

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/190159/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Zhang, Xinlei, Liu, Fan, Zhang, Chuanyi, Ji, Xiaoying, Wang, Wenhui, Zhou, Wei and Zhou, Jun 2026. Rethinking multilabel remote sensing image classification under dual-distribution instance-dependent noise. IEEE Transactions on Geoscience and Remote Sensing 10.1109/tgrs.2026.3740916

Publishers page: <https://doi.org/10.1109/tgrs.2026.3740916>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



Rethinking Multilabel Remote Sensing Image Classification under Dual-Distribution Instance-Dependent Noise

Xinlei Zhang, Fan Liu, *Senior Member, IEEE*, Chuanyi Zhang, *Member, IEEE*, Xiaoying Ji, Wenhui Wang, Wei Zhou, *Senior Member, IEEE*, and Jun Zhou, *Fellow, IEEE*

Abstract—Multilabel scene classification (MLC) is a fundamental task in remote sensing (RS). However, its performance is often limited by noisy annotations, where subtractive noise occurs when inconspicuous targets are missed and additive noise arises when visually confusing categories are falsely added. To simulate such annotation errors, the recently proposed Instance-Dependent Multilabel Noise (IDMN) method introduces the first instance-dependent noise generation strategy for RS MLC, driven by the intuition that hard label instances are more susceptible to mislabeling. Specifically, it leverages a single probability distribution of present labels to derive thresholds for both subtractive and additive noise generation. This design determines additive noise by requiring absent-label occurrence probabilities to exceed an irrelevant present-label threshold, leaving its quantity difficult to control and overlooking potentially confusing absent labels. To address this limitation, we propose Dual-Distribution Instance-Dependent Multilabel Noise (DD-IDMN), which explicitly decouples the score distributions of present and absent labels for calibrated noise generation. Empirical analysis under DD-IDMN reveals that additive noise yields more severe performance degradation than subtractive noise. To combat this DD-IDMN challenge, we propose the Prompt-guided Activation Interference Rectification Network (PAIR-Net). Through cross-modal attention with learnable text prompts, it learns more discriminative RS feature representations. Furthermore, we introduce an activation-aware dynamic filtering module to identify unreliable present labels for suppressing the more destructive additive noise under DD-IDMN. Experiments on UCMerced, AID, and MLRSNet demonstrate that PAIR-Net consistently outperforms representative noise-robust multilabel methods under complex instance-dependent noise. Our code is available at <https://github.com/BasilLeave/PAIR-Net>.

Index Terms—multilabel noise, multilabel scene classification, noisy labels, remote sensing.

I. INTRODUCTION

Remote sensing (RS) multilabel scene classification is an important task in RS image analysis, aiming to assign relevant semantic categories to a given RS image [1]–[3]. It supports various applications, such as urban planning [4]–[6], environmental monitoring [7]–[9], and disaster management [10], [11]. Despite its practical importance, constructing accurately

	GT:	field, grass, pavement, trees, water
	Random:	field, grass, pavement, trees, ~~water~~ , tank
	IDMN:	field, ~~grass~~ , pavement, trees, water, dock
	DD-IDMN:	field, ~~grass~~ , pavement, trees, water, dock , ship
	GT:	buildings, cars, grass, pavement, trees
	Random:	~~buildings~~ , cars, grass, pavement, trees, sea
	IDMN:	buildings, cars, grass, pavement, trees
	DD-IDMN:	buildings, cars, grass, pavement, trees, chaparral

Fig. 1. Comparison of noise injection methods, where green strikethrough and red labels denote subtractive and additive noise, respectively. Random noise flips labels independently of image content. IDMN may miss subtle hard negatives since it selects additive noise only when absent categories show stronger visual evidence than dominant present labels. In contrast, DD-IDMN mines more plausible hard negatives, including white floating objects resembling "ships" and shrub-like regions mistaken for "chaparral".

annotated multilabel datasets remains expensive, since multiple land-cover categories may coexist within a single image [12]–[14]. Consequently, real-world annotations inevitably contain noisy labels and training deep neural networks (DNNs) with such corrupted supervision can lead to severe performance degradation due to the memorization effect [15]–[17].

In multilabel classification, label noise is commonly divided into additive, subtractive, and mixed noise. Specifically, additive noise refers to absent categories incorrectly annotated as present, whereas subtractive noise denotes present categories omitted from the annotations. In practice, mixed noise is the most complex scenario, where additive and subtractive errors coexist within the corrupted dataset. Since real-world annotation errors are difficult to obtain and control, existing studies typically inject synthetic noise into clean multilabel RS datasets to evaluate noise-robust learning methods [18]–[21]. Early approaches mainly follow instance-independent corruption strategies, where label flipping is independent of image content. For example, uniform noise [18] flips labels with equal probability across predefined categories, while distribution-aware noise [19] assigns category-specific corruption probabilities according to label occurrence frequencies. Although these methods are simple, they ignore the visual characteristics of each image, limiting their ability to capture annotation errors associated with image content in RS images.

To overcome this limitation, recent studies have focused on instance-dependent noise [22]–[25], where label corruption depends on image content. In RS multilabel classification, the Instance-Dependent Multilabel Noise (IDMN) method [21] is a pioneering noise generation strategy, driven by the intuition that hard labels are more likely to be mislabeled. Benefiting

Corresponding author: Fan Liu (fanliu@hhu.edu.cn).

Xinlei Zhang, Fan Liu, Xiaoying Ji, and Wenhui Wang are with the College of Computer Science and Software Engineering, Hohai University, and the Key Laboratory of Water Big Data Technology of the Ministry of Water Resources, Nanjing 210098, China.

Chuanyi Zhang is with the College of Artificial Intelligence and Automation, Hohai University, Changzhou, 213200, China.

Wei Zhou is with the School of Computer Science and Informatics, Cardiff University, Cardiff CF24 4AG, United Kingdom.

Jun Zhou is with the School of Information and Communication Technology, Griffith University, Nathan, Queensland 4111, Australia.

from large-scale image-text pretraining, CLIP [26] has demonstrated superior zero-shot classification capability [27], [28], enabling category relevance estimation from image content for instance-dependent noise modeling. Thus, IDMN employs CLIP to compute a prediction score for each image-category pair, with a higher score indicating higher estimated occurrence probability. For subtractive noise generation, the key is to identify hard positives and flip their labels from 1 to 0. A hard positive refers to a present category that receives a low prediction score, indicating that it is visually difficult to recognize. In contrast, additive noise generation relies on hard negatives. A hard negative is an absent category with a high prediction score, indicating that it is visually confusing and more likely to be falsely annotated as present. IDMN flips these hard negatives from 0 to 1 when their scores exceed a threshold derived from confident present labels.

However, the additive noise generation mechanism in IDMN exhibits an inherent bias. Intuitively, additive errors mainly arise from visual confusion caused by deceptive localized features, rather than global category saliency. As illustrated in Fig. 1, small white objects floating on water are easily mistakenly annotated as "ship" due to local visual ambiguity, even though they are less conspicuous than the dominant true target "water". Since such localized patterns lack the complete semantic structures of prominent targets (*e.g.*, vast "water" or dense "buildings"), they naturally yield lower CLIP prediction scores than dominant targets. By forcing absent categories to surpass an irrelevant present-label threshold, IDMN may overlook these visually confusing absent labels. Crucially, given the discrepancy between the CLIP prediction distributions of present and absent labels, this cross-distribution thresholding bias makes the generated noise quantity difficult to control, potentially resulting in an unbalanced evaluation benchmark.

