned projection/broadcast operator matching the sequence length and feature dimension. Finally, $\tilde{V}_{s,k,c}$ is reshaped back to the 2D layout:

$$\tilde{F}_{s,k,c} = \text{Reshape}\left(\tilde{V}_{s,k,c}\right) \in \mathbb{R}^{H_s \times W_s \times C}. \tag{8}$$

We apply QGF independently at both scales using scale-specific parameters. The resulting pair $\{\tilde{F}_{1/8,k,c}, \tilde{F}_{1/4,k,c}\}$ provides query-conditioned features for two subsequent uses. First, the query-conditioned features are fed into a lightweight prompt–response estimation branch, which produces preliminary prompt-specific class responses $\{P_{k,c}^{\text{pre}}(x)\}_{k=1}^K$ and yields the shared alignment-risk map U_0. Second, the same query-conditioned features are passed to the MoE decoder for risk-aware routing and final dense prediction.

3.4. Risk-Guided MoE Decoder

Remote sensing scenes are strongly scale-mixed: thin linear structures, medium-sized objects, and large homogeneous regions often coexist within the same image. A single decoding pathway with a fixed feature transformation is therefore insufficient: it tends to over-smooth slender structures while failing to maintain stable semantics over large areas. We address this with a mixture-of-experts decoder that adaptively allocates decoding capacity across different geometric regimes, as illustrated in Figure 2.

The decoder contains three parallel expert branches (default $E = 3$), denoted as Expert-S, Expert-M, and Expert-L. In our implementation, the three experts share the same lightweight topology but maintain independent parameters. They are intended to provide complementary decoding preferences for slender/local-detail, intermediate-scale, and large-context patterns.

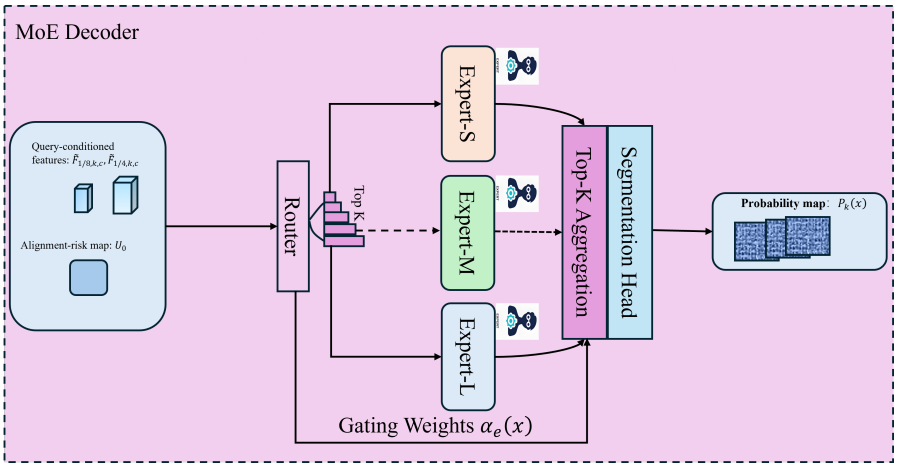


Figure 2. Architecture of the risk-guided MoE decoder. The decoder takes query-conditioned multi-scale features $\{\tilde{F}_{1/8,k,c}, \tilde{F}_{1/4,k,c}\}$ together with the alignment-risk map U_0 as input. A lightweight router predicts location-dependent Top-k expert weights, and the outputs of the three experts are aggregated before the shared segmentation head produces prompt-specific predictions.

Before expert routing, we apply a lightweight prompt–response estimation branch to the query-conditioned features and obtain preliminary prompt-specific probability maps $\{P^{\text{pre}}_{k,c}(x)\}_{k=1}^{K}$ for the queried category c. We quantify their cross-prompt dispersion by

$$U_0(x) = \text{Var}_k\Big(P^{\text{pre}}_{k,c}(x)\Big), \tag{9}$$

which defines an explicit alignment-risk map. High values of U_0 indicate spatial locations where prompt agreement is weak and pixel–text alignment is unstable.

For each prompt-category pair (k, c), the router predicts location-dependent mixing weights by combining three signals: high-resolution structural cues from $\tilde{F}_{1/4,k,c}$, coarser semantic context from $\tilde{F}_{1/8,k,c}$, and the alignment-risk map U_0. We first upsample $\tilde{F}_{1/8,k,c}$ to the spatial resolution of $\tilde{F}_{1/4,k,c}$ and then fuse the three inputs:

$$R_{k,c} = \psi\big(\tilde{F}_{1/4,k,c},\, \text{Up}(\tilde{F}_{1/8,k,c}),\, U_0\big), \tag{10}$$

where $\psi(\cdot)$ denotes channel projection followed by feature fusion. The inclusion of U_0 provides an explicit risk prior for expert selection in ambiguous regions. A routing head $G(\cdot)$ then produces expert logits at each spatial location:

$$A(x) = G(R_{k,c})(x) \in \mathbb{R}^E. \tag{11}$$

We use sparse Top-k routing in both training and inference. Let $\Omega(x) \subseteq \{1, \ldots, E\}$ denote the selected expert index set at location x, obtained from the Top-k entries of $A(x)$ (default $k = 2$). The routing weights are normalized by a softmax over the selected experts:

$$\alpha_e(x) = \frac{\exp(A_e(x))}{\sum_{j\in\Omega(x)} \exp(A_j(x))}, \qquad e \in \Omega(x), \tag{12}$$

and $\alpha_e(x) = 0$ for $e \notin \Omega(x)$.

Each expert processes the same query-conditioned multi-scale features:

$$Z_e(x) = E_e\big(\tilde{F}_{1/8,k,c},\, \tilde{F}_{1/4,k,c}\big)(x), \qquad e = 1, \ldots, E, \tag{13}$$

where $Z_e(x) \in \mathbb{R}^{C_d}$ denotes the decoded feature produced by the e-th expert at location x. The final decoded feature is obtained by weighted aggregation:

$$\hat{F}_{k,c}(x) = \sum_{e=1}^{E} \alpha_e(x)\, Z_e(x) = \sum_{e \in \Omega(x)} \alpha_e(x)\, E_e\big(\tilde{F}_{1/8,k,c},\, \tilde{F}_{1/4,k,c}\big)(x). \tag{14}$$

The aggregated feature $\hat{F}_{k,c}$ is then passed to a shared segmentation head H_{seg}:

$$\mathbf{S}_k(x) = H_{\text{seg}}\big(\hat{F}_{k,c}(x)\big), \qquad \mathbf{P}_k(x) = \text{Softmax}\big(\mathbf{S}_k(x)\big), \tag{15}$$

where $\mathbf{S}_k(x) \in \mathbb{R}^{|\mathcal{V}|}$ and $\mathbf{P}_k(x) \in \mathbb{R}^{|\mathcal{V}|}$ denote the prompt-specific logits and probabilities, respectively. These prompt-specific logits and probabilities correspond to the MoE-decoded final predictions and are used in the subsequent calibration stage. The alignment-risk map U_0 is estimated earlier from the preliminary prompt-specific responses defined in Equation (9).

The labels "S/M/L" denote intended geometric preferences rather than hard role constraints imposed by an explicit specialization loss. The behavior of the three branches is analyzed in the experiments through expert activation maps and region-wise routing statistics on thin-structure categories.

3.5. Prompt-Ensemble Logit Calibration

Prompt formulation affects dense open-vocabulary predictions. Even when multiple prompts are constructed to describe the same semantic category, their predictions may diverge in visually ambiguous regions. We therefore use the previously estimated alignment-risk cue U_0 to calibrate prompt-specific logits before refinement. The overall architecture of the prompt-ensemble logit calibration module is shown in Figure 3.

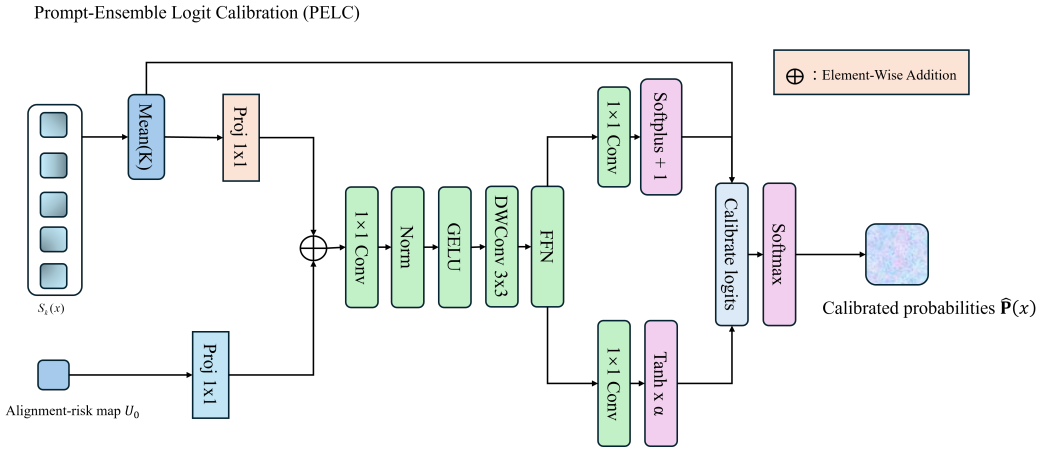


Figure 3. Prompt-ensemble logit calibration (PELC). Prompt-specific logits $\{\mathbf{S}_k(x)\}_{k=1}^{K}$ are aggregated and combined with the externally estimated alignment-risk map U_0. A lightweight calibration head predicts spatially adaptive scaling and bias terms to produce calibrated logits and calibrated probabilities.

