

Article

Multimodal Uncertainty-Aware Gating Fusion and Iterative Feedback Refinement for HSI-LiDAR Open-Set Classification

Davaajargal Myagmarsuren , Haibin Wu *  and Aili Wang 

Heilongjiang Province Key Laboratory of Laser Spectroscopy Technology and Application, Harbin University of Science and Technology, Harbin 150080, China; 2212200001@stu.hrbust.edu.cn (D.M.); aili925@hrbust.edu.cn (A.W.)

* Correspondence: woo@hrbust.edu.cn

Highlights

What are the main findings?

- An Uncertainty-Aware Gating Fusion mechanism dynamically weights HSI-LiDAR contributions according to modality reliability, while a compact two-signal EVT (Weibull)-based post hoc calibrator combines gating uncertainty and refinement variance to achieve reliable known/unknown discrimination under a strict open-set protocol.
- An Iterative Feedback Refinement module captures convergence dynamics across multiple refinement steps: known samples converge stably while unknown samples exhibit higher feature-change variance, providing an interpretable and complementary open-set signal.

What are the implications of the main findings?

- The proposed framework accurately detects and rejects previously unseen land-cover classes without retraining, reducing false acceptances of unknown samples by up to 32 percentage points over single-modality baselines and enabling safer automated land-cover mapping in dynamically changing environments.

Abstract

Open-set classification for remote sensing requires models that simultaneously achieve high accuracy on known land-cover types and reliably detect novel classes absent from the training distribution—a capability essential for real-world deployment where new classes routinely emerge. Existing multimodal fusion approaches for hyperspectral imagery (HSI) and LiDAR are primarily designed for closed-set scenarios and lack robust uncertainty modeling for unknown detection. We propose a post hoc calibrated multimodal open-set framework with three tightly integrated components. First, an Uncertainty-Aware Gating Fusion (UAGF) module dynamically weights HSI and LiDAR features per sample based on modality reliability and produces a gating uncertainty signal reflecting fusion confidence. Second, an Iterative Feedback Refinement (IFR) module progressively refines fused representations over multiple iterations and captures convergence dynamics, where stable convergence indicates known samples while high feature-change variance identifies potential unknowns. Third, a compact two-signal open-set detector combines gating uncertainty and refinement variance through an EVT (Weibull)-based post hoc calibration mechanism fitted exclusively on known validation samples. The framework follows a strict zero-unknown-supervision protocol: the multimodal backbone is trained using only known-class samples, and the open-set decision threshold is derived solely from the known validation score distribution. This design decouples representation of



Academic Editor: Salah Bourennane

Received: 9 May 2026

Revised: 8 June 2026

Accepted: 10 June 2026

Published: 12 June 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

learning from open-set decision learning, improving robustness and avoiding the objective conflicts that arise in joint training. Comprehensive experiments on three benchmark datasets—Houston2013, Muufl, and Augsburg—demonstrate that the proposed method achieves 92.79%, 84.47%, and 80.99% overall accuracy and 76.48%, 63.91%, and 56.81% unknown accuracy, outperforming the closest multimodal competitor HyLiOSR by up to 32.4 pp in unknown accuracy while maintaining competitive closed-set performance.

Keywords: open-set recognition; multimodal fusion; uncertainty estimation; hyperspectral imaging; LiDAR data; contrastive learning; post hoc calibration; out-of-distribution detection

1. Introduction

Deep learning has fundamentally transformed land-cover classification in remote sensing. This has led to more accurate mapping, which is useful in precision agriculture, urban planning, and disaster response [1–4]. Recent developments, particularly multi-sensor fusion architectures, especially those using transformers [5–7], have significantly improved accuracy by combining the detailed spectral information from hyperspectral imagery (HSI) with the three-dimensional structural data provided by light detection and ranging (LiDAR).

Building on these foundations, subsequent architectures have explored increasingly sophisticated fusion strategies—from cascaded transformers and cross-modal attention to uncertainty-aware dynamic weighting—each demonstrating that explicitly modeling spectral–geometric interactions and modality reliability yield more robust representations [8–12].

Despite these advances, virtually all existing HSI–LiDAR fusion methods operate under the closed-set assumption: every sample encountered at inference is assumed to belong to one of the classes seen during training. In practice, this assumption rarely holds. Real-world remote sensing deployments encounter novel land-cover types, seasonal transitions, and anomalous observations that were entirely absent from the training distribution. Forcing such samples into known categories produces confidently incorrect predictions—a failure mode that has been empirically documented in closed-set hyperspectral classifiers, where out-of-distribution inputs routinely receive high SoftMax confidence scores despite belonging to entirely unseen classes [13,14]. This behavior represents a critical failure mode for safety-critical applications such as environmental monitoring and urban change detection, where a misclassified novel land-cover type can propagate systematic errors across downstream analyses.

Open-set recognition (OSR) was formalized by Scheirer et al. [15] as the task of simultaneously achieving high closed-set accuracy and reliably rejecting samples from unknown classes. OpenMax [16] extended the SoftMax output with Weibull-calibrated activation vectors to model the probability of belonging to a known class, while open-set neural networks [17] and generative approaches such as G-OpenMax [18] introduced explicit unknown modeling. However, these methods were designed for natural-image benchmarks with single-modality input. Their direct application to hyperspectral images is problematic: HSI feature spaces are high-dimensional and semantically dense, causing OpenMax-style methods to underperform relative to even a simple maximum-SoftMax baseline, as demonstrated by Bisgin et al. [13]. Critically, none of these methods exploit the complementary multimodal structure of HSI–LiDAR pairs, nor do they produce the calibrated probabilistic outputs required for reliable threshold-based open-set decisions.

The OSR and out-of-distribution (OOD) detection literature has recently converged on several key insights. Energy-based scoring [19] provides a theoretically grounded OOD signal without retraining. Post hoc calibration [20,21] decouples open-set decision-threshold selection from representation learning, improving stability when training and deployment distributions diverge. Training-agnostic post hoc detectors [22] offer particular practical value because they can be applied to any pre-trained backbone without architectural modification. Meanwhile, within the hyperspectral remote sensing community, contrastive representation learning [23,24], prototype-based scoring [25], and staged progressive training have produced increasingly competitive OSR results, though these methods target single-modality HSI and do not address the multimodal fusion problem.

To our knowledge, HyLiOSR [26] is the only prior work that jointly addresses open-set recognition and HSI–LiDAR fusion. It employs a staged autoencoder with per-class Gaussian prototypes and generates virtual unknown samples for classifier training. While effective, HyLiOSR requires careful hyperparameter tuning for prototype fitting and virtual sample generation, and its open-set decision threshold is fixed rather than learned. Our work addresses these limitations through a fundamentally different design philosophy.

We propose a post hoc calibrated open-set multimodal fusion framework that introduces three tightly integrated components. First, an Uncertainty-Aware Gating Fusion module dynamically weights HSI and LiDAR contributions per sample by estimating modality-specific reliability through gated cross-modal attention, producing a gating uncertainty signal. Second, an Iterative Feedback Refinement module progressively refines the fused representation over multiple iterations, where the convergence behavior of feature updates is explicitly modeled. Stable convergence corresponds to semantically coherent known samples, while unstable refinement dynamics indicate potential unknown samples.

Unlike prior approaches that rely on multiple heterogeneous signals, our final design adopts a compact two-signal formulation. Specifically, gating uncertainty and refinement dynamics are used as complementary uncertainty cues. These signals are further processed by a post hoc EVT (Weibull)-based calibrator [27,28], which estimates the likelihood of a sample belonging to known classes. The decision threshold is determined from the distribution of known validation samples, ensuring strict open-set compliance without requiring unknown supervision.

Rather than training an end-to-end open-set detector, we deliberately decouple representation learning from open-set decision learning. The multimodal backbone is trained with a closed-set objective combining cross-entropy, supervised contrastive loss [29], a quantile-stratified boundary separation loss, and a VICReg-inspired diversity regularizer [30] applied to the uncertainty signals. A post hoc EVT (Weibull)-based calibrator is fitted on validation-set scores to estimate known-vs-unknown likelihood, with the open-set decision threshold selected to maximize balanced accuracy between known and unknown samples. This staged design follows the training-agnostic post hoc paradigm [20] and avoids the objective conflicts that arise when classification and open-set decision losses are optimized simultaneously [14].

- The main contributions of this work are as follows:
- Uncertainty-Aware Gating Fusion with modality-discrepancy uncertainty. We propose a gating mechanism that adaptively fuses HSI and LiDAR features based on modality reliability. The discrepancy between modality gates provides a principled uncertainty signal reflecting fusion confidence.
- Iterative feedback refinement is a source of uncertainty. We introduce an iterative refinement process that captures feature convergence behavior across multiple steps. The resulting refinement dynamics provide an interpretable indicator for distinguishing known and unknown samples.

- Compact two-signal open-set detection with EVT calibration. We design a minimal yet effective open-set detector that leverages gating uncertainty and refinement dynamics, followed by an EVT (Weibull)-based post hoc calibration with fixed aggregation weights, ensuring stability and fairness.

The remainder of this paper is organized as follows: Section 2 reviews the work related to open-set recognition and multimodal fusion for remote sensing; Section 3 describes the proposed architecture, the training goals, and the implementation details; Section 4 describes the experimental setup and evaluation protocols, and presents comprehensive results, ablation studies, and comparative analyses; Section 5 is the discussion; and Section 6 presents key conclusions.

2. Related Works

2.1. Open-Set Recognition

Open-set recognition was formally introduced by Scheirer et al. [15], who defined the open-set risk and proposed the Compact Abating Probability (CAP) model to measure open-set decision performance. OpenMax [18] subsequently extended the SoftMax layer with per-class Weibull distributions fitted to deep activation vectors, allowing samples with low likelihood under all known class distributions to be flagged as unknown. While conceptually appealing, OpenMax was designed for natural-image benchmarks and degrades substantially in high-dimensional, semantically dense feature spaces such as those arising from hyperspectral imaging [13].

Subsequent methodologies have investigated more complex decision boundaries within the context of open-set recognition (OSR). Conditional Gaussian models [31] and hyper-spherical embeddings [32] exploit geometric properties inherent in the feature space to improve the separation between known and unknown classes. Activation shaping [33] reduces the response of pre-trained networks to out-of-distribution (OOD) inputs by constraining deep activations during the inference phase. Prototype-based approaches [25] combine reconstruction error with prototype distance, yielding an interpretable two-component open-set decision score. Adversarial Reciprocal Point Learning [34] determines per-class anchor points that are maximally separated from known class clusters, leading to strong performance on ImageNet-scale benchmarks. Generative methods, such as G-OpenMax [18], employ GANs to generate unknown-class samples for training augmentation; however, this approach entails significant computational expenses and is susceptible to GAN training instability.

Energy-based out-of-distribution (OOD) detection, as described in [19], introduced the free energy score. This score, which has a solid theoretical basis, is an alternative to using the maximum SoftMax probability. This method showed better separation between in-distribution and out-of-distribution samples. Furthermore, negative reciprocal score training [34] employs contrastive objectives designed explicitly for the task of unknown open-set decision. Recent studies in remote sensing OOD detection have investigated training-free outlier exposure methods to enhance the detection of unknown samples without retraining, highlighting the crucial role of robust post hoc uncertainty estimation for practical uses [35].

The Trusted Multi-View Classification (TMC) framework [36] provides Dempster–Shafer evidence fusion across views, yielding uncertainty estimates that adapt to per-view reliability—a conceptual predecessor to our gating uncertainty signal, though TMC targets discrete view fusion rather than continuous feature-level weighting.

Within the hyperspectral remote sensing community, supervised contrastive learning has emerged as a powerful representation objective for OSR. OSC-SCL [23] trains a prototype–distance detector for unknown detection using contrastive features, while

LDCLNet [24] combines dual-level contrastive learning with language-guided supervision to improve feature discriminability. GDAN [37] employs a GAN backbone with Bayesian uncertainty scoring, and reconstruction–prototype fusion [25] uses reconstruction error to complement prototype distance. The open-set domain adaptation work of Bi et al. [38] applies weighted GANs with dynamic thresholding to cross-domain HSI scenarios. Spectral decomposition via fractional Fourier transforms [39] improves feature discriminability for open-set HSI classification by routing samples through frequency-specialized branches. Recent advances in open-set hyperspectral classification exploit reconstruction and contrastive learning to improve known–unknown separation under partially unlabeled settings [40]. Additionally, meta-learning and prototype-guided representation learning techniques have been introduced, focusing on improving feature compactness for known classes while enhancing separability from unknown samples [41]. However, all these methods operate on single-modality HSI data and do not address the challenge of fusing uncertain signals from multiple sensor types. Our work extends the contrastive representation paradigm of OSC-SCL and LDCLNet to the multimodal HSI–LiDAR setting while adding the iterative feedback and uncertainty gating mechanisms.

Post hoc calibration methods have gained prominence as a training-agnostic complement to end-to-end OSR. Temperature scaling [27] adjusts SoftMax confidence without retraining; ODIN [42] combines temperature scaling with input perturbation to widen the gap between in- and out-of-distribution scores. A recent survey of OOD detection [22] emphasizes that post hoc detectors are the most practically deployable approach because they require no architectural changes. Our calibration stage complements uncertainty signals through an EVT-based calibration mechanism fitted on validation data, a strategy that generalizes the single-score post hoc paradigm to complementary multi-signal fusion. The open-set recognition methods reviewed above share a critical limitation for remote sensing applications: they are designed exclusively for single-modality inputs and cannot exploit the complementary information structure of HSI–LiDAR pairs. At the same time, a distinct body of work has advanced multimodal fusion architectures for HSI–LiDAR classification, achieving strong closed-set performance but without addressing the open-set challenge. The following section reviews this fusion literature, identifying the specific capabilities—and the specific gap—that motivates the joint design pursued in this work.

2.2. Multimodal Fusion for Remote Sensing

Early HSI–LiDAR fusion methods followed three paradigms: early fusion raw input, late fusion of per-modality decisions, and intermediate feature-level fusion [43]. Deep learning introduced substantially more expressive architectures. Hierarchical random walk bifurcation CNNs [44] and graph neural networks that jointly model spatial, spectral, and height information [45] exploit the complementary spatial, spectral, and geometric information available from the two sensors. Spectral–spatial tokenization transformers [7] apply self-attention to joint spectral–structural tokens, while cascaded transformer architectures [8] progressively integrate modality-specific representations.