To address this limitation, we introduce Dual-Distribution Instance-Dependent Multilabel Noise (DD-IDMN), which decouples the prediction score distributions of present and absent labels for noise injection. Empirical evaluations reveal that additive noise yields more severe performance degradation than subtractive noise. To combat the DD-IDMN challenge, we propose the Prompt-guided Activation Interference Rectification Network (PAIR-Net). It introduces cross-modal semantic interaction with learnable text prompts to extract robust feature representations. To suppress the more destructive additive noise observed under DD-IDMN, we further develop an activation-aware dynamic filtering module. Since additive noisy labels lack visual instances, their activations tend to be diffuse and are more likely to be contained within activation regions of present categories. Motivated by this observation, we design a Spatial Containment Score (SCS) to quantify cross-class semantic interference. It measures the spatial coverage of the target activation region by the union of activation regions from other present labels. A high SCS indicates that the label lacks independent visual evidence and is likely to be additive noise. To avoid mistakenly identifying RS categories with natural spatial inclusion relationships as noise, we construct a semantic protection set with an LLM (*i.e.*, ChatGPT [29]). Although the commonly used label co-occurrence matrix [30], [31] can capture statistical label relationships, its estimates are

susceptible to noisy training annotations. In contrast, the LLM can effectively leverage superior reasoning ability [32], [33] to identify more stable physical inclusion relationships. During training, PAIR-Net dynamically corrects unreliable positive labels with high SCS to mitigate the impact of additive noise. The main contributions are summarized as:

- We identify the cross-distribution thresholding bias in IDMN and propose Dual-Distribution Instance-Dependent Multilabel Noise (DD-IDMN). By decoupling the score distributions of present and absent labels, DD-IDMN enables controllable instance-dependent noise modeling and provides a challenging benchmark for evaluating robust RS multilabel classification methods.
- Empirical analysis under DD-IDMN reveals that additive noise causes more severe degradation than subtractive noise. This finding highlights false positive suppression as a key robustness challenge beyond instance-independent noisy supervision in RS multilabel classification.
- We propose PAIR-Net, a label rectification framework for the DD-IDMN challenge. Unlike generic robust learning methods that mainly rely on loss regularization [34] or collaborative training [20], PAIR-Net exploits category-wise visual semantic evidence from activation maps to improve robustness under noisy multilabel supervision.

The remainder of this article is organized as follows. In Section II, we review studies related to our work. In Section III, we present our DD-IDMN and PAIR-Net in detail. Section IV provides comprehensive experimental results and in-depth analyses of PAIR-Net. Section V discusses the limitations. Finally, the conclusion of this work is drawn in Section VI.

II. RELATED WORK

A. Remote Sensing Scene Classification

Remote sensing scene classification has received sustained attention, with recent efforts addressing complex RS scene structures [35]–[37], scale variation [38]–[40], and multi-modal observations [41]–[43]. Notably, RASNet [44] represents each RS image as a continuous implicit function, improving adaptability to variations in scale and resolution. To capture global scene context efficiently, RSMamba [45] introduces state space modeling into RS image classification. Recently, Tri-CoMamba [46] extends Mamba-based modeling to multisource RS classification by fusing spatial, spectral, and modal information. These methods have advanced RS image classification in the single-label setting. However, RS scenes often contain multiple semantic categories within one image, making multilabel classification more challenging, especially under noisy annotations. Accordingly, we investigate robust RS multilabel scene classification under label noise.

B. Remote Sensing Multilabel Noise Injection Methods

Existing noise injection methods for RS multilabel image classification mainly follow an instance-independent setting, where label corruption is determined without image content. For instance, uniform noise [18] randomly flips labels with the same probability across images and categories. Partial uniform

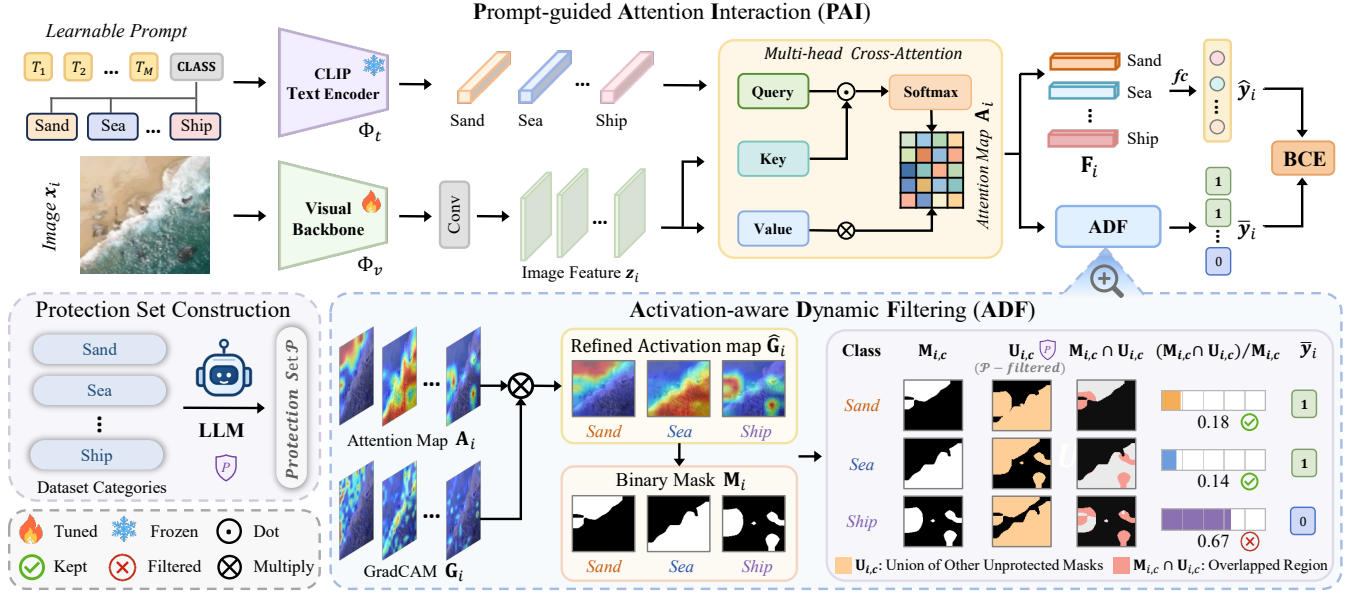


Fig. 2. Illustration of the proposed Prompt-guided Activation Interference Rectification Network (PAIR-Net). First, the Prompt-guided Attention Interaction (PAI) module exploits learnable text prompts and cross-modal attention to enhance class-specific RS representations and generate attention maps for subsequent GradCAM refinement. Then, the Activation-aware Dynamic Filtering (ADF) module identifies and rectifies potential additive noisy labels with diffuse activations covered by genuine present categories. The Protection Set Construction module protects category pairs with natural spatial inclusion relationships from mistaken filtering. Dashed modules (*i.e.*, Protection Set Construction and ADF) are employed only during training.

noise [20] applies uniform flipping only to a randomly selected subset of training images. Distribution-aware noise [19] assigns category-wise flipping probabilities according to the empirical label distribution, making frequent categories more likely to be corrupted. Although these methods are simple and controllable, they poorly capture annotation errors related to image content in RS images [47]–[49]. To alleviate this limitation, IDMN [21] introduces the first instance-dependent RS multilabel noise generation method, motivated by the intuition that hard instances are more likely to be mislabeled. However, the noise level in IDMN is not explicitly controlled, since its thresholds are derived solely from the score distribution of present labels. Therefore, we aim to generate instance-dependent noise with controllable levels for more reliable and fair evaluation of noise-robust training methods.