For a queried concept c, Equation (9) defines the alignment-risk map U_0 from the preliminary class-specific responses $\{P_{k,c}^{\text{pre}}(x)\}_{k=1}^{K}$. In parallel, we aggregate prompt-specific logits by averaging over the prompt ensemble:

$$\mathbf{S}(x) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{S}_k(x), \tag{16}$$

where $\mathbf{S}_k(x) \in \mathbb{R}^{|\mathcal{V}|}$ denotes the prompt-specific logit vector and $\mathbf{S}(x) \in \mathbb{R}^{|\mathcal{V}|}$ denotes the ensemble-averaged logits. In PELC, calibration is performed on the prompt-specific logits $\{\mathbf{S}_k(x)\}_{k=1}^{K}$, while the previously estimated U_0 is used as an external spatial risk cue.

We then calibrate the logits using a lightweight calibration head $\phi(\cdot)$:

$$[a(x), b(x)] = \phi(\mathbf{S}(x), U_0(x)), \tag{17}$$

$$\hat{\mathbf{S}}(x) = a(x)\,\mathbf{S}(x) + b(x), \tag{18}$$

$$\hat{\mathbf{P}}(x) = \mathrm{Softmax}\big(\hat{\mathbf{S}}(x)\big). \tag{19}$$

Here, $\phi(\cdot)$ is implemented as a shallow convolutional predictor. In the default setting, $a(x)$ and $b(x)$ are scalar fields shared across channels for computational efficiency.

The calibrated probability map of the queried concept is denoted by

$$\hat{P}(x) \triangleq \hat{P}_c(x), \tag{20}$$

where $\hat{P}_c(x)$ is extracted from the calibrated probability vector $\hat{\mathbf{P}}(x)$. In our implementation, the prompt sets are constructed from fixed templates and curated synonym lists to reduce semantic drift across prompts.

3.6. Structure-Aware Refinement

Even after MoE decoding and PELC-based calibration, thin and elongated targets (e.g., roads and rivers) may still exhibit local breaks, jagged boundaries, and topological discontinuities due to mixed pixels, occlusions, and sensor noise. To correct these errors, we introduce a lightweight structure-aware refinement stage.

As shown in Figure 4, the refinement module takes three inputs: the calibrated probability map of the queried category $\hat{P}$, the high-resolution fused feature $F_{1/4}^{\mathrm{fuse}}$, and the alignment-risk map U_0. The update is applied selectively rather than uniformly: regions with high prompt-dispersion risk receive stronger structural correction, while low-risk regions are updated only weakly. We implement this stage using an Edge-Conditioned Gated Propagation (ECGP) block. Boundary and connectivity cues are extracted from the high-resolution fused feature $F_{1/4}^{\mathrm{fuse}}$, whereas the spatial gating signal is predicted solely from U_0. Boundary-sensitive features are first extracted from $F_{1/4}^{\mathrm{fuse}}$ using a lightweight convolutional projection. A gating map is then predicted solely from U_0 and used to modulate the propagation strength at each spatial location. The propagated feature update is fused with the current prediction through a residual refinement step.

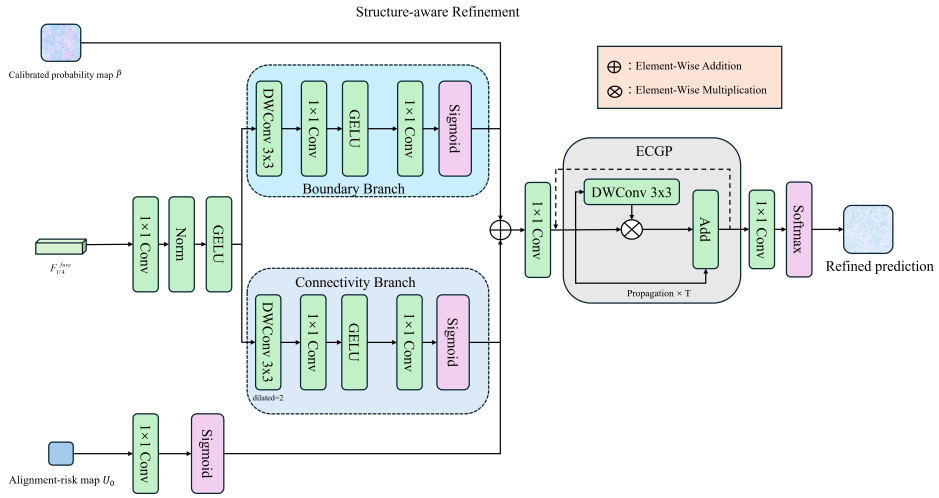


Figure 4. Structure-aware refinement module. The refinement module takes the calibrated queried-class probability map $\hat{P}$, the high-resolution fused feature $F_{1/4}^{\mathrm{fuse}}$, and the alignment-risk map U_0 as inputs. Guided by U_0, the module enhances boundary quality and connectivity of thin structures through gated propagation.

The propagation is applied for a fixed number of iterations ($T = 2$ in the default setting). The refinement stage outputs the final refined logits M, which are supervised by the segmentation loss defined in Section 3.7.

3.7. Loss Function

UGDMoE is trained with standard pixel-level supervision on the available training label space. Let $\mathcal{C}_{\text{train}}$ denote the supervised category set used during training, and let $Y \in \{0, \ldots, |\mathcal{C}_{\text{train}}| - 1\}^{H \times W}$ denote the ground-truth semantic mask. We use Ω_{valid} to denote the set of valid pixels used for loss computation. Although the alignment-risk map U_0 is estimated without annotations, the full model is optimized in the supervised label space $\mathcal{C}_{\text{train}}$.

Let $M \in \mathbb{R}^{H \times W \times |\mathcal{C}_{\text{train}}|}$ denote the final refined logits produced by the refinement stage. The primary segmentation loss is

$$L_{\text{seg}} = L_{\text{CE}}(M,\ Y; \Omega_{\text{valid}}) + \lambda_{\text{dice}}\, L_{\text{Dice}}(M,\ Y; \Omega_{\text{valid}}), \tag{21}$$

where L_{CE} enforces pixel-wise semantic discrimination and L_{Dice} improves overlap under class imbalance, particularly for thin or fragmented structures.

To stabilize the interaction between PELC and the refinement stage, we apply auxiliary supervision to the calibrated prediction before refinement. Let $\hat{\mathbf{P}} \in \mathbb{R}^{H \times W \times |\mathcal{C}_{\text{train}}|}$ denote the calibrated probability map restricted to the supervised classes for loss computation. We define

$$L_{\text{cal}} = L_{\text{CE}}(\hat{\mathbf{P}},\ Y; \Omega_{\text{valid}}) + \lambda_{\text{dice}}\, L_{\text{Dice}}(\hat{\mathbf{P}},\ Y; \Omega_{\text{valid}}). \tag{22}$$

To reduce degenerate expert usage, we employ a lightweight routing balance regularizer on the normalized routing weights. Let $\alpha_e(x)$ denote the routing weight of expert e at spatial location x (Equation (12)). The mini-batch average routing usage is computed as

$$\bar{p}_e = \frac{1}{|\Omega_{\text{batch}}|} \sum_{x \in \Omega_{\text{batch}}} \alpha_e(x), \tag{23}$$

where Ω_{batch} denotes the set of valid spatial locations in the current mini-batch. The balance loss is defined as

$$L_{\text{bal}} = \sum_{e=1}^{E} \left(\bar{p}_e - \frac{1}{E} \right)^2. \tag{24}$$

In the default setting, the training objective is

$$L = L_{\text{seg}} + \lambda_{\text{cal}} L_{\text{cal}} + \lambda_{\text{bal}} L_{\text{bal}}, \tag{25}$$

where λ_{dice}, λ_{cal}, and λ_{bal} control the contributions of the corresponding terms. In ablation studies, we disable L_{cal} or L_{bal} by setting the associated coefficient to zero.

4. Experiments

4.1. Experimental Setup

4.1.1. Benchmark

All experiments are conducted under the OVRSISBench benchmark [54], which evaluates open-vocabulary remote sensing semantic segmentation under cross-dataset transfer. OVRSISBench casts multiple remote sensing datasets into an open-vocabulary setting, where category semantics are specified by text descriptions and the model is required to generalize beyond the training label space.