Attention-based fusion has become the dominant paradigm for modeling inter-modal dependencies. Cross-modal self-attention [46] explicitly captures relationships between hyperspectral spectral signatures and LiDAR structural features, allowing the network to dynamically weight features across modalities. Pan et al. [10] further showed that transformer-based enhancement of cross-modal attention, combined with learned spectral–structural token alignment, produces more discriminative joint representations than conventional convolutional fusion under varying spatial resolution conditions. Dual-view hierarchical networks [47] maintain shallow and deep representations from both sensors and fuse them through learned interaction modules. More recently, Liu et al. [9]

extended the dual-view paradigm by combining hybrid feature extraction with a trusted decision fusion mechanism, demonstrating that explicit spectral–geometric interaction modeling yields consistent gains over single-view baselines across multiple benchmark datasets. Deep three-stream convolutional neural networks [48] and information pooling networks [49,50] offer different but related architectural perspectives.

Uncertainty-aware fusion represents a more recent advancement in the field. The ADFM framework [11] is specifically engineered to evaluate the trustworthiness of diverse data sources during integration, thereby mitigating the detrimental impacts of sensor data characterized by either noise or absence. Furthermore, adaptive multi-stage frameworks [12] can learn the reliability of individual components, with these reliability weights undergoing modification throughout the model’s training process. Taken together, methods [8–12,46,47] collectively establish that explicitly modeling spectral–geometric interactions and per-modality reliability is essential for robust HSI–LiDAR fusion—a principle that directly motivates the uncertainty-aware gating mechanism proposed in this work.

Orthogonal Total Variable Component Analysis (OTVCA) [51] amalgamates spatial and elevation data through the systematic extraction of features. Multi-level interactive fusion with competitive learning (MLIF-AL) [52] generates cross-modal interaction signals across feature levels. The interleaving perception CNN (IP-CNN) [53] employs a bidirectional autoencoder to jointly reconstruct both modalities, providing unsupervised alignment without label supervision. Our Uncertainty-Aware Gating Fusion module extends the uncertainty-driven weighting idea of ADFM and related methods by operating at the feature level, explicitly computing per-sample modality reliability, and generating a gate-discrepancy signal that feeds directly into open-set scoring.

Iterative refinement has suggested effectiveness in multiple remote sensing tasks. IDERS [54] uses iterative feedback to progressively reduce haze artifacts in remote sensing images. ISIFT [55] improves hyperspectral image registration through a feedback loop that iteratively updates correction parameters using stable feature correspondences. These methods establish iterative processing as a powerful inductive bias for remote sensing, but they do not extend to multimodal fusion or leverage refinement dynamics as an open-set signal.

Very recent studies on multimodal open-set HSI–LiDAR have started to explore geometry-aware representation learning and cross-modal consistency for unknown detection [56]. The closest related work in the HSI–LiDAR open-set setting is HyLiOSR [26], which introduces a hierarchical two-stage framework combining a per-class Gaussian autoencoder with a virtual-unknown augmented discriminative classifier. HyLiOSR demonstrates the viability of multimodal OSR but relies on sensitive hyperparameters for Gaussian prototype fitting and virtual sample generation and employs a fixed latent-likelihood open-set decision threshold. In contrast, the proposed framework avoids these limitations through architecturally learned uncertainty signals, a data-adaptive post hoc calibration stage, and a decoupled training objective—design choices described in detail in the next section.

Taken together, the two research threads reviewed in this section reveal a structural gap: open-set recognition methods provide principled unknown detection but ignore multimodal structure, while HSI–LiDAR fusion methods achieve high closed-set accuracy but assume a closed world. Bridging this gap requires an architecture that simultaneously learns reliable cross-modal representations and generates calibrated uncertainty signals suitable for open-set scoring—precisely the objective of the proposed framework.

3. Methodology

3.1. Problem Formulation and Architecture Overview

Let $D_{\text{train}} = \{(x_i^{\text{HSI}}, x_i^{\text{LiDAR}}, y_i)\}_{i=1}^N$ denote the training set, where $x_i^{\text{HSI}} \in \mathbb{R}^{P \times P \times c_H}$ represents a hyperspectral patch with spatial size $P \times P$ and c_H spectral bands, $x_i^{\text{LiDAR}} \in \mathbb{R}^{P \times P \times c_L}$ is the corresponding LiDAR patch with c_L channels, and $y_i \in y_{\text{known}} = \{1, \dots, K\}$ denotes the label from K known classes. During deployment, the model encounters samples from unknown classes, i.e., $y \notin y_{\text{known}}$.

The objective of multimodal open-set classification is to learn a decision function $f: (x^{\text{HSI}}, x^{\text{LiDAR}}) \rightarrow y_{\text{known}} \cup \{\text{unknown}\}$ that achieves high known-class classification accuracy while reliably rejecting samples that do not belong to the known training categories. The proposed framework consists of four sequential components, as illustrated in Figure 1. Sections 3.2–3.5 describe each component in detail; Figure 2 provides an expanded view of the Iterative Feedback Refinement module.

- (1) Modality-specific encoders extract modality-aware embeddings $h_{\text{HSI}}, h_{\text{LiDAR}} \in \mathbb{R}^d$, where d is the shared embedding dimension.
- (2) Uncertainty-aware gating fusion adaptively combines the two-modality embeddings according to their estimated reliability and produces a fused representation h_{fused} together with a gating uncertainty signal u_{gate} .
- (3) Iterative feedback refinement updates the fused feature over T iterations to obtain a refined representation h_{refined} and a feedback uncertainty signal u_{feedback} .
- (4) Post hoc open-set scoring and calibration use the refined feature and uncertainty signals to compute known/unknown evidence. The backbone is trained only for representation learning and closed-set discrimination, whereas the final open-set decision is obtained through a post hoc calibrator fitted on validation scores.

Figure 1 shows the general framework.

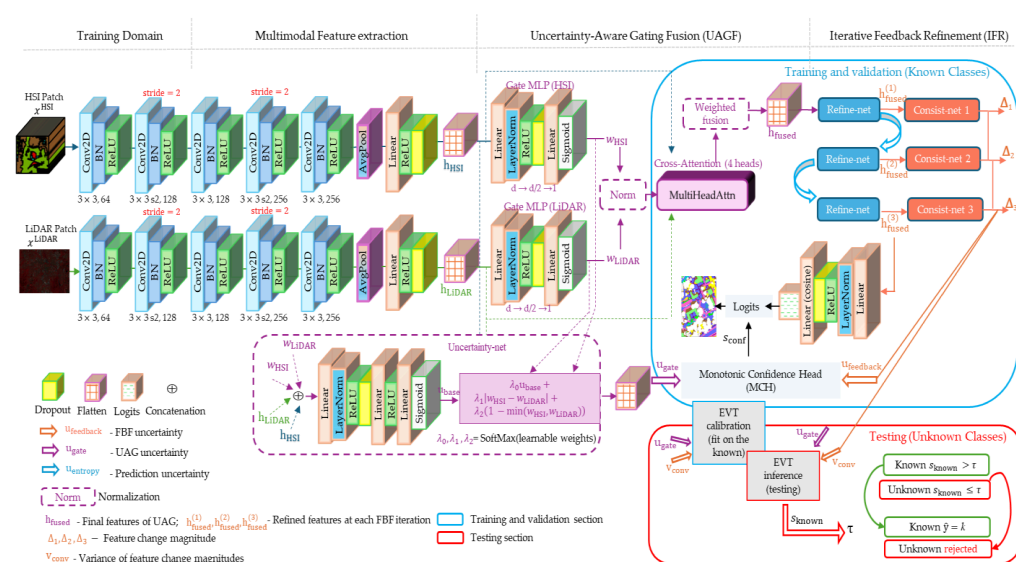


Figure 1. Architecture of the uncertainty-aware multimodal fusion framework for open-set remote sensing classification.

During training, the Uncertainty-Aware Gating Fusion module encodes and fuses HSI and LiDAR patches separately, yielding the gating uncertainty u_{gate} . The Iterative Feedback Refinement block refines the fused feature over numerous iterations, resulting in feedback-related uncertainty descriptors u_{feedback} and refinement-change statistics. A cosine classifier generates closed-set logits from which the normalized entropy uncertainty is calculated.

The monotonic confidence head (MCH) combines u_{gate} and u_{feedback} to produce a training-time confidence score, s_{conf} . For open-set recognition, the final post hoc EVT calibrator uses a compact two-signal form $[u_{\text{gate}}, \log(v_{\text{conv}} + \varepsilon)]$, where v_{conv} represents the variance of refinement changes. The calibrated known-class score s_{known} is compared to a threshold τ to differentiate between known and unknown samples.

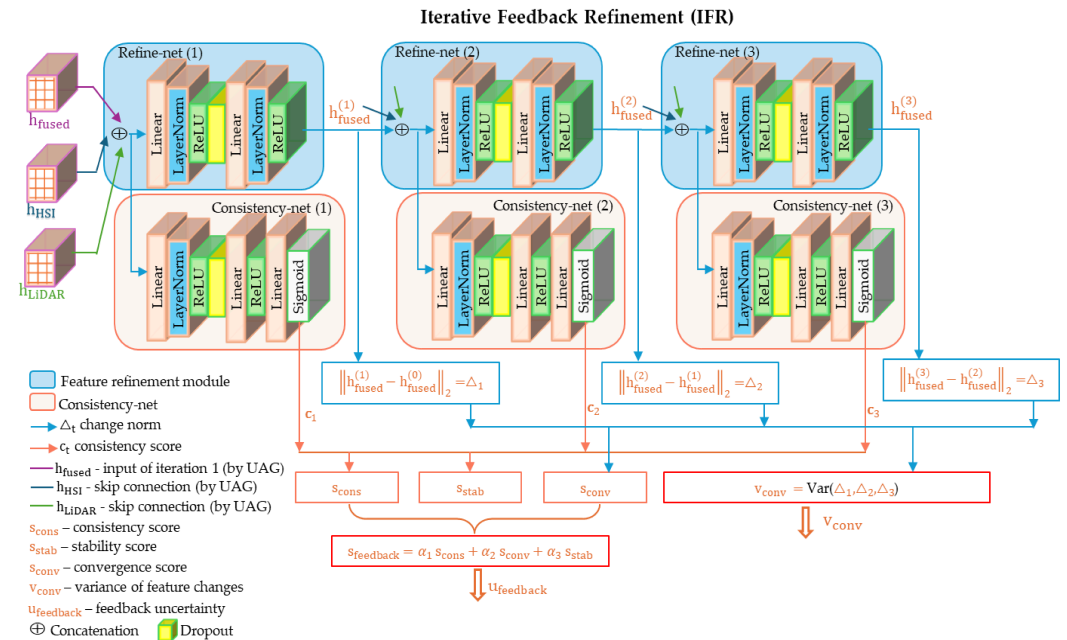


Figure 2. Illustration of the Iterative Feedback Refinement mechanism in the proposed multimodal framework.

3.2. Modality-Specific Encoders

Each modality is processed by a lightweight convolutional encoder to extract spatial-spectral representations. For an input patch x , two convolutional stages are used:

$$\begin{aligned} z_1 &= \text{MaxPool}(\text{ReLU}(\text{BN}(\text{Conv}_{3 \times 3}(x)))), \\ z_2 &= \text{MaxPool}(\text{ReLU}(\text{BN}(\text{Conv}_{3 \times 3}(z_1)))). \end{aligned} \quad (1)$$

Global average pooling then produces a spatially invariant descriptor

$$z_{\text{pool}} = \text{GlobalAvgPool}(z_2). \quad (2)$$

The pooled descriptor is projected into a shared embedding space

$$h_m = \text{ReLU}(\text{BN}(W_m z_{\text{pool}})), \quad (3)$$

where $m \in \{\text{HSI}, \text{LiDAR}\}$ and W_m is a learnable projection matrix. The resulting embeddings $h_{\text{HSI}}, h_{\text{LiDAR}} \in \mathbb{R}^d$ provide modality-specific semantic representations for the subsequent fusion stage.

3.3. Uncertainty-Aware Gating Fusion

The two modality embeddings are not equally reliable for every sample. The gating module estimates a scalar confidence score for each modality:

$$w_m = \sigma(\text{MLP}_m(h_m)), m \in \{\text{HSI}, \text{LiDAR}\} \quad (4)$$

where $\sigma(\cdot)$ denotes the sigmoid function and MLP_m is a two-layer perceptron. The weights are normalized as $\hat{w}_m = w_m / (w_{\text{HSI}} + w_{\text{LiDAR}} + \varepsilon)$ such that $\hat{w}_{\text{HSI}} + \hat{w}_{\text{LiDAR}} = 1$. Larger

values of w_m indicate higher confidence in modality m . To account for cross-modal interactions, the modality embeddings are refined using a lightweight attention mechanism

$$[h'_{\text{HSI}}, h'_{\text{LiDAR}}] = \text{Attn}(h_{\text{HSI}}, h_{\text{LiDAR}}). \quad (5)$$

The fused representation is then computed as

$$h_{\text{fused}} = \hat{w}_{\text{HSI}} h'_{\text{HSI}} + \hat{w}_{\text{LiDAR}} h'_{\text{LiDAR}}. \quad (6)$$

Instead of relying on normalized dominance measures, the proposed method directly exploits raw gating statistics to better capture modality disagreement. Specifically, the gating uncertainty is defined as

$$u_{\text{gate}} = \lambda_0 u_{\text{base}} + \lambda_1 |w_{\text{HSI}} - w_{\text{LiDAR}}| + \lambda_2 (1 - \min(w_{\text{HSI}}, w_{\text{LiDAR}})), \quad (7)$$

where

$$u_{\text{base}} = \sigma(\text{MLP}_{\text{gate}}([h_{\text{HSI}}; h_{\text{LiDAR}}; \hat{w}_{\text{HSI}}; \hat{w}_{\text{LiDAR}}])) \quad (8)$$

is a learned base uncertainty. $|w_{\text{HSI}} - w_{\text{LiDAR}}|$ captures modality disagreement, while $1 - \min(w_{\text{HSI}}, w_{\text{LiDAR}})$ penalizes cases where one modality becomes unreliable. Importantly, these two terms are computed in the raw gating space rather than the normalized space, which preserves discriminative variability between known and unknown samples. The coefficients $[\lambda_0, \lambda_1, \lambda_2]$ are jointly learned with the backbone via SoftMax normalization and do not require manual tuning; their values are implicitly encoded in the trained model weights and adapt to dataset-specific modality characteristics. This formulation provides a stable and interpretable measure of multimodal uncertainty compared to direct weight differences and serves as a key signal for subsequent open-set scoring.