C. Remote Sensing Multilabel Noise-Robust Methods

In RS multilabel classification, Aksoy et al. [20] propose a collaborative framework where two networks rank training images according to a samplewise loss and exclude images with highly corrupted label sets. To investigate different types of label noise, Burgert et al. [19] categorize multilabel noise into additive, subtractive, and mixed noise. Furthermore, they adapt several single-label robust learning methods to RS multilabel classification, including SAT [50], which refines training targets according to model predictions, ELR [34], which regularizes early learning for each label, and JoCoR [51], which selects labels with small losses via joint regularization. Additionally, recent studies explore robust learning under multilabel noise [52], [53]. Nevertheless, existing methods are primarily designed and evaluated under instance-independent noise, while their behavior under instance-dependent RS multilabel noise remains insufficiently understood.

III. METHODOLOGY

In this section, we first formulate the RS MLC task under noisy supervision. Then, we present our Dual-Distribution Instance-Dependent Multilabel Noise (DD-IDMN). Finally, we introduce PAIR-Net, a dedicated cross-modal learning framework to address this challenging noise issue.

A. Problem Definition

Let $\mathcal{D} = \{(x_i, \mathbf{y}_i)\}_{i=1}^N$ denote a clean RS multilabel training dataset, where x_i is the i -th image and $\mathbf{y}_i = [y_{i,1}, \dots, y_{i,C}] \in \{0, 1\}^C$ is its clean label vector over C predefined categories. Here, $y_{i,c} = 1$ indicates that category c is present in image x_i , while $y_{i,c} = 0$ denotes its absence. In practice, only a noisy dataset $\tilde{\mathcal{D}} = \{(x_i, \tilde{\mathbf{y}}_i)\}_{i=1}^N$ is available, where $\tilde{\mathbf{y}}_i$ may contain both subtractive and additive noise. Subtractive noise occurs when an existing category is omitted from the annotation, *i.e.*, $y_{i,c} = 1 \rightarrow \tilde{y}_{i,c} = 0$, whereas additive noise occurs when an absent category is falsely annotated as present, *i.e.*, $y_{i,c} = 0 \rightarrow \tilde{y}_{i,c} = 1$. These two noise types may coexist in the same label vector, forming mixed noise with both missed real categories and fabricated false categories. The goal is to train a robust multilabel classification model from the noisy supervision $\tilde{\mathbf{y}}_i$.

B. DD-IDMN

To address the thresholding bias in IDMN, we propose the Dual-Distribution Instance-Dependent Multilabel Noise (DD-IDMN) generation mechanism. Following Distribution-aware noise [19], we set $K_\eta = \lfloor N_{\text{pos}} \times \eta \rfloor$ as the target quantity for additive and subtractive noise at a global noise rate $\eta \in (0, 1)$, where $N_{\text{pos}} = \sum_{i=1}^N \sum_{c=1}^C y_{i,c}$ is the number of clean positive labels. Mixed noise therefore contains K_η corruptions of each type, totaling $2K_\eta$ corrupted label entries.

Following IDMN [21], we employ a frozen CLIP model (*i.e.*, ViT-B/32) to compute the prediction score $s_{i,c} = \text{CLIP}(x_i, c)$ for each image and category pair. Unlike IDMN, DD-IDMN explicitly separates the score distributions of present and absent labels. Specifically, all prediction scores are partitioned by the clean labels $y_{i,c}$ into the present label score set $\mathcal{S}_{\text{pos}}$ and the absent label score set $\mathcal{S}_{\text{neg}}$:

$$\mathcal{S}_{\text{pos}} = \{s_{i,c} \mid y_{i,c} = 1\}, \quad \mathcal{S}_{\text{neg}} = \{s_{i,c} \mid y_{i,c} = 0\}, \quad (1)$$

where $i \in \{1, \dots, N\}$ denotes the training image index and $c \in \{1, \dots, C\}$ denotes the category index.

Driven by the intuition that hard label instances are more likely to be mislabeled, DD-IDMN selects visually difficult labels from the decoupled score distributions for noise injection. For present labels, lower prediction scores indicate categories that are truly present but visually inconspicuous and difficult to recognize, and thus more likely to be missed. Therefore, the lowest scoring labels in the present score set $\mathcal{S}_{\text{pos}}$ are selected as hard positives for subtractive noise. For absent labels, higher prediction scores indicate categories that are more likely to be falsely annotated as present. Accordingly, the highest scoring labels in the absent score set $\mathcal{S}_{\text{neg}}$ are selected as hard negatives for additive noise. Formally, the subtractive threshold ρ_{sub} and additive threshold ρ_{add} are defined as the K_η -th values of the sorted present and absent score sets:

$$\rho_{\text{sub}} = \mathcal{S}_{\text{pos}}^{\text{asc}}[K_\eta], \quad \rho_{\text{add}} = \mathcal{S}_{\text{neg}}^{\text{desc}}[K_\eta], \quad (2)$$

where $\mathcal{S}_{\text{pos}}^{\text{asc}}$ denotes $\mathcal{S}_{\text{pos}}$ sorted in ascending order, and $\mathcal{S}_{\text{neg}}^{\text{desc}}$ denotes $\mathcal{S}_{\text{neg}}$ sorted in descending order. Thus, present labels with scores below ρ_{sub} are selected as hard positives and flipped from 1 to 0 to generate subtractive noise, while absent labels with scores above ρ_{add} are selected as hard negatives and flipped from 0 to 1 to form additive noise. Since multiple labels may share the boundary score, we further sample labels whose scores equal the corresponding thresholds until the target quantity is exactly matched, yielding the hard positive set $\mathcal{H}_{\text{pos}}$ and hard negative set $\mathcal{H}_{\text{neg}}$ with $|\mathcal{H}_{\text{pos}}| = |\mathcal{H}_{\text{neg}}| = K_\eta$. Formally, the generated noisy label $\tilde{y}_{i,c}$ is defined as:

$$\tilde{y}_{i,c} = \begin{cases} 0, & \text{if } (i, c) \in \mathcal{H}_{\text{pos}}, \\ 1, & \text{if } (i, c) \in \mathcal{H}_{\text{neg}}, \\ y_{i,c}, & \text{otherwise.} \end{cases} \quad (3)$$

By deriving the score thresholds from the decoupled present and absent score distributions, DD-IDMN aligns the numbers of injected subtractive and additive noisy labels with the same target quantity specified by the global noise rate η . This design provides a controllable and balanced synthetic benchmark for evaluating noise-robust multilabel learning methods.