To evaluate open-vocabulary generalization under realistic domain shifts, OVRSISBench adopts a cross-dataset transfer protocol: a model is trained on a large-scale source dataset and

then directly evaluated on multiple target datasets whose label spaces overlap only partially with the source vocabulary. The benchmark covers eight representative datasets—UAVid [55], VDD [56], DLRSD [57], iSAID [22], LoveDA [21], Potsdam, UDD5 [58], and Vaihingen—spanning aerial orthophotos, UAV imagery, and bird's-eye-view driving scenes. Figure 5 shows representative examples from all eight datasets and illustrates the considerable diversity in imaging perspective, scene layout, semantic taxonomy, and annotation style, which together make cross-dataset open-vocabulary transfer particularly challenging.

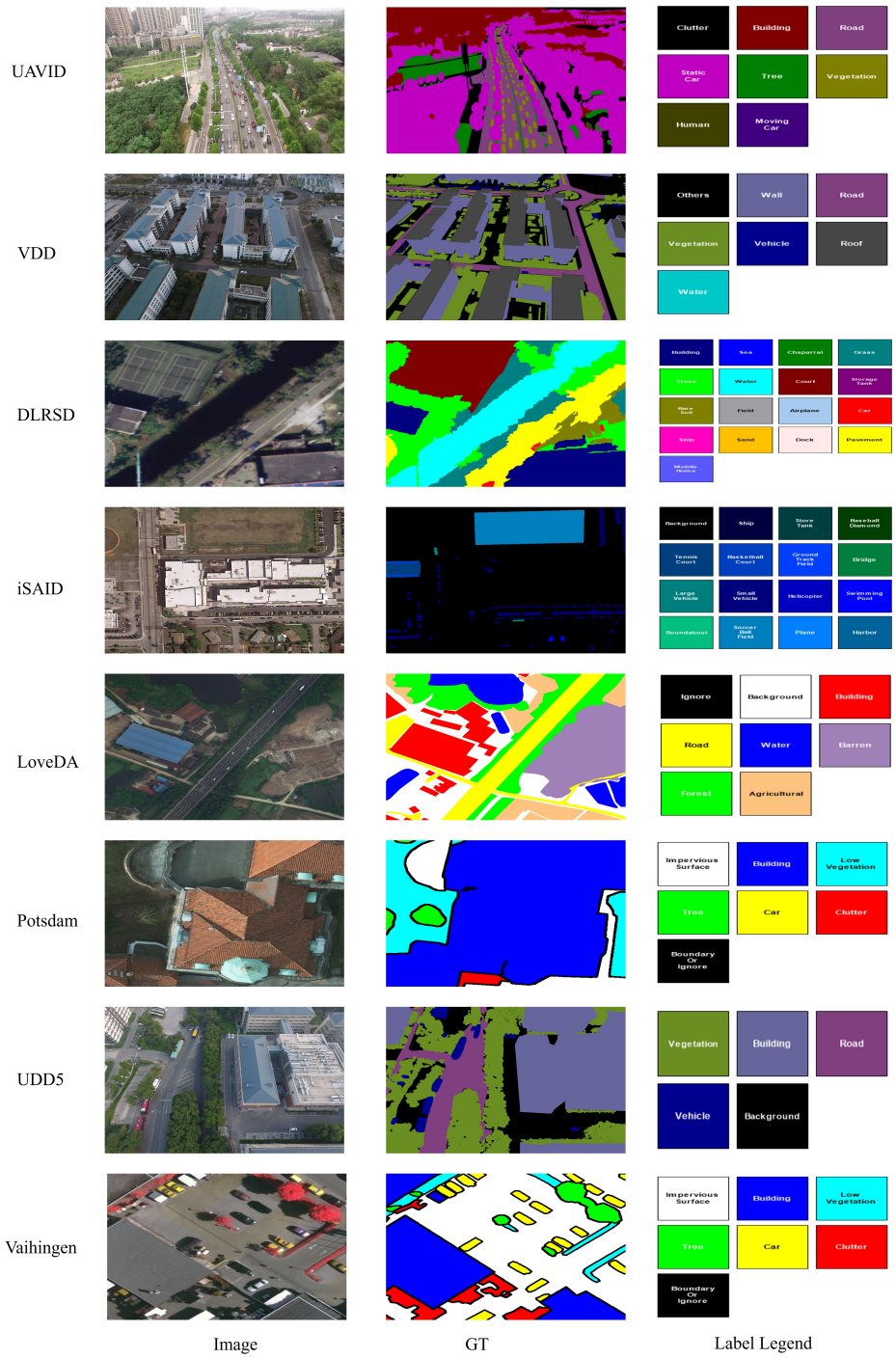


Figure 5. Representative examples of the eight datasets in OVRSISBench. For each dataset, we show one RGB image, its corresponding colorized ground-truth annotation, and the dataset-specific semantic label legend. The examples highlight substantial differences in imaging perspective, scene composition, semantic taxonomy, and annotation style across datasets, which together make cross-dataset open-vocabulary transfer particularly challenging.

Following the standard OVRSISBench protocol, we consider two training configurations, with DLRSD and iSAID used as the source dataset, respectively. In each configuration, the model is trained on the official training split of the source dataset, hyperparameters are selected on its validation split, and the trained model is then evaluated directly on the benchmark-defined evaluation splits of all target datasets without additional fine-tuning.

4.1.2. Datasets

Table 1 summarizes the dataset statistics and splits used in OVRSISBench. These datasets differ substantially in spatial resolution, imaging condition, scene composition, and semantic taxonomy, which makes cross-dataset evaluation particularly challenging.

DLRSD contains 7002 remote sensing images annotated into 17 categories, covering diverse land-cover types and man-made structures. We follow OVRSISBench and use 5601 images for training and 1401 for validation.

iSAID is a large-scale dataset with complex and densely populated scenes. It contains 24,439 images across 15 categories. In OVRSISBench, 18,076 images are used for training and 6363 for validation.

UAVid is captured from unmanned aerial vehicles and focuses on urban scene understanding. It includes 300 images with pixel-wise annotations over 8 categories. We use 200 images for training and 100 for validation.

UDD5 provides bird's-eye-view urban driving scenes with pixel-level annotations for 5 categories. It contains 4198 images in total, with 3149 images for training and 1049 for validation.

Table 1. OVRSISBench dataset statistics and splits. Potsdam and Vaihingen are used as evaluation targets in the cross-dataset transfer protocol.

Dataset	Images	Classes	Train/Val
DLRSD	7002	17	5601/1401
iSAID	24,439	15	18,076/6363
Potsdam	20,102	6	Evaluation only
Vaihingen	2254	6	Evaluation only
UAVid	300	8	200/100
UDD5	4198	5	3149/1049
LoveDA	18,000	7	14,168/3832
VDD	7992	7	5994/1998

LoveDA is a large-scale land-cover dataset collected from both rural and urban areas, consisting of 18,000 images labeled into 7 classes. We use 14,168 images for training and 3832 for validation.

VDD is designed for studying domain adaptation in remote sensing and contains 7992 images across 7 categories. We use 5994 images for training and 1998 for validation.

Potsdam and Vaihingen are widely used high-resolution aerial segmentation benchmarks. Both datasets contain 6 semantic categories and feature fine-grained structures. In OVRSISBench, they are used as target datasets for cross-dataset evaluation, and we follow the benchmark-defined evaluation split for reporting results.

4.1.3. Evaluation Metrics

We evaluate all methods using three standard metrics: mean Intersection-over-Union (mIoU), frequency-weighted Intersection-over-Union (fwIoU), and mean per-class accuracy (mAcc).

For each class, Intersection-over-Union is computed and then averaged over all n classes:

$$\text{mIoU} = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FP_i + FN_i}, \tag{26}$$

where TP_i, FP_i, and FN_i denote the number of true-positive, false-positive, and false-negative pixels for class i, respectively.

To account for class-frequency imbalance, fwIoU weights the IoU of each class by its pixel frequency:

$$\text{fwIoU} = \sum_{i=1}^{n}\frac{TP_i + FN_i}{\sum_{j=1}^{n}(TP_j + FN_j)} \cdot \frac{TP_i}{TP_i + FP_i + FN_i}. \tag{27}$$

We also report mean per-class accuracy:

$$\text{mAcc} = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FN_i}. \tag{28}$$

Unless otherwise specified, all metrics are computed on the test set and reported in percentage.

4.2. Implementation Details

The proposed framework is implemented in PyTorch 2.1.1 with Detectron2. All experiments follow the OVRSISBench protocol and use the same training and evaluation pipeline across datasets.

The visual backbone is CLIP ViT-B/16. Following Section 3, we extract three intermediate transformer features and use a lightweight FPN-style neck to produce the multi-scale feature maps $F_{1/8}^{\text{fuse}}$ and $F_{1/4}^{\text{fuse}}$. A lightweight prompt–response estimation branch is applied to the query-conditioned features to generate preliminary prompt-specific class responses, from which the alignment-risk map U_0 is computed. The MoE decoder contains three parallel expert branches (Expert-S, Expert-M, and Expert-L) with identical lightweight topology and independent parameters. Risk-aware routing is implemented by a lightweight routing head that takes fused structural and semantic features together with U_0 as input. The refinement stage takes the calibrated queried-class probability map, the high-resolution fused feature, and U_0 as inputs, as described in Section 3.6.