3.4. Iterative Feedback Refinement

Our Iterative Feedback Refinement module is, to the best of our knowledge, the first work to use the convergence profile of iterative multimodal refinement—specifically, the trajectory of feature-change norms and cross-modal consistency scores across iterations—as a quantitative uncertainty indicator for open-set detection.

Starting from the initial fused representation $h^{(0)} = h_{\text{fused}}$, the model performs T iterative refinement steps:

$$h^{(t+1)} = \text{MLP}_{\text{refine}}\left([h^{(t)}; h_{\text{HSI}}; h_{\text{LiDAR}}]\right), \quad t = 0, \dots, T - 1. \quad (9)$$

Each refinement step fully updates the fused feature without residual connections. This design enforces progressive integration of multimodal information. To quantify refinement dynamics, we derive three complementary signals from the iterative process:

(1) Feature convergence:

$$\Delta^{(t)} = \|h^{(t+1)} - h^{(t)}\|_2, \quad (10)$$

and mean change magnitude:

$$\bar{\Delta} = \frac{1}{T} \sum_{t=1}^T \Delta^{(t)}. \quad (11)$$

A convergence score is defined as

$$s_{\text{conv}} = \exp\left(-\frac{\bar{\Delta}}{T_s}\right), \quad (12)$$

where T_s is a learnable temperature parameter. Since the feature-change magnitude appears with a negative sign in the exponent, smaller feature changes produce larger convergence scores, indicating more stable refinement dynamics.

- (2) Consistency. At each iteration, a consistency estimator produces scores $s_{\text{cons}}^{(t)}$. The mean consistency score is defined as

$$s_{\text{cons}} = \frac{1}{T} \sum_{t=1}^T s_{\text{cons}}^{(t)}. \quad (13)$$

- (3) Stability. The standard deviation of consistency scores is used to measure stability

$$\sigma_{\text{cons}} = \text{std}(s_{\text{cons}}^{(t)}), \quad (14)$$

which is subsequently converted into a stability score:

$$s_{\text{stab}} = \sigma\left(\frac{1}{\sigma_{\text{cons}} + \varepsilon}\right). \quad (15)$$

Finally, these components are combined into a feedback-based known score:

$$s_{\text{feedback}} = \alpha_1 s_{\text{cons}} + \alpha_2 s_{\text{conv}} + \alpha_3 s_{\text{stab}}, \quad (16)$$

where $\alpha_1, \alpha_2, \alpha_3$ are empirically set and fixed across all datasets to ensure a consistent evaluation protocol. The feedback uncertainty is then obtained as

$$u_{\text{feedback}} = 1 - s_{\text{feedback}}. \quad (17)$$

In Figure 2, starting from the fused representation $h_{\text{fused}}^{(0)}$, the model performs T refinement iterations using lightweight refinement networks. At each iteration, the feature change magnitude is computed to characterize the refinement dynamics. In parallel, a consistency network produces iteration-wise consistency scores c_t , which are aggregated to form a global consistency estimate. These refinement statistics are used to derive two types of signals: (1) training-time feedback uncertainty u_{feedback} , computed from convergence and consistency measures, and (2) a post hoc calibration signal v_{conv} , defined as the variance of feature change magnitudes across iterations. The latter captures instability in the refinement process and serves as a key indicator for open-set detection in the EVT-based calibrator.

3.5. Multi-Signal Open-Set Scoring and Post Hoc Calibration

The proposed framework performs open-set recognition through a post hoc calibration mechanism that operates on compact and interpretable uncertainty signals derived from the multimodal fusion process. The closed-set classifier produces class probabilities as

$$p = \text{SoftMax}(\text{Classifier}(h_{\text{refined}})). \quad (18)$$

Instead of relying on classifier confidence alone, the final implementation employs a compact two-signal representation constructed from complementary uncertainty sources:

$$r(x) = [u_{\text{gate}}, \log(v_{\text{conv}} + \varepsilon)] \in \mathbb{R}^2, \quad (19)$$

denoting the score vector extracted for the sample x . u_{gate} is the gating uncertainty and v_{conv} denotes the variance of feature-change magnitudes across the iterative refinement process:

$$v_{\text{conv}} = \text{Var}\left(\left\{\Delta^{(t)}\right\}_{t=1}^T\right). \quad (20)$$

which characterizes whether the refinement process converges or diverges. This signal is later used in post hoc open-set calibration. This formulation captures both convergence dynamics and consistency behavior, providing a robust indicator of representation reliability for open-set detection.

This quantity captures the stability of the refinement dynamics, where larger variance indicates inconsistent refinement dynamics and thus higher likelihood of unknown samples. The logarithmic transformation stabilizes the dynamic range of v_{conv} and improves numerical robustness during calibration. To estimate the probability that a sample belongs to the known class set, a post hoc calibrator based on Extreme Value Theory (EVT) is employed. For each signal, values are first normalized using statistics computed from known-class validation samples and then transformed according to their directionality (i.e., whether higher values indicate known or unknown samples). A Weibull cumulative distribution function is fitted to each transformed signal. Let p_{gate} and p_{conv} denote the Weibull-based known-class probabilities derived from u_{gate} and v_{conv} , respectively. The final known-class score is computed using a weighted geometric aggregation:

$$s_{\text{known}} = \exp \left(\left(\frac{1}{(w_{\text{gate}} + w_{\text{conv}})} \right) (w_{\text{gate}} \log p_{\text{gate}} + w_{\text{conv}} \log p_{\text{conv}}) \right). \quad (21)$$

The coefficients w_{gate} and w_{conv} are aggregation weights. In the final implementation, fixed equal weights $w_{\text{gate}} = w_{\text{conv}} = 0.5$ are used for stability and fairness. Since $w_{\text{gate}} = w_{\text{conv}}$, Equation (21) reduces to the geometric mean of the two calibrated probabilities.

Importantly, the calibrator is trained using only known-class validation samples, ensuring strict open-set compliance without exposure to unknown data. Two uncertainty signals are used in the post hoc calibrator: gating uncertainty (u_{gate}) and convolutional variance (v_{conv}). Prior to EVT fitting, v_{conv} is transformed as $(\log(v_{\text{conv}} + 10^{-6}))$ to reduce scale variation. Each signal is independently normalized to $[0, 1]$ using min–max normalization with clipping to stabilize score distributions across datasets. The fitting direction is selected according to the semantic interpretation of each signal, using a negative direction for gating uncertainty and a positive direction for convolutional variance. The decision threshold τ is a data-adaptive threshold derived from the distribution of known validation scores via a fixed quantile rule:

$$\tau = \text{quantile}(\{g(r(x)) | x \in V_{\text{known}}\}, \delta), \quad (22)$$

The threshold τ is derived exclusively from known validation EVT scores, ensuring that threshold selection remains independent of unknown-class information. Although δ is constant, the resulting threshold τ remains fully data adaptive as it is derived from the empirical score distribution. The final prediction rule is given by:

$$\hat{y} = \begin{cases} \arg \max_k p_k, & s_{\text{known}} \geq \tau, \\ \text{unknown}, & s_{\text{known}} < \tau, \end{cases} \quad (23)$$

where $p \in \mathbb{R}^k$ is the closed-set class posterior; $s_{\text{known}} \in [0, 1]$ is the post hoc calibrated known-class confidence score. This formulation enables a principled and interpretable separation between known and unknown samples while maintaining strict open-set conditions.

3.6. Training Objective

The proposed framework separates representation learning from open-set decision calibration. The backbone is trained only on known-class data, while open-set detection

is handled post hoc. The optimization follows a staged training strategy. During the warm-up stage, the model is trained using cross-entropy classification and supervised contrastive loss:

$$L_{\text{total}} = L_{\text{CE}} + \lambda_{\text{con}} \cdot L_{\text{con}}, \quad (24)$$

where L_{CE} is classification loss; λ_{con} controls the contribution of the contrastive term. After warm-up, additional regularization terms are gradually introduced to improve representation robustness and uncertainty-aware separation:

$$L_{\text{total}} = L_{\text{CE}} + \lambda_{\text{con}} \cdot L_{\text{con}} + \alpha_{\text{sep}} \lambda_{\text{sep}} L_{\text{sep}}, \quad (25)$$

where L_{sep} encourages separation between reliable and boundary adjacent known samples, $\alpha_{\text{sep}} \in [0, 1]$ is a warm-up scheduler gradually increased during training, and λ_{sep} is the weighting coefficient controlling the contribution of the separation objective. This staged design stabilizes training and produces feature representations that are well-suited for post hoc open-set calibration.

3.6.1. Cross-Entropy Loss

For a batch of known-class training samples B_{known} , the classification loss is

$$L_{\text{CE}} = - \left(\frac{1}{|B_{\text{known}}|} \right) \sum_{i \in B_{\text{known}}} \log(p_{y_i}^{(i)}), \quad (26)$$

where $p_{y_i}^{(i)}$ denotes the predicted probability of the correct class for the sample i .

3.6.2. Supervised Contrastive Loss

Representative learning in the context of OSR has benefited from supervised contrastive objectives [29], which cluster same-class embeddings while separating inter-class representations, naturally improving the compactness of known-class feature manifolds. We adopt supervised contrastive learning as the backbone representation objective, combined with a novel quantile-stratified boundary separation loss that explicitly anchors the confidence head for easy- and hard-known samples. Cosine classifier normalization [57], which decouples feature norm from classification score, provides a stable logit scale across varying input distributions—an important property for the energy-based post hoc score, which is sensitive to logit magnitude.

To improve class compactness and inter-class separation in the multimodal embedding space, supervised contrastive learning is applied to the refined features. Let

$$z_i = \frac{h_{\text{refined}, i}}{\|h_{\text{refined}, i}\|_2} \quad (27)$$

be the normalized feature of the sample i , and let

$$P(i) = \{j \neq i | y_j = y_i\} \quad (28)$$

be the set of samples in the batch that share the same label. The supervised contrastive loss is

$$L_{\text{con}} = - \sum_i \left(\frac{1}{|P(i)|} \right) \sum_{j \in P(i)} \log \left(\frac{\exp \left(\frac{\text{sim}(z_i, z_j)}{\tau_c} \right)}{\sum_{k \neq i} \exp \left(\frac{\text{sim}(z_i, z_k)}{\tau_c} \right)} \right), \quad (29)$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, and τ_c is the contrastive temperature. This term encourages samples from the same known class to form compact clusters while increasing separation between different known classes. Such organization provides a more stable representation space for subsequent post hoc open-set calibration.

3.6.3. Quantile-Stratified Boundary Separation Loss

Although supervised contrastive learning improves feature compactness, it does not explicitly regulate confidence behavior near open-set boundaries. To further improve uncertainty-aware separation, a quantile-stratified boundary separation loss is introduced to distinguish reliable known samples from boundary-adjacent uncertain samples. Let the normalized entropy uncertainty used for training-time boundary separation be defined as

$$u_{\text{ent}, i} = -\frac{1}{\log K} \sum_{k=1}^K p_k^{(i)} \log(p_k^{(i)} + \varepsilon) \quad (30)$$

where $p_k^{(i)}$ represents the predicted probability of class k for the sample i , and ε is a small constant introduced for numerical stability. Let the aggregated uncertainty of the sample i be defined as

$$u_i = \frac{u_{\text{feedback}, i} + u_{\text{ent}, i}}{2}, \quad u_i = \frac{u_i - \min(u)}{\min(u) - \min(u) + \varepsilon} \quad (31)$$

where $u_{\text{feedback}, i}$ and $u_{\text{ent}, i}$ denote the feedback uncertainty and normalized entropy uncertainty, respectively. Based on the normalized uncertainty $\tilde{u}_i$, known samples are partitioned into easy-known and hard-known groups using lower and upper uncertainty quantiles, respectively. Samples with lower uncertainty are treated as reliable known samples, whereas highly uncertain samples are regarded as boundary-adjacent hard-known samples. Then, the boundary separation loss is formulated as

$$L_{\text{sep}} = \max(0, \delta_e - c_i) + \max(0, c_i - \delta_h) + \left(c_i - \left(1 - \tilde{u}_i\right)\right)^2 \quad (32)$$

where c_i denotes the confidence score produced by the MCH; δ_e and δ_h represent confidence margins that encourage high confidence for reliable samples and discourage overconfident predictions for uncertain samples, respectively.

For clarity and reproducibility, the overall inference workflow of the proposed framework is summarized in Algorithm 1, which explicitly describes the execution order and data flow from multimodal feature extraction to the final open-set decision.

The architecture and training objectives described in Section 3 are now evaluated on three benchmark datasets under a strict zero-unknown-supervision protocol. Section 4 details the experimental setup, presents quantitative comparisons, and provides ablation evidence for each component's contribution.

Algorithm 1. Inference Procedure of the Proposed Open-Set Multimodal Fusion

Input: Hyperspectral patch (x_i^{HSI}), LiDAR patch (x_i^{LiDAR}), trained network (f), calibrated threshold (τ).

Output: Predicted label $\tilde{y}_i \in y_{\text{known}} \cup \{\text{unknown}\}$

1. Modality-specific feature extraction. Extract modality-specific features:
 $[h_i^{\text{HSI}} \leftarrow E_{\text{HSI}}(x_i^{\text{HSI}}), h_i^{\text{LiDAR}} \leftarrow E_{\text{LiDAR}}(x_i^{\text{LiDAR}})]$
2. Multimodal feature fusion. Generate fused multimodal representation:
 $h_{\text{fusion}, i} = F(h_i^{\text{HSI}}, h_i^{\text{LiDAR}})$

for ($t = 1, \dots, T$) : $[h_{\text{refined}, i}^{(t)} = R(h_{\text{refined}, i}^{(t-1)})]$

Algorithm 1. *Cont.*

3. Iterative feedback refinement. Refine the fused representation through the feedback module: Initialize: $[h_{\text{refined},i}^{(0)} \leftarrow h_{\text{fusion},i}]$
4. Uncertainty estimation. Compute uncertainty signals from the refined representation: $[u_{\text{gate},i}, u_{\text{feedback},i}]$
5. Confidence prediction. Predict monotonic confidence scores using the MCH module: $[c_i \leftarrow \text{MCH}(h_{\text{refined},i})]$
6. Post hoc open-set calibration. Compute calibrated open-set score: $[s_i \leftarrow \text{Calibrator}(u_{\text{gate},i}, u_{\text{feedback},i}, c_i)]$
7. Open-set decision. If $[s_i < \tau]$ then $[\tilde{y}_i \leftarrow \text{unknown}]$
otherwise: $[\tilde{y}_i \leftarrow \arg\max_{y \in y_{\text{known}}} p(y | h_{\text{refined},i}^{(T)})]$
8. Return $\tilde{y}_i$.