C. PAIR-Net

To address the instance-dependent noisy supervision simulated by DD-IDMN, we propose the Prompt-guided Activation Interference Rectification Network (PAIR-Net), a robust multilabel classification framework for RS. As illustrated in Fig. 2, it consists of two components: Prompt-guided Attention Interaction (PAI) and Activation-aware Dynamic Filtering (ADF). Specifically, the PAI module integrates learnable text prompts

with cross-modal attention to learn class-discriminative RS representations. During training, the ADF module refines GradCAM [54] with cross-modal attention maps and leverages the refined activations to identify unreliable present labels.

1) *Prompt-guided Attention Interaction*: To improve class discrimination under noisy supervision, we introduce a Prompt-guided Attention Interaction (PAI) module that leverages the semantic priors of a pretrained vision-language model (*i.e.*, CLIP [26]). In PAI, the frozen CLIP text encoder Φ_t produces category-specific text representations serving as semantic anchors, while a trainable RS visual encoder Φ_v extracts spatial feature maps from each input image x_i .

Specifically, for each category c , we construct a learnable prompt $\mathbf{P}_c$ by concatenating M learnable context tokens $\mathbf{T} = [T_1, \dots, T_M]$ with the category name embedding E_c :

$$\mathbf{P}_c = [T_1, \dots, T_M, E_c]. \quad (4)$$

Then, $\mathbf{P}_c$ is fed into the frozen CLIP text encoder Φ_t to obtain the RS category semantic embedding $\mathbf{t}_c$:

$$\mathbf{t}_c = \Phi_t(\mathbf{P}_c). \quad (5)$$

where $\mathbf{t}_c \in \mathbb{R}^d$ denotes the textual representation and d is the embedding dimension. Meanwhile, given an input image x_i , the visual encoder Φ_v extracts a spatial feature map $\mathbf{h}_i$:

$$\mathbf{h}_i = \Phi_v(x_i), \quad (6)$$

where $\mathbf{h}_i \in \mathbb{R}^{H \times W \times d_v}$, with H , W , and d_v denoting the spatial height, spatial width, and visual channel dimension, respectively. Since d_v may differ from the text embedding dimension d , we employ a 1×1 convolution to align the visual features with the text embedding space:

$$\mathbf{z}_i = \text{Conv}_{1 \times 1}(\mathbf{h}_i), \quad (7)$$

where $\mathbf{z}_i \in \mathbb{R}^{H \times W \times d}$ denotes the projected visual feature map. To aggregate spatial features guided by textual category semantics, we reshape $\mathbf{z}_i$ into a token sequence in $\mathbb{R}^{HW \times d}$ and perform cross-modal attention. Specifically, for each category c , the textual embedding $\mathbf{t}_c$ is projected as the query, while the visual token sequence $\mathbf{z}_i$ is projected as key and value to obtain the cross-modal representation $\mathbf{F}_{i,c} \in \mathbb{R}^d$:

$$\mathbf{F}_{i,c} = \text{Softmax}\left(\frac{Q_c K_i^\top}{\sqrt{d}}\right) V_i, \quad (8)$$

where $Q_c = \mathbf{t}_c W_q$, $K_i = \mathbf{z}_i W_k$, and $V_i = \mathbf{z}_i W_v$ denote the query, key, and value projections, respectively, with W_q , W_k , and W_v being learnable matrices. Finally, the predicted probability $\hat{y}_{i,c}$ for image x_i is computed as:

$$\hat{y}_{i,c} = \sigma(\mathbf{w}_c^\top \mathbf{F}_{i,c} + b_c), \quad (9)$$

where $\sigma(\cdot)$ is the sigmoid function, and $\mathbf{w}_c$ and b_c are the classifier weight and bias for category c , respectively. By injecting category-specific textual priors into visual representation learning, PAI improves RS class discrimination under noisy supervision. Additionally, the averaged multi-head cross-modal attention weights are reshaped into a category-specific spatial attention map $\mathbf{A}_{i,c} \in \mathbb{R}^{H \times W}$, which highlights regions relevant to category c . This attention map provides localization guidance for refining the initial GradCAM in the subsequent Activation-aware Dynamic Filtering module.

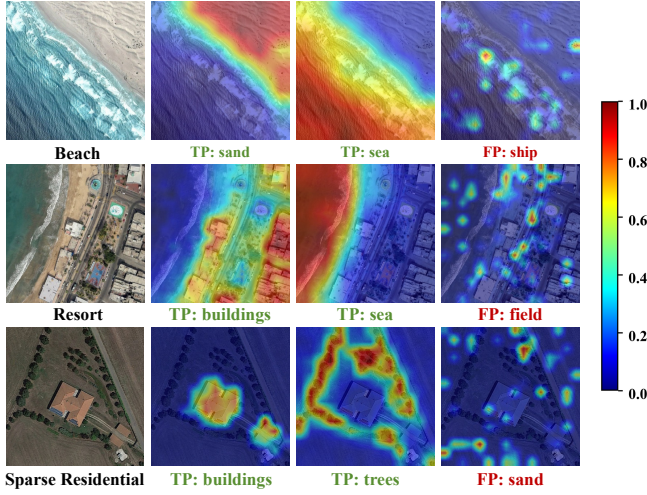


Fig. 3. Comparison of true positive (TP) and false positive (FP) activation patterns on representative AID scenes during training. TP activation maps reveal spatially coherent responses aligned with category semantics, whereas FP maps are scattered and largely overlap with TP-activated regions. The color represents the intensity of feature activation, where red indicates high activation levels and blue denotes inactive background.

2) *Activation-aware Dynamic Filtering*: Intuitively, in an ideal representation, co-occurring true targets in an RS image should preserve distinct physical boundaries. Accordingly, their activation regions are expected to be spatially coherent with limited overlap. In contrast, since false positive labels lack corresponding instances, fitting such labels tends to yield diffuse activations over unrelated regions. As illustrated in Fig. 3, we visualize the activation behavior of true positive (TP) and false positive (FP) labels during training. The TP activation maps are generally coherent and aligned with visual regions, whereas FP activations are scattered and lack an independent class-specific region. Motivated by this observation, we exploit spatial semantic interference among category activation maps to identify additive noise. Specifically, if the activation region of a positive category is largely contained within activations of other present categories, the corresponding label is more likely to be a false positive than a true label supported by independent category-aligned visual content.

To characterize this activation behavior, we compute the GradCAM $\mathbf{G}_{i,c}$ on the visual feature map $\mathbf{z}_i$. For category c and channel k , the importance weight $\alpha_{i,c}^k$ is obtained by globally averaging the gradients of the predicted probability $\hat{y}_{i,c}$ with respect to $\mathbf{z}_{i,p,q}^k$ over spatial locations (p, q) :

$$\mathbf{G}_{i,c} = \text{ReLU} \left(\sum_k \alpha_{i,c}^k \mathbf{z}_i^k \right), \alpha_{i,c}^k = \frac{1}{HW} \sum_{p=1}^H \sum_{q=1}^W \frac{\partial \hat{y}_{i,c}}{\partial \mathbf{z}_{i,p,q}^k}, \quad (10)$$

where $\hat{y}_{i,c}$ denotes the predicted probability, and $\mathbf{z}_i^k$ is the k -th channel of the projected visual feature map. To obtain more focused activation maps, we utilize the attention map $\mathbf{A}_{i,c}$ to suppress irrelevant spatial responses:

$$\hat{\mathbf{G}}_{i,c} = \mathbf{G}_{i,c} \odot \mathbf{A}_{i,c}, \quad (11)$$

where $\odot$ denotes element-wise multiplication. The refined activation map $\hat{\mathbf{G}}_{i,c}$ suppresses non-target responses and provides reliable activation localization for our dynamic noise filtering.