We optimize the model using AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$ and weight decay 0.05. A cosine learning-rate schedule is adopted with linear warm-up in the first epoch. To preserve the pretrained representation while enabling downstream adaptation, we use two learning rates: the pretrained CLIP backbone is updated with a learning rate of 2×10^{-6}, while the newly introduced modules—including the neck, QGF, prompt–response estimation branch, MoE router, expert branches, PELC head, and refinement module—are trained with a learning rate of 2×10^{-4}. Training is conducted end-to-end with a batch size of 4 for 30,000 iterations on NVIDIA A100 GPUs (80 GB).

The default objective function consists of the main segmentation loss L_{seg}, the calibration-stage auxiliary loss L_{cal}, and the routing balance regularizer L_{bal}. Their coefficients are fixed to $\lambda_{\text{cal}} = 0.3$ and $\lambda_{\text{bal}} = 0.05$ in all experiments. Unless otherwise specified, all other hyperparameters are kept unchanged across different training sources. Following the standard cross-dataset transfer setting of OVRSISBench, models trained on the source domain (DLRSD or iSAID) are evaluated directly on target benchmarks without additional fine-tuning.

4.3. Experimental Results

Tables 2 and 3 report the cross-dataset evaluation results of UGDMoE under two training sources, DLRSD and iSAID. Across both settings, UGDMoE achieves the best average performance, indicating that the proposed design improves open-vocabulary transfer beyond a single training distribution.

With DLRSD as the source dataset (Table 2), UGDMoE reaches the best overall average performance, with 46.97 m-mIoU and 63.31 m-mACC. The gains are distributed across multiple target datasets rather than concentrated on a single domain. In particular, UGDMoE improves over prior OVRSIS baselines on LoveDA, UAVid, UDD5, Vaihingen, and VDD, while remaining competitive on all remaining targets. These results are consistent with the design of the model: risk-aware routing helps allocate expert capacity to structurally fragile or semantically ambiguous regions, PELC suppresses unstable prompt-induced predictions, and the refinement stage improves the continuity of thin structures.

Table 2. Main results of UGDMoE on 8 OVRSIS datasets with DLRSD as the training set. Best results are bold, second-best are underlined. m-mIoU and m-mACC denote average performance across all datasets.

Method	Type	DLRSD		iSAID		LoveDA		Potsdam		UAVid		UDD5		Vaihingen		VDD		Avg	
		mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	m-mIoU	m-mACC
SCAN [40]	OVS	52.42	72.43	44.28	67.25	23.17	35.36	27.45	39.22	20.28	34.43	34.14	43.25	15.23	29.45	29.24	45.57	30.78	45.87
SAN [16]	OVS	86.45	91.25	49.56	67.25	25.33	37.54	37.25	46.28	23.53	38.14	37.23	48.45	39.22	48.33	35.83	53.25	41.80	53.81
SED [32]	OVS	87.68	91.24	51.23	68.24	24.55	36.83	29.35	37.95	21.33	35.64	35.73	45.15	39.02	58.62	32.53	51.34	40.18	53.13
Cat-Seg [17]	OVS	88.68	93.34	53.34	70.86	28.64	38.73	35.75	49.03	25.73	40.54	40.24	51.65	42.30	60.65	39.14	55.85	44.23	57.58
OVRS	OVRSIS	88.85	93.64	52.65	69.59	31.53	59.82	36.44	50.17	24.13	34.83	40.82	54.24	43.50	63.31	37.23	56.34	44.39	60.24
GSNet [59]	OVRSIS	86.02	91.48	53.73	71.57	32.52	60.23	37.85	52.35	24.22	35.03	40.92	57.04	44.13	62.38	37.34	57.04	44.56	60.89
RSKT-Seg [54]	OVRSIS	**93.49**	**96.49**	54.32	71.72	33.23	57.41	**38.44**	**56.69**	25.72	40.89	42.10	56.39	42.69	63.29	39.69	58.60	46.21	62.69
UGDMoE	OVRSIS	93.35	96.42	**54.38**	**71.75**	**34.17**	**60.45**	38.21	56.22	**26.85**	**41.23**	**43.78**	**57.82**	**44.92**	**63.54**	**40.12**	**59.03**	**46.97**	**63.31**

With iSAID as the source dataset (Table 3), UGDMoE again achieves the best overall average, reaching 37.47 m-mIoU and 58.52 m-mACC. The advantage remains visible on several out-of-domain targets, including DLRSD, Potsdam, UAVid, UDD5, and VDD. This suggests that the proposed framework does not rely on a specific source-domain prior, but improves a more general capability: stabilizing dense pixel–text alignment under prompt variation and domain shift.

Table 3. Main results of UGDMoE on 8 OVRSIS datasets with iSAID as the training set. Best results are bold, second-best are underlined. m-mIoU and m-mACC denote average performance across all datasets.

Method	Type	DLRSD		iSAID		LoveDA		Potsdam		UAVid		UDD5		Vaihingen		VDD		Avg	
		mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	m-mIoU	m-mACC
SCAN	OVS	21.44	53.26	64.28	85.46	18.50	38.20	28.32	52.47	14.32	26.00	16.22	32.51	14.23	34.25	20.18	36.70	24.69	44.86
SAN	OVS	20.54	49.32	87.22	92.54	22.33	44.22	24.72	**56.54**	15.01	26.50	21.35	37.66	22.49	50.78	26.42	42.30	30.01	49.98
SED	OVS	23.80	50.36	94.32	96.84	23.24	45.13	11.85	23.87	15.21	26.80	22.42	39.15	12.61	25.73	28.50	44.20	28.99	44.01
Cat-Seg	OVS	28.80	59.56	94.77	96.96	25.11	48.25	23.90	49.49	16.10	26.90	24.32	42.77	21.74	51.25	30.16	47.20	33.11	52.80
OVRS	OVRSIS	32.25	60.35	**94.86**	**97.06**	27.98	49.01	26.39	50.15	16.24	27.00	31.88	55.32	**28.80**	**54.20**	31.01	55.30	36.18	56.05
GSNet	OVRSIS	31.50	63.05	93.11	95.98	27.21	49.53	28.50	52.00	15.98	27.10	32.24	56.20	25.10	52.80	32.07	55.21	35.71	56.48
RSKT-Seg	OVRSIS	31.57	61.66	93.96	96.63	**29.62**	**51.01**	28.57	51.69	17.36	35.75	34.01	57.38	25.55	52.66	33.40	56.50	36.76	57.91
UGDMoE	OVRSIS	**32.73**	**63.28**	94.37	96.72	29.06	50.88	**29.15**	52.17	**17.68**	**36.07**	**34.37**	**58.13**	28.27	53.76	**34.12**	**57.18**	**37.47**	**58.52**

Across both training settings, the improvements are most evident on datasets that stress different aspects of open-vocabulary dense prediction. On Potsdam and Vaihingen, where fine boundaries coexist with large homogeneous regions, UGDMoE improves both mIoU and mACC, consistent with the S/M/L branch design for scale-mixed decoding. On UAVid, which contains narrow structures and small objects, the gains indicate that risk-aware routing and structure refinement reduce fragmentation and over-smoothing. On UDD5 and VDD, where domain gaps and texture confusion are more pronounced, the improvements suggest that the alignment-risk cue helps suppress semantic drift and over-expansion in ambiguous regions.

4.4. Ablation Study

4.4.1. Ablation on UGDMoE Components

Table 4 reports the component-wise ablation of UGDMoE on UAVid and VDD under two transfer settings. Starting from a CLIP-only baseline based on direct pixel–text matching, the full model yields substantial gains in both mIoU and mACC across all settings. Under DLRSD training, for example, the mIoU increases from 18.24 to 26.85 on UAVid and from 28.76 to 40.12 on VDD. Similar gains are observed when the source dataset is switched to iSAID.

Table 4. Ablation study of core components of UGDMoE on UAVid and VDD datasets (DLRSD and iSAID training). Best results are shown in bold.