4. Experiment Results and Analysis**4.1. Datasets Description**

We evaluate our methodology using three benchmark remote sensing datasets.

- The Houston2013 dataset is from the University of Houston campus and the surrounding metropolitan area. The dataset includes hyperspectral data at 144 different wavelengths (380–1050 nm), a Digital Surface Model (DSM) created using LiDAR, and detailed labels for 15 land-cover types. The dataset has a pixel size of 349–1905 pixels and a spatial resolution of 2.5 m.
- The Muufl dataset was contributed by the University of Southern Mississippi’s Gulf Park Campus. The dataset includes 64 hyperspectral images (375–1050 nm), LiDAR data with elevation, and a hyperspectral image with a spatial resolution of 1 m and a pixel size of 325×220 pixels. There are eleven land cover classes based on the ground truth document.
- The Augsburg dataset was compiled over the entire city of Augsburg. It contains 180 spectral bands with wavelengths from 0.4 to 2.5 μm . The image size is 332×485 pixels, and the ground sampling distance is 30 m. The dataset provides the final designation of seven different land-cover classes. Table 1 lists the different samples with legends for each data class.

Table 1. The information of labeled samples on three datasets.

№	Houston2013		Muufl		Augsburg	
	Class Name	Number	Class Name	Number	Class Name	Number
C1	Grass healthy	1246	Trees	21,049	Forest	11,740
C2	Grass stressed	1244	Mostly grass	3926	Residential area	29,516
C3	Grass synthetic	697	Mixed ground	5624	Industrial area	3712
C4	Trees	1244	Dirt and sand	1160	Commercial area	1592
C5	Soil	1569	Road	6034	Water	1487
C6	Residential	1253	Water	344		
C7	Commercial	1243				
C8	Road	1236				
C9	Highway	1227				
C10	Railway	1205				
C11	Parking lot 1	1233				
Unknown 1	Water	325	Building shadow	2115	Low plants	23,633
Unknown 2	Parking lot 2	468	Building	5523	Allotment	519
Unknown 3	Tennis court	428	Sidewalk	1276		
Unknown 4	Running track	660	Yellow curb	167		
Unknown 5			Cloth panels	269		

4.2. Setting Details Implementation

All experiments were conducted using the hardware and software configuration summarized in Table 2. The implementation was developed PyTorch 2.8.0 and executed on an NVIDIA RTX 5070 GPU with 12 GB of memory. The AdamW optimizer, set with a learning rate of 5×10^{-4} and a weight decay of 0.01, is used to train the model. In addition, the training process is limited to a maximum of 200 epochs, using a batch size of 32. A fixed spatial patch size of 15×15 was adopted for all datasets to jointly exploit spectral–spatial contextual information while maintaining consistent experimental settings across datasets.

Table 2. Experimental environment and hyperparameters.

Category	Component	Specification	Parameter	Value
Hardware	GPU	NVIDIA RTX 5070 (12 GB)	Optimizer	AdamW
	CPU	Intel Core i5-14600KF	Learning rate	5×10^{-4}
	Memory	32 GB RAM	Weight decay	0.01
Software	Framework	PyTorch 2.8.0	Batch size	32
	Programming language	Python 3.10.17	Patch size	15×15
	CUDA version	12.8	Training epochs	200
Model configuration	Feature dimension	64	Gradient clip	1.0
	Attention heads	4	Early stop patience	40
Strategy	Learning rate scheduler	Warm-up + cosine decay	$\alpha_1, \alpha_2, \alpha_3$	0.6, 0.2, 0.2

To stabilize the optimization process and reduce the risk of premature overconfidence in open-set recognition, we use a learning-rate schedule that combines a warm-up phase with cosine decay. We also employ gradient clipping, set to a value of 1.0, and implement early stopping based on validation performance. The feature embedding dimension is set to 64, and the multi-head attention module uses four attention heads.

All experiments were conducted using a single GPU. To facilitate reproducibility, the code used in the implementation, as well as the dataset partition protocol, will be made publicly available. Training employs a warm-up + cosine decay schedule to stabilize optimization and reduce early overconfidence.

4.3. Open-Set Evaluation Protocol

To ensure fair and reproducible comparison with existing methods, all experiments follow a strict zero-unknown-supervision protocol: no unknown-class samples are used at any stage of backbone training, uncertainty signal learning, or post hoc calibrator fitting. The decision threshold τ is derived exclusively from the distribution of known validation scores. This protocol is more stringent than several prior works that expose models to unknown samples during threshold selection [26,38], and it represents the evaluation standard we advocate for future multimodal remote sensing open-set benchmarks. In practice, the number of unknown-class samples used during both training and validation was fixed to zero, ensuring that unknown categories remained completely unseen prior to testing.

4.3.1. Training/Validation/Test Split

A fixed multi-unknown open-set protocol was adopted for all datasets, where designated unknown classes were jointly excluded from training and validation and reserved exclusively for testing. Specifically, the unknown classes were defined as {Water, parking lot 2, tennis court, running track} for Houston2013, {Building shadow, building, sidewalk, yellow curb, cloth panels} for Muufl, and {Low plants, allotment} for Augsburg, while the remaining classes were treated as known classes. Therefore, unknown classes were evaluated jointly within a single open-set scenario, rather than individually through multiple leave-class-out experiments. For each remaining known class, 20 labeled samples

were used for training and 20 labeled samples for validation, following the low-label regime commonly adopted in hyperspectral open-set learning. The validation set contains known samples only, while all unknown samples were reserved exclusively for testing. All experiments used a fixed seed list [0, 1, 2, 3, 4] which was predefined before experimentation. For each seed, the model is trained independently, the best checkpoint is selected via validation performance, and a seed-specific post hoc calibrator is fitted on the corresponding validation split. Test metrics are then computed for each seed using that checkpoint–calibrator pair. The reported results represent the mean and standard deviation of these five independent test evaluations.

Due to the patch-based representation strategy, pixels near image boundaries are excluded since a complete $P \times P$ neighborhood cannot be formed; only spatially valid pixels are therefore used in all experiments. The resulting effective sample counts are summarized in Table 3.

Table 3. Training, validation, and test sample partition under the strict open-set protocol.

Dataset	Train	Validation	Test (Known)	Test (Unknown)	Total Test
Houston2013	220	220	12,577	1881	14,458
MuufI	120	120	38,537	9350	47,887
Augsburg	100	100	47,847	24,152	71,999

For fair comparison, all baseline methods were adapted to the same strict open-set evaluation protocol while preserving their original core architectures, optimization strategies, and decision mechanisms without introducing additional modules or post hoc modifications. In particular, the original WGDt dynamic thresholding mechanism and loss design were preserved without additional post hoc calibration or architectural modifications.

4.3.2. Evaluation Metrics

The performance of the open-set classification is evaluated using both closed-end and open-end portfolio indicators represented in Table 4. To ensure reproducibility and reduce variance due to random initialization and sampling, all experiments are repeated five independent times using different random seeds; results are reported as mean $\pm$ standard deviation across the five runs.

Table 4. Evolution metrics.

Metric	Meaning	Way
Closed-Set Metrics		
Overall accuracy (OA)	Standard classification accuracy on known test samples only	$OA = \frac{\text{correctly classified known samples}}{\text{total known test samples}} \times 100\%$
Average accuracy (AA)	Mean classification accuracy across all classes.	$AA = \frac{1}{C} \sum_{i=1}^C \frac{TP_i}{N_i} \times 100\%$ where C is the number of known classes, TP_i denotes the number of correctly classified samples of class i , and N_i denotes the total number of samples belonging to class i .
Cohen's Kappa	Agreement between predicted and true labels corrected for chance.	$k = \frac{p_o - p_e}{1 - p_e}; p_e = \sum_{i=1}^C \frac{N_i^{(pred)} N_i^{(true)}}{N^2};$ where p_o is observed accuracy and p_e is chance agreement. $N_i^{(pred)}$ predicted samples of class i , $N_i^{(true)}$ true samples of class i , and N total samples.
Open-set metrics		
Known accuracy	Proportion of known samples correctly accepted as known at the operating threshold.	Known Accuracy = $\frac{\text{correctly accepted known samples}}{\text{total known test samples}} \times 100\%$
Unknown accuracy	Proportion of unknown samples that are correctly rejected at the operating threshold.	Unknown Accuracy = $\frac{\text{correctly rejected unknown samples}}{\text{total unknown samples}} \times 100\%$

Table 4. Cont.

Metric	Meaning	Way
Area under the ROC curve (AUROC)	Measures the ability to discriminate between known and unknown samples across all possible thresholds. True Positive Rate (TPR) measures the proportion of unknown samples correctly detected as unknown, while False Positive Rate (FPR) measures the proportion of known samples incorrectly rejected as unknown.	$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR})d(\text{FPR})$ In this context, unknown samples are considered the positive class. $\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$; $\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$; TP = unknown samples correctly detected as unknown FN = unknown samples incorrectly accepted as known FP = known samples incorrectly rejected as unknown TN = known samples correctly accepted as known
False Positive Rate at 95% True Positive Rate (FPR95)	Measures the false open-set decision rate of known samples when the system correctly identifies 95% of unknown samples.	$\text{FPR95} = \text{FPR} \mid \text{TPR} = 0.95$
HOS (Harmonic Open-set Score)	Measures the balanced trade-off between known-class recognition and unknown-class rejection.	$\text{HOS} = \frac{2 \times \text{Known Accuracy} \times \text{Unknown Accuracy}}{\text{Known Accuracy} + \text{Unknown Accuracy}}$ where Known Accuracy and Unknown Accuracy denote known-class accuracy and unknown-class accuracy, respectively. A higher HOS indicates a more balanced open-set performance.

4.4. Architecture and Training Hyperparameters Sensitivity

4.4.1. Effect of Hyperparameters on Open-Set Recognition Performance

We conducted experiments to optimize the tuning parameters by varying key hyperparameters independently and keeping the others constant. The results for Houston2013 are shown in Figure 3. The red bars indicate the baseline configuration carried forward during sequential tuning, while the blue bars represent the tested candidate values at each step. Results show that performance is sensitive to feature dimensionality and separation-related parameters, with optimal performance achieved at moderate values, while excessively large dimensions or weights degrade open-set discrimination. When analyzing the baseline with feature dimension = 256, dropout = 0.1, contrastive weight = 0.3, alpha separation = 0.3, and separation weight = 0.15, it was found that feature dimension = 64 showed the highest results and all other baselines showed the best computational performance. Therefore, we will use these hyperparameters consistently in all future experiments.

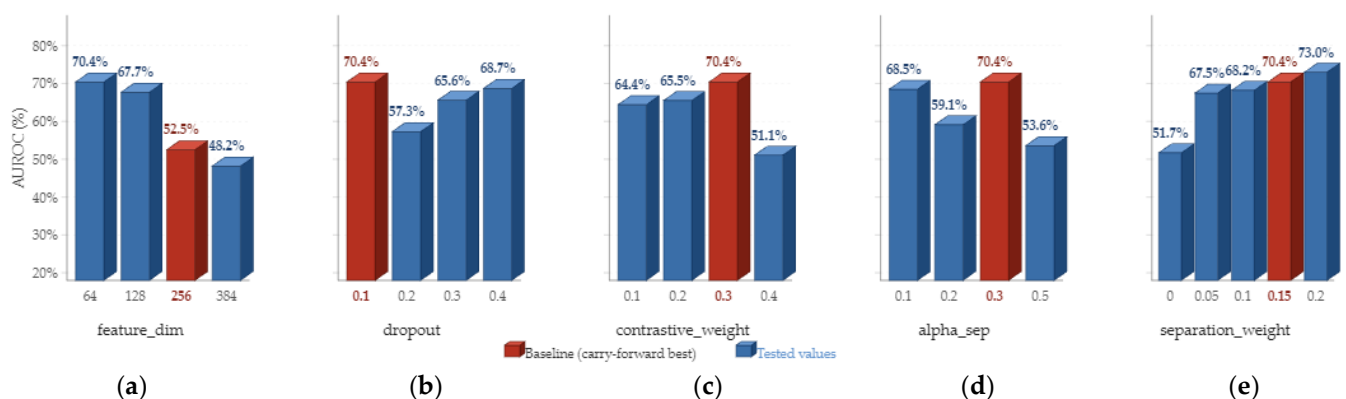


Figure 3. Sensitivity analysis of key hyperparameters in the proposed open-set multimodal framework on the Houston2013 dataset (AUROC, %). (a) Impact of feature dimension; (b) dropout; (c) contrastive weight; (d) alpha separation; (e) separation weight.

4.4.2. Iterative Feedback Refinement Count Analysis

Table 5 reports the effect of varying the feedback iteration count n on open-set recognition performance, computational cost, and model size. At $n = 1$, the lack of iterative refinement yields near-random open-set detection (AUROC = 44.60%, unknown accuracy = 38.66%), despite a competitive OA of 92.40%, confirming that a single forward pass is not sufficient for reliable known/unknown discrimination. Increasing to $n = 2$ yields the largest improvements in open-set metrics (AUROC = 72.20%, unknown accuracy = 77.08%), albeit OA drops to 90.78%, imply-

ing that two iterations induce mild instability in closed-set classification. At $n = 3$, the model achieves the highest OA (92.79%), the lowest inference time (5.91 ms), and retains competitive open-set performance (AUROC = 70.53%, unknown accuracy = 76.48%) simultaneously, indicating that three iterations reach a stable equilibrium between closed-set and open-set objectives. Further increasing n to 5 and 7 results in gradual degradation across all metrics, with inference time increasing to 8.21 ms and 9.96 ms respectively, and parameter count increasing from 2.42 M at $n = 1$ to 2.62 M at $n = 7$ due to the additional feedback pathway parameters.

Table 5. The effect of the feedback iteration count n on open-set recognition performance is examined. Results are mean $\pm$ standard deviation over 5 random seeds.

Metrics	$n = 1$	$n = 2$	$n = 3$	$n = 5$	$n = 7$
OA (%)	92.40	90.78	92.79	89.58	90.52
Unknown accuracy (%)	38.66	77.08	76.48	57.50	65.66
AUROC (%)	44.60	72.20	70.53	57.46	66.14
Inference time (ms)	6.05	6.00	5.91	8.21	9.96
Parameter (M)	2.42	2.45	2.48	2.55	2.62

Bold values denote the best result per column.