Algorithm 1 PAIR-Net

Require: Training dataset $\tilde{\mathcal{D}} = \{(x_i, \tilde{\mathbf{y}}_i)\}_{i=1}^N$, semantic protection set $\mathcal{P}$, network parameters Θ , warm-up epochs T_w , filtering epochs T_f , total epochs T_{total} , batch size B .

- 1: Initialize $\tilde{\mathbf{y}}_i^{(T_w+1)} \leftarrow \tilde{\mathbf{y}}_i$
- 2: **for** $t = 1, 2, \dots, T_{total}$ **do**
- 3: # Warm-up Training
- 4: **if** $t \leq T_w$ **then**
- 5: **for** $iteration = 1, 2, \dots$ **do**
- 6: Fetch $\mathcal{B} = \{(x_i, \tilde{\mathbf{y}}_i)\}_{i=1}^B$ from $\tilde{\mathcal{D}}$;
- 7: Calculate $\mathcal{L}^{(t)} = \frac{1}{|\mathcal{B}|} \sum \ell^{(t)}(x_i, \tilde{\mathbf{y}}_i)$; $\triangleright$ Eq. (16)
- 8: Update Θ by optimizing $\mathcal{L}^{(t)}$.
- 9: **end for**
- 10: **end if**
- 11: # Robust Training
- 12: **if** $T_w < t \leq T_{total}$ **then**
- 13: **for** $iteration = 1, 2, \dots$ **do**
- 14: Fetch $\mathcal{B} = \{(x_i, \tilde{\mathbf{y}}_i^{(t)})\}_{i=1}^B$;
- 15: # Prompt-guided Attention Interaction
- 16: Obtain cross-modal feature $\mathbf{F}_i$ $\triangleright$ Eq. (8)
- 17: **if** $t \leq T_w + T_f$ **then**
- 18: # Activation-aware Dynamic Filtering
- 19: Obtain cross-modal attention map $\mathbf{A}_i$;
- 20: Refine GradCAM $\mathbf{G}_i$ with $\mathbf{A}_i$; $\triangleright$ Eq. (11)
- 21: Generate binary mask $\mathbf{M}_i$; $\triangleright$ Eq. (12)
- 22: Calculate $\mathbf{SCS}_i$; $\triangleright$ Eq. (14)
- 23: Obtain rectified labels $\tilde{\mathbf{y}}_i^{(t+1)}$; $\triangleright$ Eq. (15)
- 24: **else**
- 25: Retain current labels: $\tilde{\mathbf{y}}_i^{(t+1)} \leftarrow \tilde{\mathbf{y}}_i^{(t)}$;
- 26: **end if**
- 27: Calculate $\mathcal{L}^{(t)} = \frac{1}{|\mathcal{B}|} \sum \ell^{(t)}(x_i, \tilde{\mathbf{y}}_i^{(t)})$; $\triangleright$ Eq. (16)
- 28: Update Θ by optimizing $\mathcal{L}^{(t)}$.
- 29: **end for**
- 30: **end if**
- 31: **end for**

Then, the refined activation map $\hat{\mathbf{G}}_{i,c}$ is binarized by Otsu's thresholding [55], which automatically determines the threshold $\theta_{i,c}$ for each category-specific activation map. Formally, the binary mask $\mathbf{M}_{i,c}$ at each spatial location (p, q) follows:

$$\mathbf{M}_{i,c}(p, q) = \begin{cases} 1, & \hat{\mathbf{G}}_{i,c}(p, q) \geq \theta_{i,c}, \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

To quantify cross-class semantic interference among positive labels, we propose the Spatial Containment Score (SCS), which measures how much the activation of target category c is covered by other present categories in image x_i . A high SCS indicates target activation largely contained within other category activations, suggesting weak class discrimination and a higher likelihood of additive noise. Conversely, a low SCS indicates distinct target activation and category focusing, suggesting that the label is more likely to be a true positive.

Since geographical entities in RS imagery naturally exhibit spatial inclusion relationships, directly treating high SCS values as noisy labels may suppress hard positives. For example, airplanes often appear on pavement, and ships are located

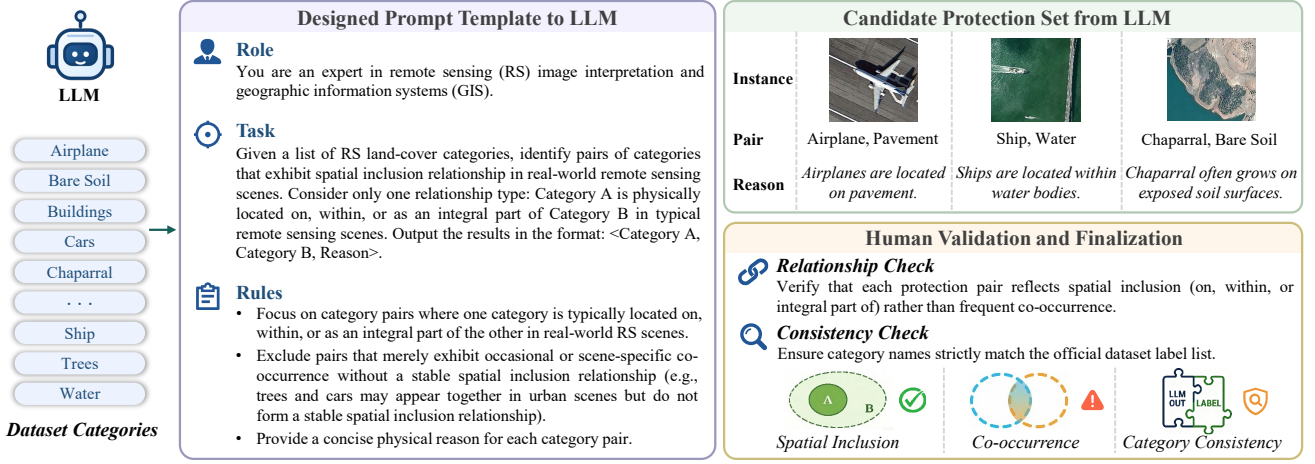


Fig. 4. Construction pipeline of the RS-specific semantic protection set $\mathcal{P}$. Specifically, we utilize an LLM (*i.e.*, ChatGPT powered by GPT-5.5 [29]) to identify category pairs with stable physical inclusion relationships from the predefined dataset categories. The generated candidates are then manually verified through a *Relationship Check*, which distinguishes genuine physical inclusion (*e.g.*, Ship–Water) from frequent co-occurrence without a stable spatial inclusion relation (*e.g.*, Sand–Sea), and a *Consistency Check*, which maps all category names to the official dataset labels.