Method	UAVid(DLRSD)		VDD(DLRSD)		UAVid(iSAID)		VDD(iSAID)	
	mIoU	**mACC**	**mIoU**	**mACC**	**mIoU**	**mACC**	**mIoU**	**mACC**
Baseline (CLIP only)	18.24	32.15	28.76	48.32	10.53	24.87	22.41	42.65
w/o MoE	21.32	34.18	33.25	50.46	12.05	27.34	26.84	47.92
w/o Refinement	24.10	37.52	37.03	54.81	15.21	31.85	30.95	52.37
w/o PELC	24.57	36.94	37.86	55.23	14.83	30.12	31.42	53.05
w/o QGF	25.23	38.76	38.45	56.72	16.12	33.28	32.36	54.91
w/o PosInt	25.68	39.42	38.97	57.34	16.54	34.15	32.89	55.63
w/o Fusion	25.41	39.05	38.62	56.98	16.31	33.72	32.57	55.24
Full (UGDMoE)	**26.85**	**41.23**	**40.12**	**59.03**	**17.68**	**36.07**	**34.12**	**57.18**

Among all components, removing the MoE decoder causes the largest performance drop. This indicates that the expert-based decoder is the main contributor to performance gains in scale-mixed remote sensing scenes, where thin targets, medium-scale objects, and large homogeneous regions must be handled jointly. Removing QGF also degrades performance consistently, showing that injecting textual priors into dense visual features is more effective than relying only on post hoc similarity matching. Removing PELC causes another clear drop, which supports the role of prompt-dispersion calibration in reducing prompt-induced prediction instability. The refinement module mainly affects UAVid, where thin structures and boundary-sensitive targets are more common, which is consistent with its role in correcting local breaks and boundary artifacts.

The early feature processing pipeline is also important. Removing positional interpolation reduces the effectiveness of the backbone on high-resolution inputs, while removing multi-scale fusion weakens the representation used by subsequent routing and calibration. These results indicate that UGDMoE benefits from both the scale-aware backbone adaptation and the explicit alignment-risk-guided decoding pipeline.

4.4.2. Ablation of MoE Design

Table 5 further studies the design of the MoE module. All models are trained on DLRSD and evaluated on UAVid and VDD. We first vary the number of homogeneous experts without geometric role differentiation. Increasing the number of experts from 1 to 3 improves performance on both datasets, with UAVid mIoU rising from 21.32 to 24.02 and VDD mIoU from 33.25 to 37.12. Increasing the number further to 4 does not provide additional gains, which suggests that merely enlarging the expert set is insufficient once redundancy appears.

We then fix the number of experts to $E = 3$ and evaluate the contribution of the three S/M/L branches by removing one branch at a time. Using all three branches gives the best results on both datasets, reaching 26.85 mIoU/41.23 mACC on UAVid and 40.12 mIoU/59.03 mACC on VDD. All two-branch variants perform worse than the full configuration, indicating that the three branches provide complementary decoding

behaviors. Among them, removing the slender-structure branch causes the most evident degradation, which is consistent with the prevalence of narrow and elongated structures in remote sensing imagery. Removing the medium-scale branch or the large-context branch also degrades performance, showing that accurate OV-RSS requires joint modeling of fine detail, intermediate-scale objects, and broad contextual regions.

Table 5. Ablation study of the MoE module on the UAVid and VDD datasets (trained on DLRSD). S/M/L denote the three expert branches for slender, medium-scale, and large-context patterns. Best results are shown in bold.

Expert Number	S	M	L	UAVid(DLRSD)		VDD(DLRSD)	
				mIoU	mACC	mIoU	mACC
1	✗	✗	✗	21.32	34.18	33.25	50.46
2	✗	✗	✗	23.10	36.02	35.74	53.10
3	✗	✗	✗	24.02	37.11	37.12	54.66
4	✗	✗	✗	23.95	37.02	36.89	54.28
3	✓	✗	✓	25.68	39.42	38.97	57.34
3	✗	✓	✓	25.41	39.05	38.62	56.98
3	✓	✓	✗	25.23	38.76	38.45	56.72
3	✓	✓	✓	**26.85**	**41.23**	**40.12**	**59.03**

Overall, the results in Table 5 show that the effectiveness of the MoE decoder does not primarily come from the number of experts but from the geometric role differentiation of the S/M/L branches.

4.4.3. Ablation on PELC Risk Estimation

Table 6 studies the estimation strategy of the alignment-risk cue U_0 by changing only its source while keeping the rest of UGDMoE unchanged. All models are trained on DLRSD. Here, "w/o PELC" removes only the calibration head while preserving the remaining decoding and refinement architecture; it is therefore different from the "No U_0" setting in Table 7, where the alignment-risk cue is removed from both routing and refinement. Under "w/o PELC", the model obtains 24.57 mIoU on UAVid and 37.86 mIoU on VDD, serving as the reference without prompt-dispersion-based calibration. Constructing U_0 from non-equivalent prompts, such as {road, car, tree}, leads to a clear performance drop to 22.47 on UAVid and 34.10 on VDD. These results indicate that the overall effectiveness of the U_0-guided calibration strategy depends not only on uncertainty estimation itself but also on whether the resulting cue reflects semantic equivalence under different prompt formulations.

Table 6. Ablation of the PELC module for alignment-risk estimation. All variants use the same UGDMoE architecture; only the source of U_0 changes. Results are on models trained on DLRSD. "w/o PELC" corresponds to the setting in Table 4.

Variant	U_0 Source	UAVid mIoU	VDD mIoU
w/o PELC	Disabled	24.57	37.86
Non-equiv. prompts	{road, car, tree} prompt ensemble	22.47	34.10
Entropy-based U_0	Prediction entropy	24.30	36.80
Dropout-based U_0	MC-dropout variance	24.90	37.20
Full (PELC)	Synonym-consistent prompt ensemble	**26.85**	**40.12**

We further compare several annotation-free alternatives for estimating U_0. Prediction entropy slightly underperforms the no-PELC setting, while MC-dropout variance brings only marginal improvements. In contrast, the proposed synonym-consistent prompt

ensemble achieves the best results, reaching 26.85 mIoU on UAVid and 40.12 on VDD. These results indicate that the effectiveness of PELC depends not merely on introducing an uncertainty-related cue but on whether the adopted U_0 correctly captures semantic agreement under equivalent prompt formulations.

Table 7. Architectural ablation of the alignment-risk cue U_0. Models are trained on DLRSD and evaluated on UAVid and VDD.

Setting	UAVid(DLRSD)		VDD(DLRSD)	
	mIoU	**mACC**	**mIoU**	**mACC**
No U_0	25.70	39.60	38.60	56.90
w/o U_0 for MoE Routing	25.92	39.88	38.97	57.34
w/o U_0 for Refinement	26.14	40.15	39.28	57.62
Visual Confidence Surrogate	22.63	35.85	34.35	51.58
Inverted U_0 (1-U_0)	22.47	35.60	34.10	51.20
Full (U_0)	**26.85**	**41.23**	**40.12**	**59.03**

4.4.4. Ablation on the Alignment-Risk Cue U_0

Table 7 analyzes the architectural role of the alignment-risk cue U_0. In this table, "No U_0" removes the cue from both the MoE routing stage and the refinement stage; "w/o U_0 for MoE routing" removes it only from routing while retaining it in refinement; and "w/o U_0 for Refinement" removes it only from refinement while retaining it in routing. When U_0 is removed from both stages, performance drops to 25.70 mIoU on UAVid and 38.60 mIoU on VDD. Removing U_0 only from the routing stage gives 25.92 mIoU on UAVid, whereas removing it only from refinement gives 26.14 mIoU. These results indicate that both stages benefit from the risk cue, with routing showing slightly higher sensitivity in this setting.

We also compare the proposed U_0 with several alternative spatial cues. Replacing it with a visual confidence surrogate causes a clear degradation, reducing mIoU to 22.63 on UAVid and 34.35 on VDD. Reversing the cue to $1 - U_0$ leads to a similarly large drop. This shows that the gains do not come from simply introducing an additional spatial weighting term but from whether the cue can accurately highlight regions with unstable pixel–text alignment. The synonym-consistent U_0 used in UGDMoE provides a more effective signal for expert routing, logit calibration, and refinement.

4.4.5. Signal Robustness and Semantic Sensitivity of U_0

Table 8 further evaluates the robustness and functional characteristics of the alignment-risk cue U_0. Adding Gaussian noise to the estimated U_0 map consistently degrades performance. With $\sigma = 0.1$ and $\sigma = 0.5$, the mIoU on UAVid drops to 24.91 and 24.80, respectively, both lower than the variant without U_0. This suggests that the decoder is sensitive to the quality of the risk cue, and that a corrupted signal may interfere with expert allocation rather than provide useful guidance.

We further compare the proposed design with controlled variants to examine what information U_0 captures. The full model slightly outperforms the Oracle U_0 variant, which uses ground-truth boundaries. This indicates that the benefit of U_0 is not limited to highlighting structural discontinuities but also comes from modeling semantic instability under open-vocabulary prediction. A similar trend appears in the masking study. Removing the cue around structural-break regions causes the largest drop, while masking texture-confused regions produces a slightly smaller degradation. These results suggest that U_0 contributes to both semantic disambiguation and structural repair, with the latter playing a slightly more critical role in the final prediction.

Table 8. Ablation study on the alignment-risk cue U_0. All models are trained on DLRSD and evaluated on UAVid and VDD. Performance is measured by mIoU (%).