Three-dimensional performance trade-off space of feedback iteration count n , where each axis represents OA (%), unknown accuracy (%), and AUROC (%), sphere size encodes inference time, and $n = 3$ (★) achieves the optimal balance—highest OA (92.79%) with competitive open-set detection (unknown accuracy = 76.48%, AUROC = 70.53%) at the lowest inference cost (5.91 ms) in Figure 4. Based on this joint analysis of accuracy, open-set detection, inference efficiency, and model compactness, $n = 3$ is selected as the optimal configuration throughout this work.

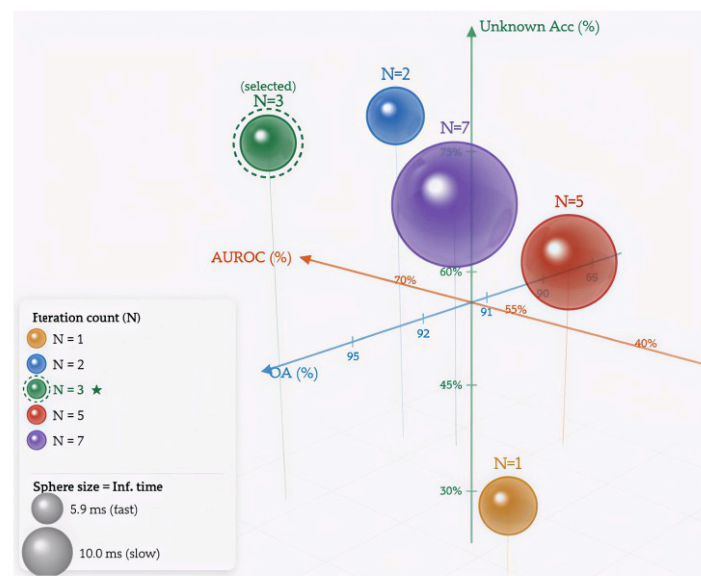


Figure 4. Three-dimensional efficiency trade-off space of iteration feedback refinement. Sphere size represents inference time, and ★ denotes the selected optimal iteration setting ($n = 3$) in the final model.

4.4.3. Sensitivity Analysis of the Quantile Level (δ)

The quantile level (δ) in Equation (22) determines the threshold used for post hoc open-set calibration. To analyze its influence, δ was varied within the range [0.30, 0.60] with an interval of 0.05 on the Houston2013, Muufl, and Augsburg datasets. Figure 5 presents the corresponding variations in HOS, known-class accuracy, and unknown-class accuracy.

As increases, unknown-class accuracy generally increases while known-class accuracy gradually decreases, reflecting the expected trade-off induced by stricter thresholding. The optimal operating point differs slightly from dataset to dataset. On Houston2013, the highest HOS is reached at $\delta = 0.50$, but Augsburg obtains the best performance at approximately $\delta = 0.45$. In contrast, Muufl indicates relatively stable behavior on a broader range, with the highest HOS reached at $\delta = 0.30$. However, the performance differences around the optimal region remain relatively small, especially within $\delta \in [0.40, 0.50]$. Given the goal of maintaining a unified protocol across datasets and avoiding dataset-specific parameter tuning, $\delta = 0.45$ was used in all experiments. This setting provides consistently competitive HOS performance while preserving a balanced trade-off between known-class recognition and unknown-class rejection across different datasets.

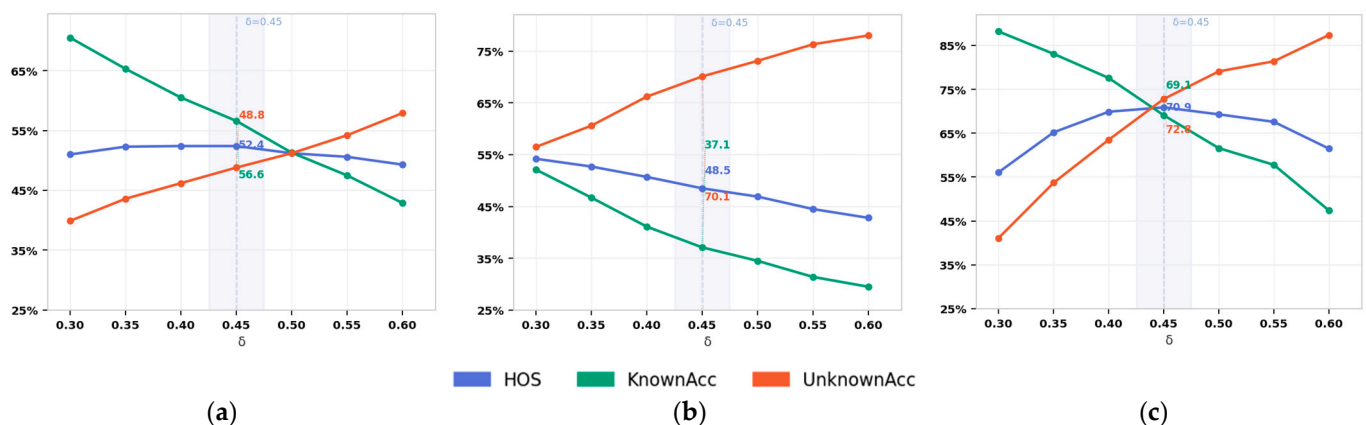


Figure 5. Sensitivity analysis of the quantile level (δ) on datasets. (a) Houston2013; (b) Muufl; (c) Augsburg. The light blue shaded region indicates the selected quantile region centered at ($\delta = 0.45$), which was adopted as the unified setting across datasets.

4.4.4. Contribution of LiDAR to Open-Set Recognition

To quantitatively investigate the contribution of LiDAR to open-set recognition, modality ablation experiments were conducted under three settings, HSI only, LiDAR only, and HSI + LiDAR multimodal fusion, using the same experimental protocol and five-seed evaluation setting. The results are summarized in Table 6.

Table 6. Quantitative contribution of LiDAR to open-set recognition performance under different modality settings.

Metrics	Houston2013			Muufl			Augsburg		
	HSI	LiDAR	HSI + LiDAR	HSI	LiDAR	HSI + LiDAR	HSI	LiDAR	HSI + LiDAR
OA (%)	79.91	44.73	92.79	78.58	56.84	84.47	77.98	41.59	80.99
Known accuracy	53.82	55.50	93.07	66.65	59.50	86.39	63.72	52.05	69.06
Unknown accuracy	43.82	56.88	76.48	56.58	36.91	63.91	38.61	44.05	56.81
HOS	47.41	55.35	83.96	60.23	42.43	73.47	41.89	45.26	62.34

Bold values denote the best performance.

Across all datasets, LiDAR demonstrates a complementary role in improving unknown-class discrimination. On Houston2013, HSI-only achieves stronger closed-set discrimination than LiDAR-only; however, multimodal fusion substantially improves unknown accuracy from 43.82% and 56.88% to 76.48%, while also achieving the highest HOS (83.96%), indicating a more balanced recognition of known and unknown samples. In Augsburg, similar behavior is observed where HSI + LiDAR increases the unknown accuracy to 56.81% and improves HOS to 62.34%, suggesting that LiDAR provides useful structural information for distinguishing unseen land-cover categories. For Muufl, HSI-only already achieves relatively strong open-set performance, indicating that spectral information is

highly discriminative in this dataset. However, multimodal fusion still improves unknown accuracy and HOS, reaching 63.91% and 73.47%, respectively. These results suggest that the contribution of LiDAR is dataset dependent but consistently complementary, particularly for improving unknown-class rejection in open-set scenarios.

4.5. Main Results on Datasets

We compare open-set classification approaches such as supervised confidence learning (OSC-SCL) [23], open-set domain adaptation and adversarial network (WGDT) [38], hyperspectral single-modal methods, and hierarchical advanced training for the collaborative multimodal open-set recognition process (HyLiOSR) [26]. To comprehensively evaluate the performance of the proposed method, we present detailed classification accuracy results for three datasets. Tables 6–8 show the classification results for Houston2013, Muufl, and Augsburg datasets with the performance of the methods using one modality (hyperspectral) and multimodal (HSI + LiDAR) data in the total metrics.

Table 7. Classification outcomes of all methodologies used in the Houston2013 dataset.

Modality	HSI		HSI + LiDAR	
Class	OSC-SCL [23]	WGDT [38]	HyLiOSR [26]	Proposed
Grass healthy	97.55	76.91	92.98	87.79
Grass stressed	97.54	75.72	94.96	99.07
Grass Synthetic	99.56	100.00	99.89	100.00
Trees	97.22	86.91	93.33	97.52
Soil	99.61	98.12	96.92	99.44
Residential	85.19	72.57	78.19	97.56
Commercial	83.40	87.36	63.50	88.99
Road	80.78	65.56	80.42	86.56
Highway	92.08	88.87	87.79	80.44
Railway	87.65	91.18	88.71	91.81
Parking lot 1	89.30	95.49	90.77	94.62
Unknown	51.30	57.59	44.12	76.48
OA (%)	86.37 ± 0.53	81.22 ± 0.81	82.01 ± 0.78	92.79 ± 0.84
AA (%)	88.43 ± 0.67	82.98 ± 0.82	84.30 ± 1.61	93.07 ± 0.80
Kappa×100	88.54 ± 0.87	79.40 ± 0.86	80.34 ± 0.89	92.05 ± 0.93
HOS	64.93	67.96	57.92	83.96

The best result is highlighted in bold.

Table 8. Classification outcomes of all methodologies used in the Muufl dataset.

Modality	HSI		HSI + LiDAR	
Class	OSC-SCL [23]	WGDT [38]	HyLiOSR [26]	Proposed
Trees	78.19	70.69	89.57	85.60
Mostly grass	65.31	68.53	80.66	76.70
Mixed ground	62.68	58.94	56.28	80.98
Dirt and sand	87.28	91.38	89.92	92.20
Road	69.73	65.05	83.19	92.03
Water	99.01	99.01	97.85	98.57
Unknown	48.94	48.51	46.48	63.91
OA (%)	68.98 ± 0.01	65.01 ± 0.01	74.32 ± 1.99	84.47 ± 1.07
AA (%)	73.02 ± 0.06	71.76 ± 0.12	78.64 ± 1.43	86.39 ± 1.15
Kappa×100	72.38 ± 0.03	54.37 ± 0.06	66.11 ± 2.26	76.88 ± 1.48
HOS	58.60	57.86	58.43	73.47

The best result is highlighted in bold.

Table 7 compares the classification results of multiple methods on the Houston2013 dataset. The proposed method achieves high performance in most classes, and especially in the multimodal case. OA = 92.79%, AA = 93.07%, and Kappa = 92.05%, which are higher than other methods. All the methods performed well for each class in the grass synthetic class and the proposed method achieved 100% accuracy. The proposed method

performed similarly to or better than OSC-SCL and HyLiOSR in trees and soil classes, showing high discrimination. In urban categories such as residential and commercial, the proposed method showed significant improvement (e.g., residential (97.56%)). In the road class, the proposed method (86.56%) outperformed WGDT (65.56%) and HyLiOSR (80.42%), demonstrating the importance of multimodal information. However, in some classes, like the highway class, the proposed method (80.44%) did not work well compared to the other methods, implying that the spectral and spatial features of this class are more complex. In the unknown class, the proposed method achieved an unknown accuracy of 76.48%, which is better than OSC-SCL (25.18%), WGDT (18.89%), and HyLiOSR (32.36%). On HOS, the proposed method obtains the highest value of 83.96%, which is comparatively higher than OSC-SCL (64.93%), WGDT (67.96%) and HyLiOSR (57.92%). This shows the proposed framework has a more balanced trade-off between known-class recognition and unknown-class rejection and has a better open-set discrimination capability on Houston2013 dataset.

Table 8 displays the results of classifying the Muufl dataset. The proposed method showed the highest overall performance using multimodal data, with OA = 84.47%, AA = 86.39% and Kappa = 76.88%. This suggests that it is significantly superior to the WGDT (OA = 65.01%) and OSC-SCL (OA = 68.98%) methods that use only hyperspectral data. The results of each class showed that the proposed method achieved the highest performance (92.20% and 92.03%) in the dirt and sand and road classes, respectively, and demonstrated that multimodal fusion can effectively combine spatial and spectral information. The proposed method (80.98%) was better than the other methods in the mixed ground class, which is usually difficult to distinguish due to the similar spectral characteristics of these classes. The proposed method attains the best HOS value of 73.47%, exceeding OSC-SCL (58.60%), WGDT (57.86%), and HyLiOSR (58.43%). The improvement suggests that the proposed uncertainty-aware framework more effectively preserves known-class performance while simultaneously improving unknown-sample rejection under challenging multimodal conditions.

In the trees and mostly grass classes, the proposed method showed consistently high performance, but HyLiOSR slightly outperformed in some cases, indicating the difference in domain adaptation strategies. In the water class, all methods showed very high accuracy (≈ 97 – 99%), indicating that this class has high spectral discrimination. The proposed method achieved 63.91% accuracy in the unknown class, which is higher than OSC-SCL (14.97%), WGDT (15.40%) and HyLiOSR (17.43%).

Table 9 suggests that the proposed method outperforms other methods based on multimodal fusion on the Augsburg dataset with OA = 79.89%, AA = 66.90%, and Kappa = 67.70%. In terms of the results for each class, the proposed method is close to WGDT and HyLiOSR in the forest class (94.02%) but slightly lower than OSC-SCL, which indicates that all methods perform well in classes with high spectral differences. The proposed method has the highest performance in the residential area class (80.25%), which indicates that the spatial differences are better utilized in multimodal information. The proposed method performs substantially improved than HyLiOSR (54.43%) in the industrial area class but lower than WGDT (65.47%), which indicates that domain adaptation-based methods may be preferred in some cases. For the commercial area class, the proposed method is close to OSC-SCL (41.46%) but outperforms WGDT and HyLiOSR. The proposed method performs best in the water class (64.35%), which indicates that multimodal fusion can effectively combine spectral and geometric information. In the unknown class, the proposed method achieves an accuracy of 56.81%, outperforming all baselines, OSC-SCL (12.51%), WGDT (6.53%) and HyLiOSR (11.33%), but unknown detection is still the most difficult part of the whole system. The proposed method achieves the best HOS value of 62.34%, outperforming OSC-SCL (50.86%), WGDT (54.08%) and HyLiOSR (46.51%).

Despite Augsburg being a more challenging open-set scenario, the proposed method also demonstrates a better trade-off between the known-class classification and the unknown detection, which further validates its robustness in complex urban environments.

Table 9. Classification outcomes of all methodologies used in the Augsburg dataset.