within water. Such pairs may produce large activation overlaps even when both labels are clean. To prevent valid spatial overlap from being treated as noisy interference, we construct a dataset-specific semantic protection set $\mathcal{P}$ using an LLM (*i.e.*, ChatGPT powered by GPT-5.5 [29]). As described in Fig. 4, the construction pipeline combines a prompt template with human validation to verify inclusion relationships and category names. Although physical inclusion relationships are inherently directional, we define $\mathcal{P}$ as a symmetric unordered set since our protection mechanism removes valid activation overlap between category pairs from the SCS calculation. Accordingly, for target present class c in image x_i , we define the epoch-dependent unprotected set $\Omega_{i,c}^{(t)}$ by excluding protected pairs with c from current present labels $\bar{y}_{i,k}^{(t)}$:

$$\Omega_{i,c}^{(t)} = \{k \mid \bar{y}_{i,k}^{(t)} = 1, k \neq c, \{c, k\} \notin \mathcal{P}\}, \quad (13)$$

where $\bar{y}_{i,k}^{(t)}$ denotes the current label state at epoch t , with $\bar{y}_{i,k}^{(t)} = 1$ indicating that category k is annotated as present for image x_i . Formally, at training epoch t , let $\mathbf{M}_{i,c}^{(t)}$ denote the binary activation mask of the target positive category c , and let $\mathbf{U}_{i,c}^{(t)}$ denote the union mask of the corresponding unprotected present categories in $\Omega_{i,c}^{(t)}$ under the current label state. The SCS is defined as the fraction of target mask pixels in $\mathbf{M}_{i,c}^{(t)}$ that overlap with the union mask $\mathbf{U}_{i,c}^{(t)}$:

$$\text{SCS}_{i,c}^{(t)} = \frac{|\mathbf{M}_{i,c}^{(t)} \cap \mathbf{U}_{i,c}^{(t)}|}{|\mathbf{M}_{i,c}^{(t)}|}, \quad \mathbf{U}_{i,c}^{(t)} = \bigcup_{k \in \Omega_{i,c}^{(t)}} \mathbf{M}_{i,k}^{(t)}, \quad (14)$$

where $\text{SCS}_{i,c}^{(t)} = 0$ if $|\mathbf{M}_{i,c}^{(t)}| = 0$ or $\Omega_{i,c}^{(t)} = \emptyset$. Finally, we rectify potential additive noisy labels with a predefined filtering threshold τ , yielding the updated label $\bar{y}_{i,c}^{(t+1)}$:

$$\bar{y}_{i,c}^{(t+1)} = \begin{cases} 0, & \bar{y}_{i,c}^{(t)} = 1 \text{ and } \text{SCS}_{i,c}^{(t)} > \tau, \\ \bar{y}_{i,c}^{(t)}, & \text{otherwise.} \end{cases} \quad (15)$$

After T_w warm-up epochs, the label state is initialized with the observed noisy labels as $\bar{y}_{i,c}^{(T_w+1)} = \tilde{y}_{i,c}$. During the

filtering stage at epoch t , we update the rectified label vector $\bar{\mathbf{y}}_i^{(t+1)}$. Only positive labels identified as unreliable are flipped from 1 to 0, while all other labels remain unchanged. Thus, the proposed Activation-aware Dynamic Filtering (ADF) module progressively suppresses additive noisy labels during training, mitigating the severe degradation caused by additive noise, as empirically demonstrated in Fig. 5.

3) *Optimization Objective*: At each training epoch t , PAIR-Net is supervised by the current label vector $\bar{\mathbf{y}}_i^{(t)}$. Given a mini-batch of B images and C predefined categories, the Binary Cross-Entropy (BCE) objective is formulated as:

$$\mathcal{L}^{(t)} = -\frac{1}{B} \sum_{i=1}^B \sum_{c=1}^C [\bar{y}_{i,c}^{(t)} \log \hat{y}_{i,c}^{(t)} + (1 - \bar{y}_{i,c}^{(t)}) \log (1 - \hat{y}_{i,c}^{(t)})]. \quad (16)$$

To present an intuitive understanding of our PAIR-Net, the overall procedure is summarized in Algorithm 1.

IV. EXPERIMENT

In this section, we present the experimental settings, including datasets, evaluation metrics, implementation details, and compared methods. We then compare DD-IDMN with different noise-injection strategies and evaluate PAIR-Net against noise-robust learning methods. Finally, comprehensive ablation studies are conducted to systematically analyze the effectiveness and characteristics of the proposed approach.

A. Datasets

We conduct experiments on three multilabel RS image datasets: UCMerced [56], AID [57], and MLRSNet [58]. UCMerced and AID were originally designed for single-label scene classification, and we adopt their public multilabel annotations from previous studies [59], [60]. MLRSNet is a large-scale RS dataset with native multilabel annotations. These datasets vary in image scale and scene complexity, enabling a comprehensive evaluation across multilabel scenarios.

UCMerced [56]. The original dataset contains 2,100 aerial images from the United States Geological Survey National

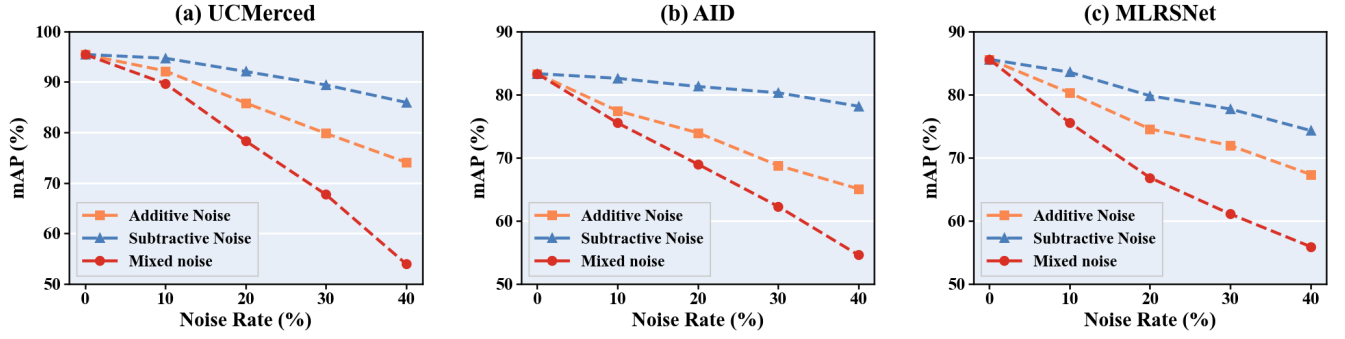


Fig. 5. Classification performance of different noise types under DD-IDMN across varying noise rates on the UCMerced, AID, and MLRSNet datasets. Under the DD-IDMN noise setting, additive noise leads to more severe performance degradation than subtractive noise across all evaluated datasets. Mixed noise yields the lowest performance, containing both K_η additive and K_η subtractive corruptions at noise rate η .

TABLE I

COMPARISON OF MULTILABEL CLASSIFICATION PERFORMANCE ACROSS SYNTHETIC MULTILABEL NOISE GENERATION METHODS AND NOISE LEVELS USING BCE AS THE BASELINE METHOD. MIX- η DENOTES THE TARGET NOISE QUANTITY K_η FOR BOTH ADDITIVE AND SUBTRACTIVE CORRUPTION. RESULTS ARE REPORTED AS THE MEAN $\pm$ STD OVER 3 RANDOM RUNS WITH SEEDS IN (0, 1, 2).