Variant	UAVid mIoU	VDD mIoU
No U_0	25.70	38.60
U_0 + Gaussian Noise ($\sigma = 0.1$)	24.91	37.02
U_0 + Gaussian Noise ($\sigma = 0.5$)	24.80	36.20
Oracle U_0 (Ground-Truth Boundaries)	26.50	39.70
Masked U_0 (Texture Confusion)	24.36	37.52
Masked U_0 (Structural Break)	24.30	36.80
Full (UGDMoE)	26.85	40.12

4.5. Visualizations

4.5.1. Comparative Qualitative Results

Figure 6 presents qualitative results on UAVid, where all models are trained on DLRSD and evaluated on representative urban scenes. Compared with existing open-vocabulary segmentation baselines, UGDMoE produces more consistent predictions in scale-mixed areas and alleviates several common failure modes, including fragmented thin structures, boundary bleeding across adjacent regions, and semantic confusion caused by repetitive textures. In particular, UGDMoE better preserves the continuity of slender targets while maintaining clearer boundaries and more coherent regional predictions. The highlighted difficult regions further illustrate the advantage of UGDMoE in challenging urban scenes, and these observations are consistent with the quantitative improvements.

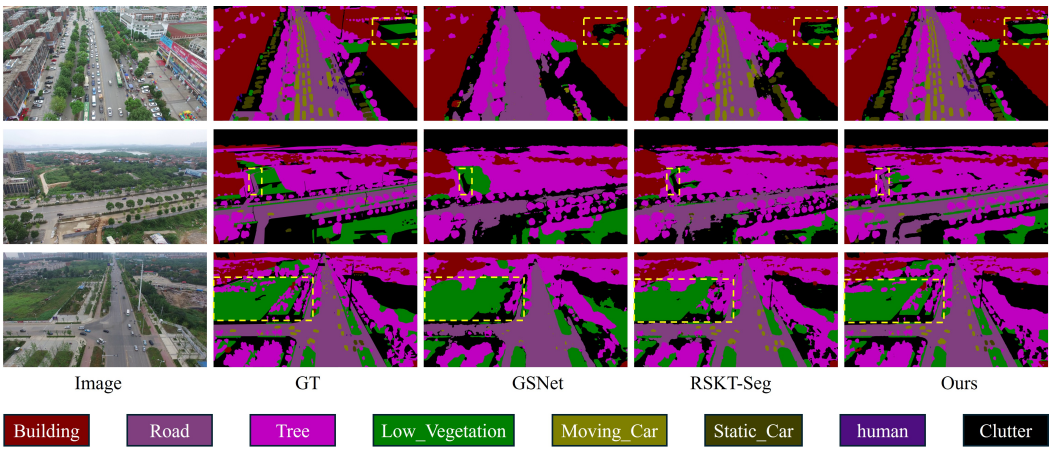


Figure 6. Qualitative comparisons on UAVid. From left to right, each row shows the input image, ground truth, predictions of representative baselines, and UGDMoE. Yellow dashed boxes highlight the most challenging regions, where UGDMoE better preserves thin-structure continuity, reduces boundary ambiguity, and yields more coherent predictions. The color legend at the bottom indicates the semantic categories used in visualization.

4.5.2. Qualitative Results on VDD

Figure 7 presents qualitative comparisons on VDD, highlighting representative scenes with complex semantic layouts and challenging inter-class boundaries. Compared with existing open-vocabulary segmentation baselines, UGDMoE produces more coherent regional predictions and cleaner transitions in difficult areas where water, vegetation, road, and roof regions are easily confused. In particular, the highlighted regions show that UGDMoE better alleviates semantic confusion and boundary bleeding across adjacent regions. These observations are consistent with the quantitative improvements and further support the effectiveness of the S/M/L expert branches together with alignment-risk-guided calibration in semantically ambiguous scenes.

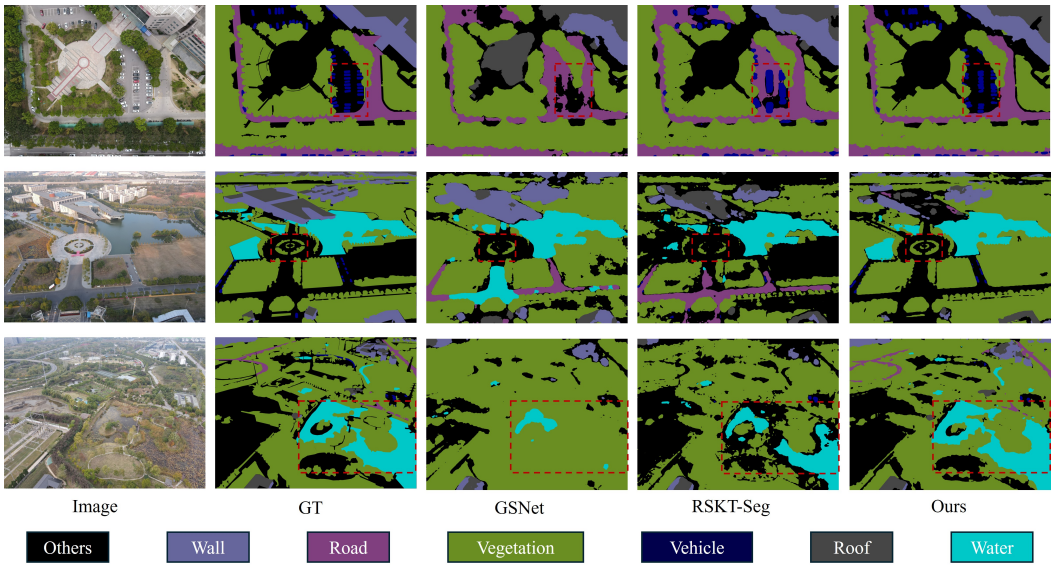


Figure 7. Qualitative comparisons on VDD. From left to right, each row shows the input image, ground truth, predictions of representative baseline methods, and UGDMoE. Red dashed boxes highlight the most representative difficult regions, where UGDMoE reduces semantic confusion, suppresses boundary bleeding across adjacent regions, and yields more coherent segmentation results. The color legend at the bottom indicates the semantic categories used in visualization.

4.5.3. Feature-Response Visualization

Figure 8 shows intermediate feature responses on UAVid. Compared with Scan and RSKT-Seg, our method exhibits slightly more structured responses along the main road trajectory and nearby geometric boundaries, while reducing part of the scattered activations in background regions. Although these responses are not final predictions, they provide supportive evidence that UGDMoE encourages a more organized intermediate representation, which is broadly consistent with the qualitative results in Figure 6.

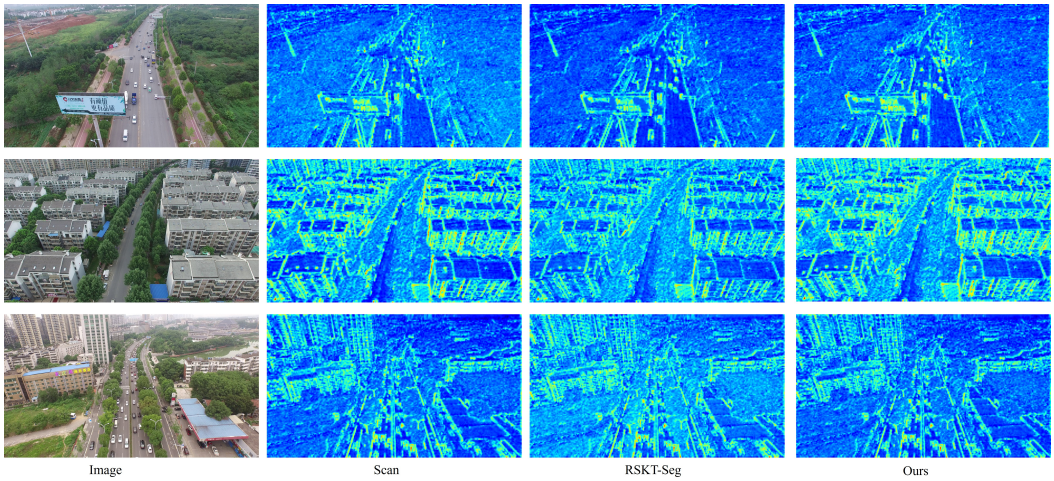


Figure 8. Intermediate feature-response visualization on UAVid. From left to right: input image, feature responses of Scan, RSKT-Seg, and Ours. Our method shows slightly more structured activations along road trajectories and nearby boundaries.

4.6. Expert Activation Maps and Geometric Specialization

To improve the interpretability of UGDMoE, we visualize the internal routing behavior using representative region-specific activation maps and the corresponding average expert preferences. Figure 9 shows three representative cases associated with slender, mid-scale, and large-scale regions, respectively. In each row, the left column provides the colorized

label visualization of the selected region, the middle column shows the corresponding expert activation map overlaid on the input image, and the right column summarizes the average expert preference computed over pixels within that region type.

The first row shows that the highest routing preference is assigned to Expert-S on slender regions, where the activation is concentrated along narrow and elongated structures. The second row indicates that mid-scale regions are more strongly associated with Expert-M, while the third row shows that broad spatially continuous regions are predominantly assigned to Expert-L. These results suggest that the three expert branches exhibit distinct scale-related tendencies rather than redundant behavior. Together with the U_0 ablation results, these visualizations support the interpretation that alignment-risk-guided routing allocates more suitable expert capacity to regions with different geometric characteristics.