Modality	HSI		HSI + LiDAR	
Class	OSC-SCL [23]	WGDT [38]	HyLiOSR [26]	Proposed
Forest	96.07	90.88	90.41	93.78
Residential area	79.33	60.46	71.44	80.96
Industrial area	34.55	65.47	20.49	63.32
Commercial area	42.95	24.83	28.22	35.52
Water	61.06	59.08	29.44	71.73
Unknown	44.30	50.28	45.48	56.81
OA (%)	66.63 ± 0.03	61.61 ± 0.04	61.38 ± 2.19	80.99 ± 3.51
AA (%)	59.71 ± 0.01	58.50 ± 0.03	47.58 ± 6.64	69.06 ± 1.50
Kappa × 100	70.30 ± 0.03	47.71 ± 0.04	46.65 ± 3.95	69.08 ± 4.94
HOS	50.86	54.08	46.51	62.34

The best result is highlighted in bold.

Overall, the proposed method improved the overall classification performance by using multimodal fusion and showed consistently high accuracy in most classes but still needs to be improved in the unknown class. Figure 6 depicts the 3D performance landscape of the surface view on comparison approaches across three benchmark datasets (Houston2013, Muufl, and Augsburg). The vertical axis represents classification performance, and the horizontal axes denote the compared methods (OSC-SCL, WGDT, HyLiOSR, and the proposed method) and datasets, respectively. The color gradient from blue to red indicates low to high performance values, which qualitatively indicates that the proposed method consistently achieves better results on all datasets and evaluation metrics.

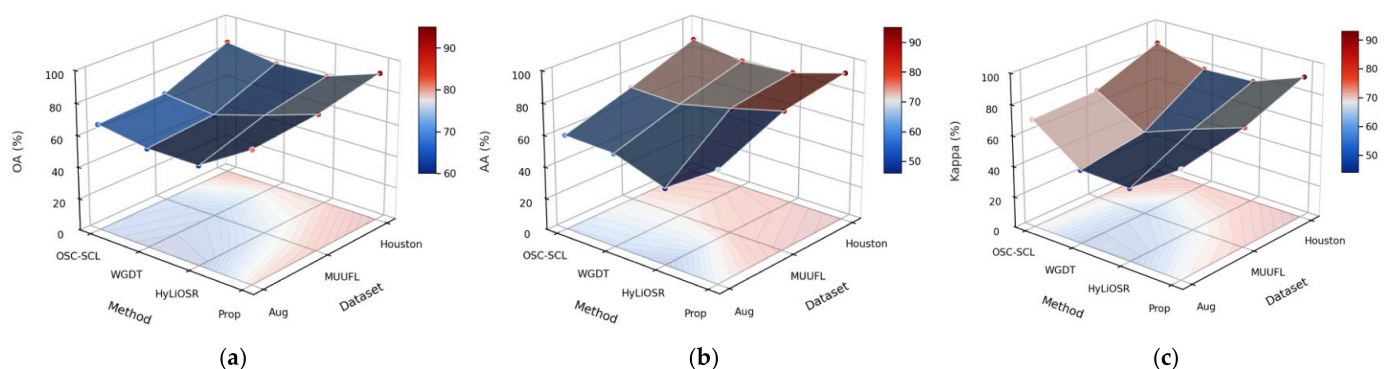


Figure 6. 3D performance landscape of surface view on comparison approaches. (a) Overall accuracy performance landscape. (b) Average accuracy performance landscape. (c) Kappa coefficient performance landscape.

Figures 7–9 present qualitative comparisons showing that the proposed method (f) produces cleaner, more spatially coherent classifications than competing approaches across all three datasets, with particularly notable improvements in boundary delineation and unknown class detection (white regions) compared to baseline methods.

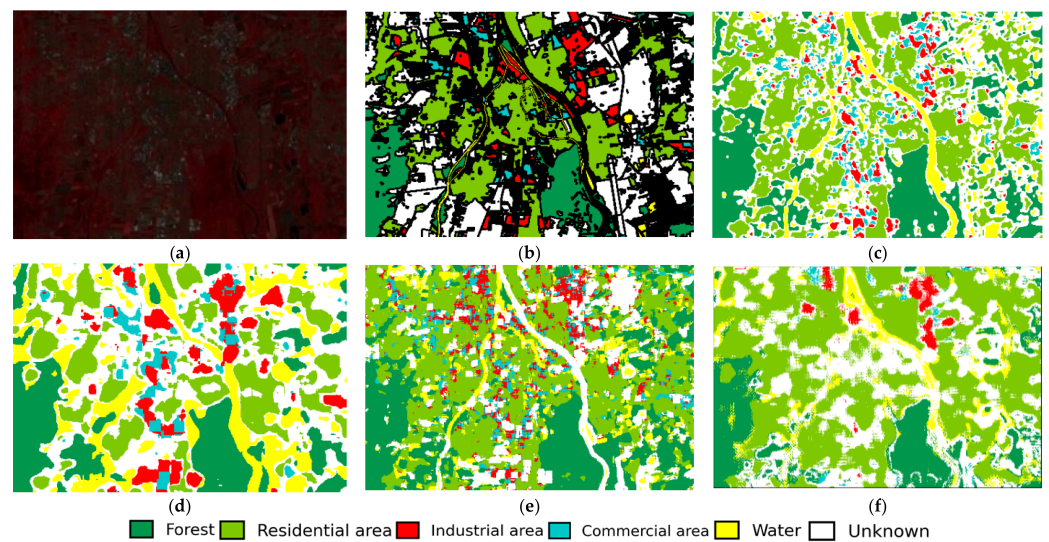


Figure 7. Classification maps of Augsburg. (a) Ground-truth map. (b) RGB origin. (c) OSC-SCL. (d) WGD. (e) HyLiOSR. (f) Proposed.

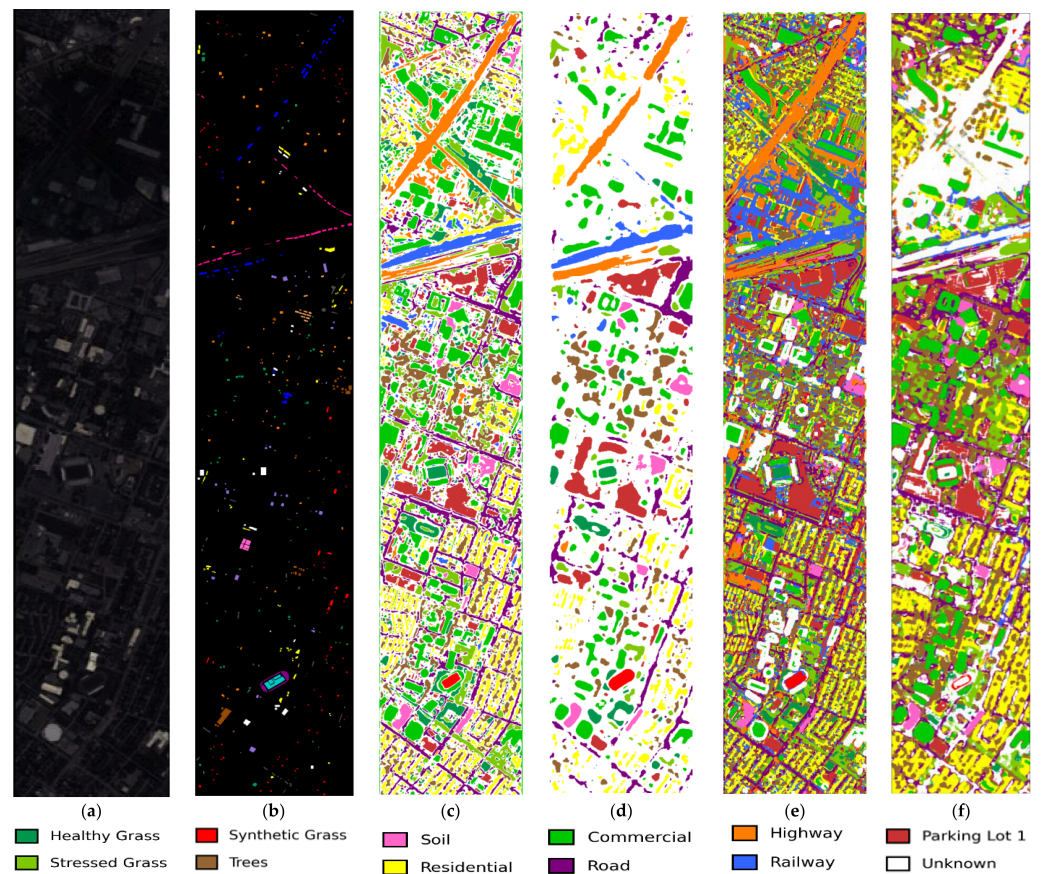


Figure 8. Classification maps of Houston2013. (a) Ground-truth map. (b) RGB origin. (c) OSC-SCL. (d) WGD. (e) HyLiOSR. (f) Proposed.

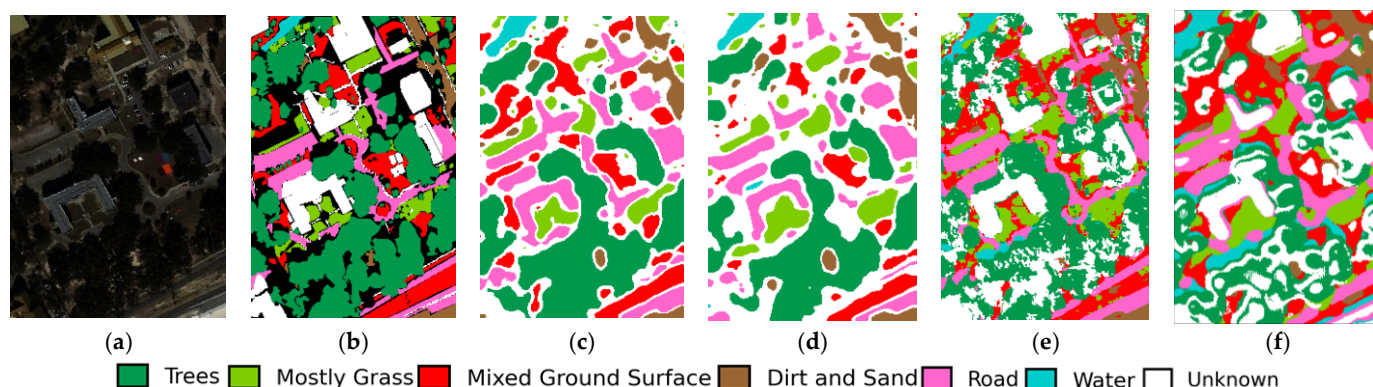


Figure 9. Classification maps of Muufl. (a) Ground-truth map. (b) RGB origin. (c) OSC-SCL. (d) WGD. (e) HyLiOSR. (f) Proposed.

To further strengthen qualitative analysis, Figure 10 visualizes modality-aware confidence heatmaps on the Muufl dataset. The HSI and LiDAR contribution heatmaps demonstrate that the proposed uncertainty-aware gating mechanism adaptively emphasizes different modalities according to local class characteristics, which reflects the complementary multimodal utilization. Vegetation-related categories, including trees and mostly grass, tend to have relatively stronger HSI contributions, reflecting the importance of spectral characteristics for discrimination. Meanwhile, structurally complex regions show increased LiDAR contribution, indicating the complementary role of elevation-related information. Moreover, the MCH confidence heatmap shows relatively higher confidence in homogeneous regions and reduced confidence near class boundaries and mixed regions, qualitatively supporting the reliability of the proposed open-set recognition mechanism.

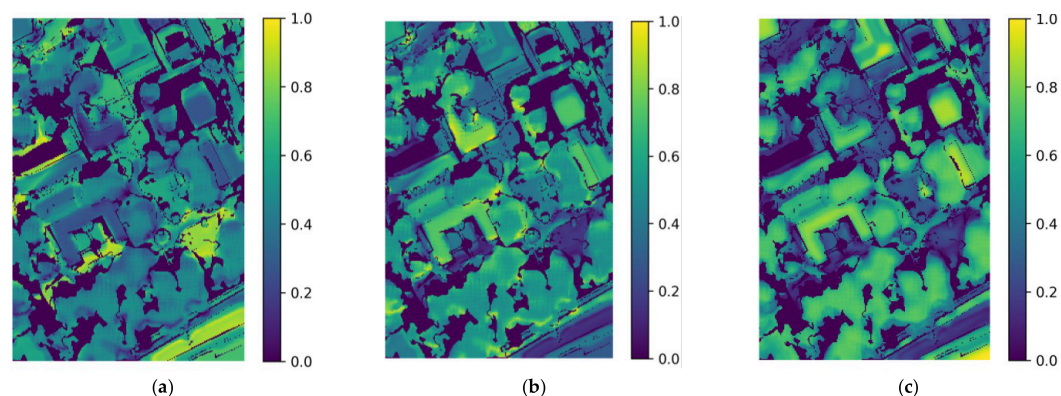


Figure 10. Modality-aware confidence heatmaps on the Muufl dataset. (a) HSI contribution heatmap; (b) LiDAR contribution heatmap; (c) MCH confidence heatmap.

Figure 11 figured the t-SNE representation of the learned feature representation for the three datasets; the gap represents the confidence difference (known–unknown). For visualization clarity, the embeddings were generated using randomly sampled test pixels, including 2500 known-class samples and 2500 unknown-class samples. The t-SNE embeddings across datasets qualitatively suggest that the proposed framework produces relatively separated known-class clusters with unknown samples (×) positioned in peripheral or inter-cluster regions of the feature space (a), and the four-category prediction outcome maps (b) qualitatively indicate that post hoc calibrator successfully identifies the majority of unknown samples as true unknowns (green) while keeping the false-acceptance rate (red) low, with threshold values of $\tau = 0.285$, $\tau = 0.252$, and $\tau = 0.456$ for Houston2013, Muufl, and Augsburg respectively; the continuous s_{known} score maps (c) further reveal that the calibrated EVT calibration produces well-calibrated probabilistic outputs rather than

binary hard decisions, and the confidence gap across the three datasets provides qualitative insight into the discriminative behavior of the two-signal representation.