Datasets	Noise Model	Mix-10%		Mix-20%		Mix-30%		Mix-40%	
		mAP $\uparrow$ (%)	Ham. $\downarrow$ (%)	mAP $\uparrow$ (%)	Ham. $\downarrow$ (%)	mAP $\uparrow$ (%)	Ham. $\downarrow$ (%)	mAP $\uparrow$ (%)	Ham. $\downarrow$ (%)
UCM	Uniform	91.97 $\pm$ 0.80	6.68 $\pm$ 0.19	88.44 $\pm$ 0.37	7.96 $\pm$ 0.03	82.57 $\pm$ 1.61	9.80 $\pm$ 0.22	75.05 $\pm$ 0.11	11.66 $\pm$ 0.25
	Partial uniform	91.93 $\pm$ 0.97	5.82 $\pm$ 0.25	90.38 $\pm$ 0.85	6.37 $\pm$ 0.01	87.17 $\pm$ 1.29	7.57 $\pm$ 0.50	85.26 $\pm$ 0.58	8.54 $\pm$ 0.19
	Distribution-aware	92.35 $\pm$ 0.32	7.02 $\pm$ 0.30	89.14 $\pm$ 1.01	9.21 $\pm$ 0.41	85.01 $\pm$ 0.88	11.74 $\pm$ 0.43	77.68 $\pm$ 0.11	15.44 $\pm$ 0.30
	IDMN	90.80 $\pm$ 0.38	7.61 $\pm$ 0.33	83.66 $\pm$ 0.79	10.17 $\pm$ 0.13	70.60 $\pm$ 0.88	13.64 $\pm$ 0.36	58.74 $\pm$ 1.33	18.06 $\pm$ 0.49
	DD-IDMN (ours)	89.65 $\pm$ 0.15	8.10 $\pm$ 0.09	78.31 $\pm$ 0.98	11.46 $\pm$ 0.12	67.71 $\pm$ 0.73	14.65 $\pm$ 0.19	53.96 $\pm$ 0.70	17.65 $\pm$ 0.18
AID	Uniform	80.19 $\pm$ 0.28	6.81 $\pm$ 0.10	75.10 $\pm$ 0.30	8.54 $\pm$ 0.13	66.77 $\pm$ 0.96	11.60 $\pm$ 0.27	55.91 $\pm$ 0.71	16.36 $\pm$ 0.52
	Partial uniform	80.58 $\pm$ 0.57	6.19 $\pm$ 0.04	77.74 $\pm$ 0.93	6.45 $\pm$ 0.16	75.08 $\pm$ 0.71	7.32 $\pm$ 0.11	72.39 $\pm$ 0.52	8.25 $\pm$ 0.46
	Distribution-aware	80.43 $\pm$ 0.61	8.23 $\pm$ 0.37	73.63 $\pm$ 0.38	11.79 $\pm$ 0.26	65.03 $\pm$ 0.20	16.08 $\pm$ 0.47	57.88 $\pm$ 0.86	20.59 $\pm$ 0.16
	IDMN	76.34 $\pm$ 0.49	9.30 $\pm$ 0.04	69.88 $\pm$ 0.07	13.34 $\pm$ 0.05	63.90 $\pm$ 0.36	17.96 $\pm$ 0.17	55.94 $\pm$ 0.26	23.53 $\pm$ 0.06
	DD-IDMN (ours)	75.85 $\pm$ 0.61	9.41 $\pm$ 0.15	68.96 $\pm$ 0.09	14.44 $\pm$ 0.28	62.29 $\pm$ 0.90	19.48 $\pm$ 0.14	54.74 $\pm$ 0.43	24.37 $\pm$ 0.09
MLRSNet	Uniform	80.31 $\pm$ 0.42	4.22 $\pm$ 0.04	76.91 $\pm$ 0.35	4.88 $\pm$ 0.05	72.00 $\pm$ 0.36	5.77 $\pm$ 0.03	65.26 $\pm$ 0.25	6.71 $\pm$ 0.01
	Partial uniform	77.23 $\pm$ 0.73	3.96 $\pm$ 0.03	72.55 $\pm$ 0.09	4.48 $\pm$ 0.05	68.79 $\pm$ 0.36	5.46 $\pm$ 0.03	65.65 $\pm$ 0.38	7.29 $\pm$ 0.16
	Distribution-aware	80.80 $\pm$ 0.47	4.44 $\pm$ 0.07	78.08 $\pm$ 0.04	5.37 $\pm$ 0.03	73.92 $\pm$ 0.14	6.50 $\pm$ 0.06	68.09 $\pm$ 0.40	7.89 $\pm$ 0.02
	IDMN	73.73 $\pm$ 0.07	4.88 $\pm$ 0.03	63.29 $\pm$ 0.12	6.83 $\pm$ 0.04	54.33 $\pm$ 0.13	9.60 $\pm$ 0.01	45.56 $\pm$ 0.12	13.60 $\pm$ 0.10
	DD-IDMN (ours)	75.53 $\pm$ 0.20	4.68 $\pm$ 0.02	66.84 $\pm$ 0.19	5.98 $\pm$ 0.01	61.09 $\pm$ 0.22	7.19 $\pm$ 0.02	55.92 $\pm$ 0.23	8.54 $\pm$ 0.05

Map, each with 256×256 pixels at 0.3 m resolution. We adopt the manually interpreted 17-category multilabel annotations from [59]. It is split into 1,680 training and 420 test images.

AID [57]. It contains 10,000 Google Earth images of 600×600 pixels at 0.5 m to 8 m resolution. Following previous studies, we use the AID multilabel subset from [60], where 3,000 images are annotated with 17 categories. This multilabel subset is split into 2,400 training and 600 test images.

MLRSNet [58]. It is a large-scale multilabel remote sensing image dataset collected from Google Earth. It contains 109,161 images of 256×256 pixels, with pixel resolutions ranging from 0.1 m to 10 m. Each image is annotated with multiple labels from 60 categories, making MLRSNet more challenging due to its larger data scale and richer category space. The dataset is split into 87,328 training and 21,833 test images.

B. Evaluation Metrics

Following IDMN [21], we report macro-averaged mean average precision (mAP) and Hamming loss (Ham.) metrics. All results are reported as mean $\pm$ std over three random runs under identical settings, with seeds in (0, 1, 2).

C. Implementation Details

Following IDMN [21], ResNet-50 initialized with ImageNet weights is adopted as the visual backbone Φ_v for all comparisons. The network is optimized with Adam [61] for 50 epochs, with an initial learning rate of 1×10^{-4} . All models are trained on a single NVIDIA RTX 3090 GPU with a batch size of 16 across all experiments. Standard data augmentation consists of random horizontal flipping with a probability of 0.5. DD-IDMN uses frozen CLIP ViT-B/32 for noise generation, while PAIR-Net employs a ViT-L/14 text encoder with an embedding dimension of 768 during training.

D. Compared Methods.

We compare DD-IDMN with various noise strategies and evaluate PAIR-Net against representative robust methods.