4.7. Limitations Analysis

Despite the robust generalization demonstrated across diverse OVRSIS benchmarks, UGDMoE has several limitations. First, semantic reasoning is still anchored in the pretrained CLIP backbone, which may limit zero-shot performance on highly specialized geospatial categories or rare land-cover types that are weakly represented in the pretraining corpus. Second, the refinement stage relies on high-resolution visual features to restore structural continuity. Under severe image degradation, such as dense cloud cover or strong urban shadows, these low-level cues may become unreliable and weaken the effectiveness of the alignment-risk-guided gating process. Third, although the S/M/L branch design is effective in most scale-mixed scenes, objects with highly ambiguous or atypical geometric patterns may still induce routing instability and lead to fragmented predictions.

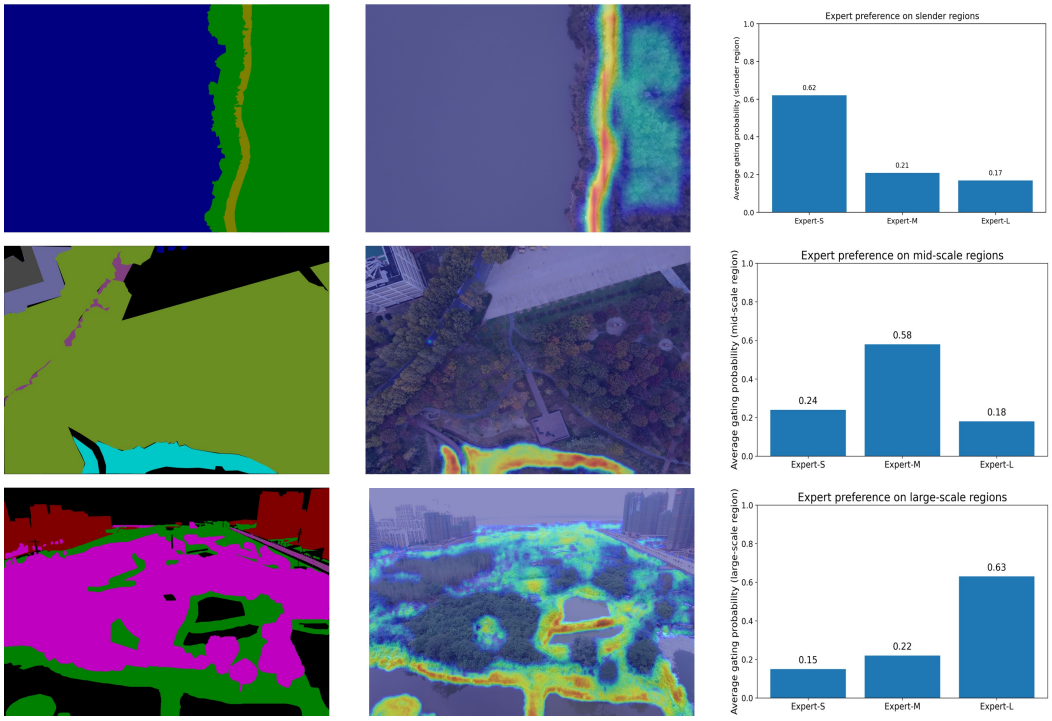


Figure 9. Representative expert activation maps and expert preferences across different geometric region types. From top to bottom: slender-scale, mid-scale, and large-scale regions. (**Left**): colorized label visualization of the selected region. (**Middle**): corresponding expert activation map overlaid on the input image (higher intensity indicates stronger routing preference). (**Right**): average expert preference aggregated over pixels within the corresponding region type, showing that Expert-S, Expert-M, and Expert-L are respectively more favored on slender, mid-scale, and large-scale regions.

5. Conclusions

In this paper, we proposed UGDMoE, a risk-aware mixture-of-experts framework for open-vocabulary remote sensing semantic segmentation. The method is designed for dense geospatial scenes in which extreme scale variation, thin or elongated structures, and repetitive textures make pixel–text alignment unstable. To address these difficulties, UGDMoE derives an explicit alignment-risk cue U_0 from synonym-consistent prompt dispersion and reuses it in three stages: risk-aware expert routing, prompt-ensemble logit calibration, and structure-aware refinement. The MoE decoder uses three expert branches with complementary geometric preferences to better handle scale-mixed regions, while the refinement stage improves the continuity and boundary quality of thin targets. Extensive experiments on eight OVRSIS benchmarks show that UGDMoE achieves state-of-the-art performance and maintains strong cross-source generalization. Future work will investigate automated prompt expansion with geospatial ontologies, multimodal and temporal extensions, and broader applications to open-world dense prediction tasks.

Author Contributions: Conceptualization, W.Q., G.J.; methodology, W.Q., G.J.; validation, W.Q., G.J.; analysis, W.Q.; resources, Q.Y., Z.G., J.Z.; writing—original draft preparation, W.Q., G.J.; writing—review and editing, W.Q., G.J.; visualization, Q.Y., Z.G., J.Z.; funding acquisition, G.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Zhejiang Province Science and Technology Plan Project (2026LDC01014(JT)); Hangzhou Natural Science Foundation (2024HZY0349); Postdoctoral Research Funding of Hangzhou International Innovation Institute of Beihang University (2025BKZ022).

Data Availability Statement: The data presented in this research are available upon request from the corresponding author.

Conflicts of Interest: Author Qiang Yuan was employed by the company Hangzhou Institute of Communications Planning Design & Research Co., Ltd., Author Zhushenyu Guo was employed by the company Hangzhou Low-Altitude Economy Industry Development Co., Ltd., Hangzhou Industrial Investment Group Co., Ltd., Author Jianfeng Zhang was employed by the company Wasu Media Network Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Yu, D.; Fang, C. Urban remote sensing with spatial big data: A review and renewed perspective of urban studies in recent decades. *Remote Sens.* **2023**, *15*, 1307. [CrossRef]
2. Neupane, B.; Horanont, T.; Aryal, J. Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis. *Remote Sens.* **2021**, *13*, 808. [CrossRef]
3. Dritsas, E.; Trigka, M. Remote sensing and geospatial analysis in the big data era: A survey. *Remote Sens.* **2025**, *17*, 550. [CrossRef]
4. Cao, Q.; Chen, Y.; Ma, C.; Yang, X. Open-vocabulary remote sensing image semantic segmentation. *arXiv* **2024**, arXiv:2409.07683.
5. Li, X.; Wen, C.; Hu, Y.; Yuan, Z.; Zhu, X.X. Vision-language models in remote sensing: Current progress and future trends. *IEEE Geosci. Remote Sens. Mag.* **2024**, *12*, 32–66. [CrossRef]
6. Lu, S.; Guo, J.; Zimmer-Dauphinee, J.R.; Nieusma, J.M.; Wang, X.; VanValkenburgh, P.; Wernke, S.A.; Huo, Y. Vision foundation models in remote sensing: A survey. *IEEE Geosci. Remote Sens. Mag.* **2025**, *13*, 190–215. [CrossRef]
7. Li, X.; Wen, C.; Hu, Y.; Zhou, N. RS-CLIP: Zero shot remote sensing scene classification via contrastive vision-language supervision. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *124*, 103497. [CrossRef]
8. Zhang, J.; Zhou, Z.; Mai, G.; Hu, M.; Guan, Z.; Li, S.; Mu, L. Text2Seg: Zero-shot remote sensing image semantic segmentation via text-guided visual foundation models. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*; ACM: New York, NY, USA, 2024; pp. 63–66.
9. Huang, B.; Li, J.; Luan, W.; Tan, J.; Li, C.; Huang, L. Expanding Open-Vocabulary Understanding for UAV Aerial Imagery: A Vision–Language Framework to Semantic Segmentation. *Drones* **2025**, *9*, 155. [CrossRef]