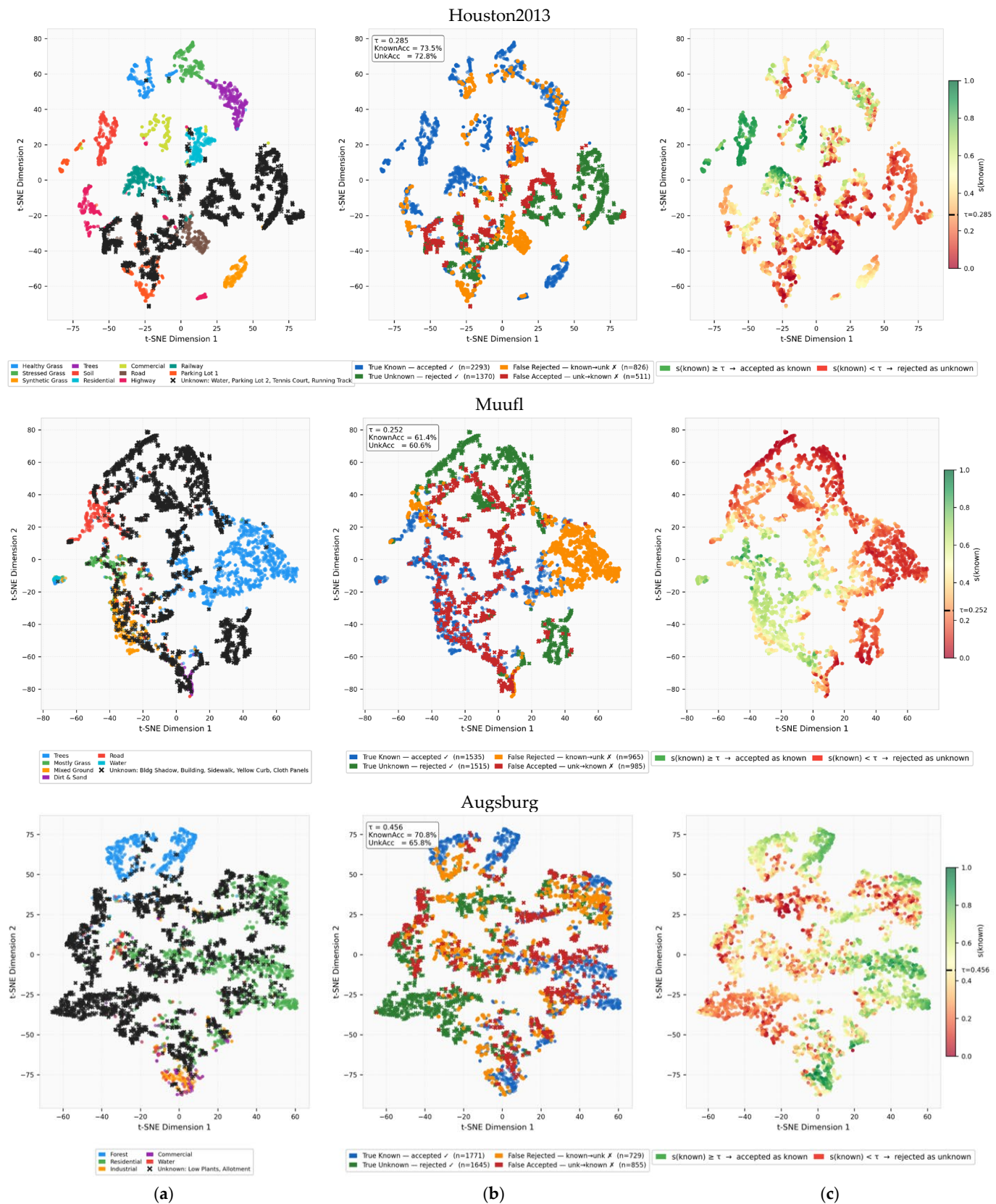


Figure 11. t-SNE visualization of open-set recognition on the three datasets (balanced best seed). (a) Ground-truth class distribution. (b) Four-category prediction outcome. (c) s_{known} score map.

4.6. Ablation Study

4.6.1. Component Contributions

Table 10 shows the results of the component ablation study on the Houston2013 dataset. The cross-entropy (CE) baseline shows a good classification performance of OA = 91.5% but unknown accuracy = 0.00% and AUROC = 50.0%, which confirms that the model without a special mechanism cannot fully recognize unknown classes. There is no change in the classification performance when adding supervised contrastive loss (SCL) (CE + SCL), and the open-set detection performance remains the same, indicating that contrastive learning alone is not sufficient for unknown detection. Adding an Uncertainty-Aware Gate Fusion (UAGF) reduces the OA to 72.7%, but unknown accuracy reaches 34.9% and AUROC is 45.7%. The decrease in OA is due to the excessive classification of known classes as unknown without the gate mechanism Iterative Feedback Refinement (IFR), which indicates the instability of the UAGF model. When IFR is added (UAGF_IFR, iter = 3), OA recovers to 91.2%, unknown accuracy improves to 40.8%, and FPR95 improves to 82.6%. This indicates that feedback refinement fine-tunes the known/unknown boundary by iteratively processing the gate decision.

Table 10. Ablation study of component contributions on Houston2013 dataset.

Ablation Variant	UAGF	IFR	MCH	Iters	OA (%)	Unknown Accuracy (%)	AUROC	FPR95
CE	-	-	-	1	91.5	0.00	50.0	100.0
CE + SCL	-	-	-	1	91.5	0.00	50.0	100.0
UAGF	✓	-	-	1	72.7	34.9	45.7	92.1
UAGF_IFR	✓	✓	-	3	91.2	40.8	53.0	82.6
UAGF_IFR_MCH	✓	✓	✓	3	92.8	76.5	70.5	67.5

Bold highlights denote the optimal values for each entry in the dataset. ✓ denotes the presence of the corresponding component, whereas—indicates its absence.

The full model with MCH loss (UAGF_IFR_MCH) achieves the highest performance in all parameters, reaching OA = 92.8%, unknown accuracy = 76.5%, AUROC = 70.5%, and FPR95 = 67.5%. The ability of MCH to compact known class clusters improves the threshold definition of the EVT calibrator and allows for a clear separation of unknown samples. In particular, the increase in unknown accuracy of +35.7 pp confirms the direct contribution of MCH loss.

While the preceding ablation identifies which components are necessary, it does not address whether additional regularization could further improve the uncertainty signals. We therefore examine the effect of VICReg regularization—a diversity penalty originally proposed to prevent signal collapse in self-supervised learning [30]—on open-set detection performance.

4.6.2. The Role of the MCH Separation Loss

VICReg was selected as the test condition because its variance-spreading objective represents the conceptual opposite of MCH's compactness objective: if MCH's intra-class compaction is the primary driver of open-set performance, then a regularizer that counteracts compaction should degrade open-set detection while leaving closed-set accuracy unchanged. Figure 12 confirms this prediction across five independent seeds on Houston2013: adding VICReg degrades unknown accuracy by 17.4 pp and AUROC by 11.3 pp with no gain in OA, and the degradation is consistent across all seeds (Figure 11b). This result provides direct evidence that the MCH-induced compactness of known-class clusters is the critical property that enables the EVT calibrator to set a discriminative threshold and that any regularization conflicting with this compactness objective is detrimental.

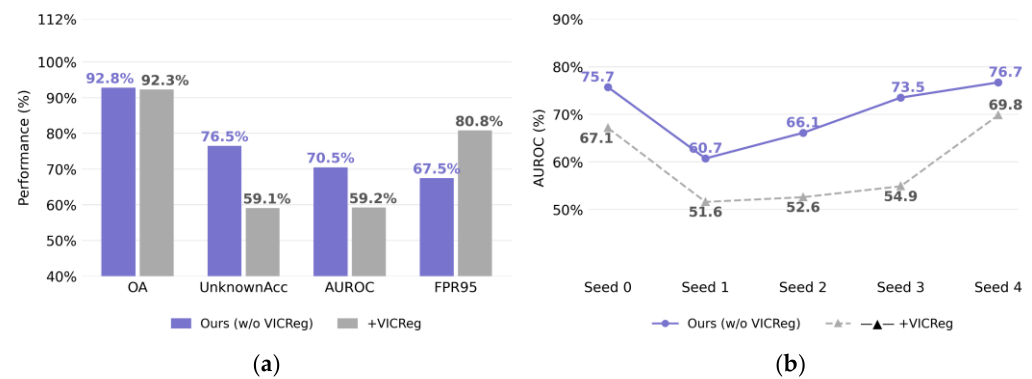


Figure 12. Impact difference of our MCH and VICReg regularization on open-set recognition (Houston2013). (a) Open-set metrics comparison. (b) Seed-level AUROC consistency.

4.6.3. Computational Complexity Analysis

Table 11 reports the computational complexity of different methods on the Houston2013 dataset in terms of model parameters, floating-point operations (FLOPs), and inference time. All methods were evaluated under the same hardware and implementation settings using identical input conditions. HyLiOSR exhibits the lowest computational cost, while OSC-SCL shows the highest inference latency (13.241 ms/sample). The proposed method introduces a larger model capacity (2.465 M parameters) to support multimodal fusion and uncertainty-aware refinement while maintaining practical inference efficiency (3.599 ms/sample). Although the proposed framework incurs additional computational overhead compared with lightweight baselines, it achieves a favorable balance between computational cost and open-set recognition performance.

Table 11. Computational complexity analysis on Houston2013 dataset.

Method	Parameters (M)	FLOPs (M)	Inference Time (ms/Sample)
OSL-SCL	0.538	879	13.241
WGDT	0.560	423	1.076
HyLiOSR	0.274	240	1.000
Proposed	2.465	152	3.599

The quantitative results establish the competitive performance of the proposed method. Section 5 examines the underlying factors that drive these results, analyzes failure modes, and situates the findings within the broader open-set recognition literature.

5. Discussion

5.1. Why Closed-Set Training Cannot Detect Unknown Samples

The ablation results reveal a fundamental limitation of purely discriminative training for open-set classification: a cross-entropy classifier with strong closed-set accuracy (OA = 91.5%) assigns high SoftMax confidence to out-of-distribution inputs because it is never penalized for doing so during training. This behavior is consistent with the theoretical analysis of Vaze et al. [14], who showed that SoftMax outputs are unreliable indicators of distributional membership. Supervised contrastive loss improves feature compactness but does not alter this behavior, as the contrastive objective optimizes relative distances between known classes without constraining the score assigned to samples outside the training distribution. This finding motivates the post hoc calibration design adopted in this work: rather than modifying the closed-set training objective to explicitly penalize unknown-like behavior, we decouple representation learning from open-set decision-making. The backbone is trained to produce discriminative known-class features,

and the EVT calibrator is subsequently fitted to distinguish the distributional tail of known samples from potential unknowns. This separation avoids the objective conflicts that arise when classification and rejection losses are optimized simultaneously.

5.2. The Stabilizing Role of Iterative Feedback Refinement

The sharp OA drop observed when UAGF is introduced without IFR (72.7%) suggests that gating uncertainty alone introduces instability: without iterative refinement to progressively consolidate multimodal representations, the gating mechanism produces high-variance uncertainty estimates that cause excessive rejection of borderline known samples. IFR resolves this tension by providing a feedback pathway that iteratively adjusts the fused representation towards a stable fixed point. Known samples, whose spectral–structural signatures are consistent with the training distribution, converge rapidly across iterations; unknown samples, which lack such consistency, exhibit larger and more variable feature changes. This behavioral difference is precisely what the convergence variance signal v_{conv} captures, providing the EVT calibrator with a complementary and interpretable indicator of distributional membership. The non-monotonic behavior across iteration counts further supports this interpretation. Two iterations establish a meaningful convergence trajectory but are insufficient for full stabilization, producing mild representation instability that slightly reduces closed-set accuracy. Three iterations achieve the critical threshold at which the feedback pathway fully stabilizes known class representations while preserving sufficient dynamic range in the convergence signal to discriminate against unknown samples.

5.3. Cross-Dataset Performance Gaps and Their Origins

The progressive performance decline from Houston2013 to Muufl to Augsburg reflects two independent sources of difficulty that are not equally addressable by the proposed framework. The first source is class imbalance. Augsburg’s Commercial area class contains fewer than 6% of the samples of the residential area class, causing the EVT calibrator to assign disproportionate weight to majority-class score statistics when estimating the known-class boundary. This imbalance is a data-level challenge that neither the UAGF nor the IFR module can overcome architecturally, and it points to a genuine limitation of the current framework in severely imbalanced settings. The second source is spectral ambiguity between known and unknown classes. In Augsburg, the low plant unknown class shares spectral characteristics with the forest and residential area known classes. Because of the gating uncertainty, a signal is computed from modality reliability disagreement rather than semantic content; it cannot distinguish a spectrally ambiguous known sample from a spectrally similar unknown sample when both modalities agree on the wrong assignment. This represents the fundamental boundary of what uncertainty-based open-set detection can achieve without access to semantic prototypes or explicit unknown modeling.

5.4. VICReg Conflict with the Separation Objective

The consistent performance degradation observed when VICReg is added provides evidence for a previously unreported interaction between variance regularization and margin-based separation losses. VICReg encourages each uncertainty signal to independently maintain high variance across the training batch, preventing signal collapse. However, the MCH separation loss simultaneously compacts known-class embeddings, reducing within-class variance as a direct consequence of its optimization objective. These two objectives are geometrically incompatible in the embedding space: VICReg pushes representations apart while MCH pulls same-class representations together. The result is a calibration landscape in which the EVT threshold τ is set lower than optimal, increasing false acceptance of unknown samples. This finding suggests that diversity regularization

for multi-signal open-set detection requires task-specific design—specifically, a formulation that preserves inter-signal diversity without counteracting the intra-class compactness that post hoc calibration depends upon.

5.5. Practical Deployment Considerations

Beyond classification accuracy, three properties are required for practical remote sensing deployment: computational efficiency, calibration stability, and adaptability to sensor degradation. The proposed framework addresses all three. At $n = 3$, inference requires 5.91 ms per sample—fast enough for real-time processing on aerial platforms operating at typical UAV flight speeds. The post hoc calibration stage is fitted once on validation data and applied without retraining, making it immediately deployable when sensor conditions change between missions. Finally, the gating mechanism explicitly down-weights unreliable modalities when LiDAR data is noisy or partially occluded, providing graceful degradation under sensor failure conditions that would cause fixed-fusion architectures to produce unpredictable outputs. Although $\delta = 0.45$ is constant across all datasets, the resulting threshold τ is fully data-adaptive, as it is derived from the empirical score distribution of known validation samples and therefore takes a different value for each dataset and validation split. The residual limitation is of a different kind: τ is computed once at training time and does not update when the deployment-time score distribution shifts due to seasonal variation, sensor aging, or geographic domain change. In such settings, the static τ may become suboptimal even though it was well-calibrated at validation time. Transductive threshold calibration [20], which estimates the deployment score distribution from unlabeled test data to update τ without unknown supervision, represents the most direct path to addressing this limitation.