10. Shojaei, S.; Bohl, T.; Palaniappan, K.; Bunyak, F. Adaptive Structure-Aware Connectivity-Preserving Loss for Improved Road Segmentation in Remote Sensing Images. In *Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*; IEEE: New York, NY, USA, 2025; pp. 1120–1128.
11. Li, F.; Zhang, Z.; Wang, X.; Wang, X.; Xu, Y. Bridging Vision Foundation and Vision–Language Models for Open-Vocabulary Semantic Segmentation of UAV Imagery. *Remote Sens.* **2025**, *17*, 3704. [CrossRef]
12. Cao, Q.; Chen, Y.; Ma, C.; Yang, X. Open-Vocabulary High-Resolution Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *663*, 5620614. [CrossRef]
13. Zhang, K.; Li, P.; Wang, J. A review of deep learning-based remote sensing image caption: Methods, models, comparisons and future directions. *Remote Sens.* **2024**, *16*, 4113. [CrossRef]
14. Qiu, X.; Zhang, Z.; Luo, X.; Zhang, X.; Yang, Y.; Wu, Y.; Su, J. Semantic Uncertainty-Awared for Semantic Segmentation of Remote Sensing Images. *IET Image Process.* **2025**, *19*, e70045. [CrossRef]
15. Lin, Y.; Chen, M.; Zhang, K.; Li, H.; Li, M.; Yang, Z.; Lv, D.; Lin, B.; Liu, H.; Cai, D. Tagclip: A local-to-global framework to enhance open-vocabulary multi-label classification of clip without training. In *Proceedings of the AAAI Conference on Artificial Intelligence*; ACM: New York, NY, USA, 2024; Volume 38, pp. 3513–3521.
16. Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; Bai, X. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2023; pp. 2945–2954.
17. Cho, S.; Shin, H.; Hong, S.; Arnab, A.; Seo, P.H.; Kim, S. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 4113–4123.
18. Zhang, X.; Zhou, C.; Huang, J.; Zhang, L. TPOV-Seg: Textually Enhanced Prompt Tuning of Vision-Language Models for Open-Vocabulary Remote Sensing Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4420117. [CrossRef]
19. Wu, J.; Shi, J.; Zhao, Z.; Liu, Z.; Zhi, R. FreeMix: Open-Vocabulary Domain Generalization of Remote-Sensing Images for Semantic Segmentation. *Remote Sens.* **2025**, *17*, 1357. [CrossRef]
20. Chen, B.; Chen, K.; Yang, M.; Zou, Z.; Shi, Z. Heterogeneous mixture of experts for remote sensing image super-resolution. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 6006405. [CrossRef]
21. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv* **2021**, arXiv:2110.08733.
22. Waqas Zamir, S.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Shahbaz Khan, F.; Zhu, F.; Shao, L.; Xia, G.S.; Bai, X. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*; IEEE: New York, NY, USA, 2019; pp. 28–37.
23. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2020; pp. 4096–4105.
24. 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 Sens.* **2024**, *16*, 2930. [CrossRef]
25. Li, J.; Sun, B.; Li, S.; Kang, X. Semisupervised semantic segmentation of remote sensing images with consistency self-training. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5615811. [CrossRef]
26. Chang, Z.; Lu, Y.; Ran, X.; Gao, X.; Wang, X. Few-shot semantic segmentation: A review on recent approaches. *Neural Comput. Appl.* **2023**, *35*, 18251–18275. [CrossRef]
27. Huang, Z.; Xiang, T.Z.; Chen, H.X.; Dai, H. Scribble-based boundary-aware network for weakly supervised salient object detection in remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2022**, *191*, 290–301. [CrossRef]
28. Zhu, J.; Guo, Y.; Sun, G.; Yang, L.; Deng, M.; Chen, J. Unsupervised domain adaptation semantic segmentation of high-resolution remote sensing imagery with invariant domain-level prototype memory. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–18. [CrossRef]
29. Garioud, A.; Gonthier, N.; Landrieu, L.; De Wit, A.; Valette, M.; Poupée, M.; Giordano, S.; Wattrelos, B. FLAIR: A country-scale land cover semantic segmentation dataset from multi-source optical imagery. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 16456–16482.
30. Wang, D.; Zhang, J.; Du, B.; Xu, M.; Liu, L.; Tao, D.; Zhang, L. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 8815–8827.
31. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*; PmLR: Cambridge, MA, USA, 2021; pp. 8748–8763.
32. Xie, B.; Cao, J.; Xie, J.; Khan, F.S.; Pang, Y. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 3426–3436.
33. Li, B.; Zhang, D.; Zhao, Z.; Gao, J.; Li, X. FGAseg: Fine-Grained Pixel-Text Alignment for Open-Vocabulary Semantic Segmentation. *arXiv* **2025**, arXiv:2501.00877.

34. Jiao, S.; Zhu, H.; Huang, J.; Zhao, Y.; Wei, Y.; Shi, H. Collaborative vision-text representation optimizing for open-vocabulary segmentation. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 399–416.
35. Li, B.; Weinberger, K.Q.; Belongie, S.; Koltun, V.; Ranftl, R. Language-driven semantic segmentation. *arXiv* **2022**, arXiv:2201.03546. [CrossRef]
36. Dong, X.; Bao, J.; Zheng, Y.; Zhang, T.; Chen, D.; Yang, H.; Zeng, M.; Zhang, W.; Yuan, L.; Chen, D.; et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer VISION and Pattern Recognition*; IEEE: New York, NY, USA, 2023; pp. 10995–11005.
37. Lüddecke, T.; Ecker, A. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 7086–7096.
38. Huynh, D.; Kuen, J.; Lin, Z.; Gu, J.; Elhamifar, E. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 7020–7031.
39. Wang, Z.; Xia, X.; Chen, Z.; He, X.; Guo, Y.; Gong, M.; Liu, T. Open-vocabulary segmentation with unpaired mask-text supervision. *arXiv* **2024**, arXiv:2402.08960.
40. Liu, Y.; Bai, S.; Li, G.; Wang, Y.; Tang, Y. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 3491–3500.
41. Bai, L.; Du, S.; Zhang, X.; Wang, H.; Liu, B.; Ouyang, S. Domain adaptation for remote sensing image semantic segmentation: An integrated approach of contrastive learning and adversarial learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5628313. [CrossRef]
42. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive mixtures of local experts. *Neural Comput.* **1991**, *3*, 79–87. [CrossRef] [PubMed]
43. Liu, Y.; Wu, C.; Guan, M.; Wang, J. MixtureRS: A Mixture of Expert Network Based Remote Sensing Land Classification. *Remote Sens.* **2025**, *17*, 2494. [CrossRef]
44. Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; Chi, E.H. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; ACM: New York, NY, USA, 2018; pp. 1930–1939.
45. Zoph, B. Designing effective sparse expert models. In *Proceedings of the 2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*; IEEE: New York, NY, USA, 2022; p. 1044.
46. Xue, F.; Zheng, Z.; Fu, Y.; Ni, J.; Zheng, Z.; Zhou, W.; You, Y. Openmoe: An early effort on open mixture-of-experts language models. *arXiv* **2024**, arXiv:2402.01739.
47. Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; Houlsby, N. Scaling vision with sparse mixture of experts. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8583–8595.
48. Yuan, X.; Lin, Z.; Kuen, J.; Zhang, J.; Wang, Y.; Maire, M.; Kale, A.; Faieta, B. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2021; pp. 6995–7004.
49. Lin, B.; Tang, Z.; Ye, Y.; Cui, J.; Zhu, B.; Jin, P.; Huang, J.; Zhang, J.; Pang, Y.; Ning, M.; et al. Moe-llava: Mixture of experts for large vision-language models. *arXiv* **2024**, arXiv:2401.15947. [CrossRef]
50. Cao, J.; You, Y.; Liu, J. Mv-moe: A visual mixture-of-experts model for optical-sar image matching. In *Proceedings of the IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*; IEEE: New York, NY, USA, 2024; pp. 9676–9679.
51. He, S.; Cheng, Q.; Huai, Y.; Zhu, Z.; Ding, J. Mixture-of-experts for semantic segmentation of remoting sensing image. In *Proceedings of the International Conference on Image Processing and Artificial Intelligence (ICIPAI 2024)*; SPIE: Bellingham, WA, USA, 2024; Volume 13213, pp. 478–483.
52. Lin, H.; Hong, D.; Ge, S.; Luo, C.; Jiang, K.; Jin, H.; Wen, C. Rs-moe: A vision-language model with mixture of experts for remote sensing image captioning and visual question answering. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5614918. [CrossRef]
53. Liu, X.; Lian, Z. Rsunivlm: A unified vision language model for remote sensing via granularity-oriented mixture of experts. *arXiv* **2024**, arXiv:2412.05679.
54. Li, B.; Dong, H.; Zhang, D.; Zhao, Z.; Gao, J.; Li, X. Exploring Efficient Open-Vocabulary Segmentation in the Remote Sensing. *arXiv* **2025**, arXiv:2509.12040. [CrossRef]
55. Lyu, Y.; Vosselman, G.; Xia, G.S.; Yilmaz, A.; Yang, M.Y. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *165*, 108–119. [CrossRef]
56. Cai, W.; Jin, K.; Hou, J.; Guo, C.; Wu, L.; Yang, W. Vdd: Varied drone dataset for semantic segmentation. *J. Vis. Commun. Image Represent.* **2025**, *109*, 104429. [CrossRef]
57. Chaudhuri, B.; Demir, B.; Chaudhuri, S.; Bruzzone, L. Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 1144–1158. [CrossRef]

58. Chen, Y.; Wang, Y.; Lu, P.; Chen, Y.; Wang, G. Large-scale structure from motion with semantic constraints of aerial images. In *Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 347–359.
59. Ye, C.; Zhuge, Y.; Zhang, P. Towards open-vocabulary remote sensing image semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*; ACM: New York, NY, USA, 2025; Volume 39, pp. 9436–9444.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.