6. Conclusions

Existing HSI–LiDAR fusion methods operate under the closed-set assumption, making them unsuitable for real-world deployments where novel land-cover classes routinely appear. This work presents an open-set classification framework based on multimodal Uncertainty-Aware Gating Fusion and Iterative Feedback Refinement for HSI and LiDAR, which addresses this limitation through three tightly integrated components: UAGF, IFR, and MCH-based EVT calibrator.

Experiments on Houston2013, Muufl, and Augsburg reveal that the proposed technique attains overall accuracies of 92.79%, 84.47%, and 80.99%, respectively, surpassing the nearest multimodal rival, HyLiOSR, by 32.36 pp, 17.43 pp, and 11.33 pp of unknown accuracy. Ablation results confirm that each component contributes meaningfully: removing IFR reduces unknown accuracy by 35.7 pp, while VICReg regularization degrades AUROC by 11.3 pp, demonstrating that the MCH separation loss is the important contributor of discriminative uncertainty.

Nevertheless, the proposed framework still has limitations. Unknown-class detection performance on Augsburg (56.81%) remains comparatively limited relative to Houston2013 and Muufl, suggesting that the two-signal EVT calibrator becomes more challenging when unknown classes are spectrally similar to known classes and multimodal separability is reduced. This observation indicates that robust unknown detection under highly overlapping spectral–spatial conditions remains an open challenge for multimodal open-set remote sensing.

Future work will investigate: (1) adaptive threshold selection conditioned on estimated class imbalance, (2) extension to hyperspectral–SAR fusion where modality reliability differences are more pronounced, and (3) temporal open-set recognition using multi-date image sequences where unknown classes emerge gradually.

Author Contributions: Conceptualization, D.M., A.W. and H.W.; methodology, D.M., A.W. and H.W.; software, D.M.; validation, D.M.; writing—original draft preparation, D.M.; writing—review and editing, A.W. and H.W.; visualization, D.M.; supervision, A.W. and H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Key Research and Development Plan Project of Heilongjiang (JD2023SJ19) and the National Key Support Project for Foreign Experts of Northeast Special Project (D20250098).

Data Availability Statement: Houston2013 and Augsburg: <https://drive.google.com/file/d/1UaeUWqTHhXzpwGHZEIcF8AHwmVo0IwaV/view> (accessed on 21 May 2021); Muufl: <https://github.com/GatorSense/MUUFLLGulfport> (accessed on 17 April 2017). The source code is publicly available at <https://github.com/Davaajargal-M/OpenSet-Multimodal-Fusion> (accessed on 8 May 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep Convolutional Neural Networks for Hyperspectral Image Classification. *J. Sens.* **2015**, *2015*, 258619. [\[CrossRef\]](#)
- Paoletti, M.E.; Haut, J.M.; Plaza, J.; Plaza, A. Deep learning classifiers for hyperspectral imaging: A review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 279–317. [\[CrossRef\]](#)
- Rasti, B.; Ghamisi, P.; Plaza, J.; Plaza, A. Fusion of hyperspectral and LiDAR data using sparse and low-rank component analysis. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 6354–6365. [\[CrossRef\]](#)
- Cai, J.; Zhang, M.; Yang, H.; He, Y.; Yang, Y.; Shi, C.; Zhao, X.; Xun, Y. A novel graph-attention based multimodal fusion network for joint classification of hyperspectral image and LiDAR data. *Expert Syst. Appl.* **2024**, *249*, 123587. [\[CrossRef\]](#)
- Roy, S.K.; Deria, A.; Hong, D.; Rasti, B.; Plaza, A.; Chanussot, J. Multimodal fusion transformer for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5515620. [\[CrossRef\]](#)
- Roy, S.; Deria, A.; Shah, C.; Haut, J.; Du, Q.; Plaza, A. Spectral-Spatial Morphological Attention Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5503615. [\[CrossRef\]](#)
- Sun, L.; Zhao, G.; Zheng, Y.; Wu, Z. Spectral-spatial feature tokenization transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5522214. [\[CrossRef\]](#)
- Wang, S.; Hou, C.; Chen, Y.; Liu, Z.; Zhang, Z.; Zhang, G. Classification of hyperspectral and LiDAR data using multi-modal transformer cascaded fusion net. *Remote Sens.* **2023**, *15*, 4142. [\[CrossRef\]](#)
- Liu, J.; Xue, X.; Zuo, Q.; Ren, J. Classification of hyperspectral-LiDAR dual-view data using hybrid feature and trusted decision fusion. *Remote Sens.* **2024**, *16*, 4381. [\[CrossRef\]](#)
- Pan, J.; Shuai, X.; Xu, Q.; Dai, M.; Zhang, G.; Wang, G. Classification of hyperspectral and LiDAR data by transformer-based enhancement. *Remote Sens. Lett.* **2024**, *15*, 1074–1084. [\[CrossRef\]](#)
- Cao, W.; Wang, X.; Xiao, Y.; Huang, W.; Wen, B.; Zhong, X. Uncertainty-Aware with Adaptive Geometric Correction for Multimodal Land-Cover Classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5654413. [\[CrossRef\]](#)
- Zhang, Y.; Gao, H.; Chen, Z.; Fei, S.; Zhou, J.; Ghamisi, P.; Zhang, B. Adaptive multi-stage fusion of hyperspectral and LiDAR data via selective state space models. *Inf. Fusion* **2025**, *125*, 103488. [\[CrossRef\]](#)
- Bisgin, H.; Palechor, A.; Suter, M.; Günther, M. Large-scale evaluation of open-set image classification techniques. *arXiv* **2024**, arXiv:2406.09112.
- Vaze, S.; Han, K.; Vedaldi, A.; Zisserman, A. Open-set recognition: A good closed-set classifier is all you need. In Proceedings of the International Conference on Learning Representations, Virtual Event, 25–29 April 2022.
- Scheirer, W.J.; de Rezende Rocha, A.; Sankota, A.; Boulton, T.E. Toward open set recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 1757–1772. [\[CrossRef\]](#)
- Bendale, A.; Boulton, T.E. Towards open set deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1563–1572.
- Shu, L.; Xu, H.; Liu, B. Doc: Deep open classification of text documents. *arXiv* **2017**, arXiv:1709.08716. [\[CrossRef\]](#)
- Ge, Z.; Demyanov, S.; Chen, Z.; Garnavi, R. Generative openmax for multi-class open set classification. *arXiv* **2017**, arXiv:1707.07418.
- Liu, W.; Wang, X.; Owens, J.; Li, Y. Energy-based out-of-distribution detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21464–21475.
- Zhang, Q.; An, D.; Xiao, T.; He, T.; Tang, Q.; Wu, Y.N.; Tighe, J.; Xing, Y. Learning for Transductive Threshold Calibration in Open-World Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 17097–17106.

21. Novello, P.; Prudent, Y.; Dalmau, J.; Friedrich, C.; Pequignot, Y. Improving out-of-distribution detection by combining existing post-hoc methods. *arXiv* **2024**, arXiv:2407.07135.
22. Lu, S.; Wang, Y.; Sheng, L.; He, L.; Zheng, A.; Liang, J. Out-of-distribution detection: A task-oriented survey of recent advances. *ACM Comput. Surv.* **2025**, *58*, 37. [[CrossRef](#)]
23. Li, Z.; Bi, K.; Wang, Y.; Fang, Z.; Zhang, J. Supervised contrastive learning for open-set hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 5509805. [[CrossRef](#)]
24. Qin, B.; Feng, S.; Zhao, C.; Li, W.; Tao, R.; Zhou, J. Language-enhanced dual-level contrastive learning network for open-set hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5508114. [[CrossRef](#)]
25. Tang, J.; Lei, S.; Liu, J.; Lv, N.; Qi, H. Dual-Branch Hyperspectral Open-Set Classification with Reconstruction–Prototype Fusion for Satellite IoT Perception. *Remote Sens.* **2025**, *17*, 3722. [[CrossRef](#)]
26. Xi, B.; Cai, M.; Li, J.; Wang, Z.; Feng, S.; Li, Y.; Chanussot, J. HyLiOSR: Staged progressive learning for joint open-set recognition of hyperspectral and LiDAR data. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5506714. [[CrossRef](#)]
27. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; pp. 1321–1330.
28. Hendrycks, D.; Mazeika, M.; Dietterich, T. Deep anomaly detection with outlier exposure. *arXiv* **2018**, arXiv:1812.04606.
29. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673.
30. Bardes, A.; Ponce, J.; LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv* **2021**, arXiv:2105.04906.
31. Sun, X.; Yang, Z.; Zhang, C.; Ling, K.-V.; Peng, G. Conditional gaussian distribution learning for open set recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 13480–13489.
32. Ming, Y.; Sun, Y.; Dia, O.; Li, Y. How to exploit hyperspherical embeddings for out-of-distribution detection? *arXiv* **2022**, arXiv:2203.04450.
33. Djurisic, A.; Bozanic, N.; Ashok, A.; Liu, R. Extremely simple activation shaping for out-of-distribution detection. *arXiv* **2022**, arXiv:2209.09858.
34. Chen, G.; Peng, P.; Wang, X.; Tian, Y. Adversarial reciprocal points learning for open set recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 8065–8081. [[CrossRef](#)]
35. Dimitrić, D.; Risojević, V. Training-free outlier exposure for remote sensing out-of-distribution detection. *Int. J. Data Sci. Anal.* **2025**, *21*, 28. [[CrossRef](#)]
36. Han, Z.; Zhang, C.; Fu, H.; Zhou, J.T. Trusted multi-view classification with dynamic evidential fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2551–2566. [[CrossRef](#)]
37. Ji, F.; Zhao, W.; Wang, Q.; Peng, R. A Global Distribution-Aware Network for Open-Set Hyperspectral Image Classification. *Remote Sens.* **2025**, *17*, 3938. [[CrossRef](#)]
38. Bi, K.; Li, Z.; Chen, Y.; Du, Q.; Ma, L.; Wang, Y.; Fang, Z.; Qi, M. Open-set domain adaptation for hyperspectral image classification based on weighted generative adversarial networks and dynamic thresholding. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5507717. [[CrossRef](#)]
39. Wu, H.; Yan, S.; Liu, C.; Wang, A.; Wang, M.; Yu, L. Hyperspectral Open-Set Classification via Frequency-Domain Rotation Enhancement and Multi-Branch Adversarial Routing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 24864–24882. [[CrossRef](#)]
40. Sun, H.; Chen, R.; Chen, Y.; Chen, W.; Xie, W.; Lu, X. Reconstruction-Contrast Coupling Learning for Open-Set Semi-Supervised Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2026**, *35*, 5032–5047. [[CrossRef](#)]
41. Wang, A.; Yan, S.; Wu, H.; Yu, L.; Molnar, G.; Wang, M. Open-Set Hyperspectral Image Classification via Meta-Learning and Prototype-Guided Decoding. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 14795–14819. [[CrossRef](#)]
42. Liang, S.; Li, Y.; Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv* **2017**, arXiv:1706.02690.
43. Hong, D.; Gao, L.; Yokoya, N.; Yao, J.; Chanussot, J.; Du, Q.; Zhang, B. More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4340–4354. [[CrossRef](#)]
44. Zhao, X.; Tao, R.; Li, W.; Li, H.C.; Du, Q.; Liao, W.; Philips, W. Joint Classification of Hyperspectral and LiDAR Data Using Hierarchical Random Walk and Deep CNN Architecture. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7355–7370. [[CrossRef](#)]
45. Hong, D.; Gao, L.; Yao, J.; Zhang, B.; Plaza, A.; Chanussot, J. Graph Convolutional Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5966–5978. [[CrossRef](#)]
46. Li, J.; Hu, Y.; Huang, X. CaSaFormer: A cross- and self-attention based lightweight network for large-scale building semantic segmentation. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *130*, 103942. [[CrossRef](#)]

47. Wang, X.; Feng, Y.; Song, R.; Mu, Z.; Song, C. Multi-attentive hierarchical dense fusion net for fusion classification of hyperspectral and LiDAR data. *Inf. Fusion* **2022**, *82*, 1–18. [[CrossRef](#)]
48. Li, H.; Ghamisi, P.; Soergel, U.; Zhu, X.X. Hyperspectral and LiDAR fusion using deep three-stream convolutional neural networks. *Remote Sens.* **2018**, *10*, 1649. [[CrossRef](#)]
49. Zafar, A.; Aamir, M.; Mohd Nawawi, N.; Arshad, A.; Riaz, S.; Alruban, A.; Dutta, A.K.; Almotairi, S. A comparison of pooling methods for convolutional neural networks. *Appl. Sci.* **2022**, *12*, 8643. [[CrossRef](#)]
50. Roy, S.K.; Deria, A.; Hong, D.; Ahmad, M.; Plaza, A.; Chanussot, J. Hyperspectral and LiDAR Data Classification Using Joint CNNs and Morphological Feature Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5530416. [[CrossRef](#)]
51. Rasti, B.; Ghamisi, P.; Gloaguen, R. Hyperspectral and LiDAR fusion using extinction profiles and total variation component analysis. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3997–4007. [[CrossRef](#)]
52. Fan, Y.; Qian, Y.; Gong, W.; Chu, Z.; Qin, Y.; Muhetaer, P. Multi-level interactive fusion network based on adversarial learning for fusion classification of hyperspectral and LiDAR data. *Expert Syst. Appl.* **2024**, *257*, 125132. [[CrossRef](#)]
53. Zhang, M.; Li, W.; Tao, R.; Li, H.; Du, Q. Information fusion for classification of hyperspectral and LiDAR data using IP-CNN. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5506812. [[CrossRef](#)]
54. Xu, L.; Zhao, D.; Yan, Y.; Kwong, S.; Chen, J.; Duan, L.-Y. IDeRs: Iterative dehazing method for single remote sensing image. *Inf. Sci.* **2019**, *489*, 50–62. [[CrossRef](#)]
55. Chen, S.; Zhong, S.; Xue, B.; Li, X.; Zhao, L.; Chang, C.I. Iterative Scale-Invariant Feature Transform for Remote Sensing Image Registration. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 3244–3265. [[CrossRef](#)]
56. Khoshbakht, A.; Aptoula, E. Hyperbolic Prototype Learning for Open-Set Joint HSI–LiDAR Classification. *IEEE Geosci. Remote Sens. Lett.* **2026**, *23*, 5504105. [[CrossRef](#)]
57. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A discriminative feature learning approach for deep face recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 499–515.

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.