Different Noise Injection Strategies. We compare DD-IDMN with four multilabel noise strategies: uniform noise [18], partial uniform noise [20], distribution-aware noise [19], and IDMN [21]. Given a noise rate η , uniform noise randomly samples label entries, whereas partial uniform noise

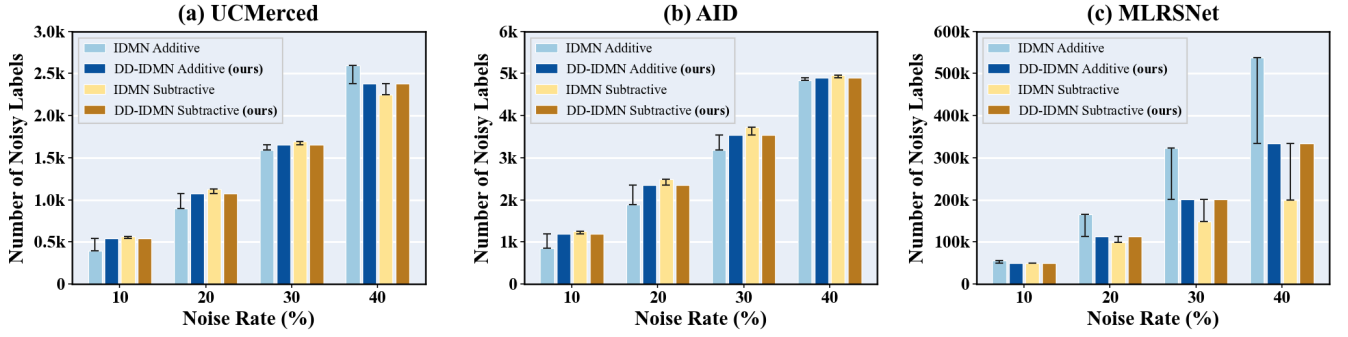


Fig. 6. Comparison of actual noisy label quantities between IDMN and our DD-IDMN across different datasets and noise rates. Following [19], the target quantity is defined as the number of present labels multiplied by the noise rate η . The vertical black bars denote deviations from the target quantity. IDMN cannot precisely control the numbers of additive and subtractive noisy labels, especially on MLRSNet. In contrast, DD-IDMN accurately matches the target quantity for both noise types, enabling a more controllable and balanced synthetic evaluation benchmark.

TABLE II

COMPARISON WITH NOISE-ROBUST METHODS UNDER MIXED DD-IDMN NOISE ON UCMERGED, AID, AND MLRSNET. MIX- η DENOTES THE TARGET NOISE QUANTITY K_η FOR BOTH ADDITIVE AND SUBTRACTIVE CORRUPTION. RESULTS ARE MEAN $\pm$ STD OVER 3 RUNS WITH SEEDS (0, 1, 2).

Datasets	Algorithms	Mix-10%		Mix-20%		Mix-30%		Mix-40%	
		mAP $\uparrow$ (%)	Ham. $\downarrow$ (%)	mAP $\uparrow$ (%)	Ham. $\downarrow$ (%)	mAP $\uparrow$ (%)	Ham. $\downarrow$ (%)	mAP $\uparrow$ (%)	Ham. $\downarrow$ (%)
UCM	BCE	89.65 $\pm$ 0.15	8.10 $\pm$ 0.09	78.31 $\pm$ 0.98	11.46 $\pm$ 0.12	67.71 $\pm$ 0.73	14.65 $\pm$ 0.19	53.96 $\pm$ 0.70	17.65 $\pm$ 0.18
	LCR	90.64 $\pm$ 0.45	7.90 $\pm$ 0.14	77.51 $\pm$ 0.67	11.37 $\pm$ 0.25	67.99 $\pm$ 0.59	14.30 $\pm$ 0.09	54.93 $\pm$ 0.74	17.50 $\pm$ 0.09
	RCML	90.60 $\pm$ 0.22	7.27$\pm$0.13	78.73 $\pm$ 0.83	10.80 $\pm$ 0.23	68.89 $\pm$ 1.50	13.67 $\pm$ 0.26	56.09 $\pm$ 1.35	17.32 $\pm$ 0.13
	SAT-ML	92.56 $\pm$ 0.13	7.61 $\pm$ 0.10	83.75 $\pm$ 0.32	10.69 $\pm$ 0.16	73.97 $\pm$ 0.09	13.12 $\pm$ 0.06	62.40 $\pm$ 0.33	16.12 $\pm$ 0.09
	ELR-ML	89.90 $\pm$ 0.55	8.00 $\pm$ 0.14	77.92 $\pm$ 1.34	11.33 $\pm$ 0.14	67.73 $\pm$ 0.45	13.12 $\pm$ 0.06	54.24 $\pm$ 0.45	17.51 $\pm$ 0.16
	JoCoR-ML	89.38 $\pm$ 1.50	7.39 $\pm$ 0.33	78.58 $\pm$ 0.55	10.75 $\pm$ 0.17	70.24 $\pm$ 0.75	12.95 $\pm$ 0.20	59.13 $\pm$ 0.53	16.40 $\pm$ 0.40
	PAIR-Net (ours)	93.34$\pm$0.70	7.36 $\pm$ 0.15	87.42$\pm$0.48	10.16$\pm$0.09	81.27$\pm$0.29	11.96$\pm$0.29	75.53$\pm$0.60	14.41$\pm$0.06
AID	BCE	75.85 $\pm$ 0.61	9.41 $\pm$ 0.15	68.96 $\pm$ 0.09	14.44 $\pm$ 0.28	62.29 $\pm$ 0.90	19.48 $\pm$ 0.14	54.74 $\pm$ 0.43	24.37 $\pm$ 0.09
	LCR	73.24 $\pm$ 0.21	9.39 $\pm$ 0.04	66.88 $\pm$ 0.84	13.96 $\pm$ 0.14	59.97 $\pm$ 0.42	18.96 $\pm$ 0.16	53.85 $\pm$ 0.14	23.76 $\pm$ 0.04
	RCML	76.57 $\pm$ 0.41	9.26 $\pm$ 0.11	70.51 $\pm$ 0.26	14.08 $\pm$ 0.11	63.16 $\pm$ 0.18	19.54 $\pm$ 0.30	56.27 $\pm$ 0.36	24.25 $\pm$ 0.11
	SAT-ML	77.23 $\pm$ 0.45	8.83$\pm$0.08	70.80 $\pm$ 0.24	13.55 $\pm$ 0.11	64.44 $\pm$ 0.36	18.46 $\pm$ 0.12	57.03 $\pm$ 0.25	23.56 $\pm$ 0.18
	ELR-ML	75.75 $\pm$ 0.65	9.43 $\pm$ 0.14	68.94 $\pm$ 0.34	14.49 $\pm$ 0.07	61.63 $\pm$ 0.85	19.49 $\pm$ 0.18	55.11 $\pm$ 0.39	24.38 $\pm$ 0.04
	JoCoR-ML	75.23 $\pm$ 0.46	8.98 $\pm$ 0.17	69.72 $\pm$ 0.43	13.48 $\pm$ 0.31	63.13 $\pm$ 0.44	18.39 $\pm$ 0.23	56.17 $\pm$ 0.14	23.06 $\pm$ 0.20
	PAIR-Net (ours)	78.43$\pm$0.67	8.85 $\pm$ 0.13	75.18$\pm$0.82	12.46$\pm$0.12	68.72$\pm$0.54	15.90$\pm$0.25	63.91$\pm$0.44	19.45$\pm$0.07
MLRSNet	BCE	75.53 $\pm$ 0.20	4.68 $\pm$ 0.02	66.84 $\pm$ 0.19	5.98 $\pm$ 0.01	61.09 $\pm$ 0.22	7.19 $\pm$ 0.02	55.92 $\pm$ 0.23	8.54 $\pm$ 0.05
	LCR	75.43 $\pm$ 0.37	4.70 $\pm$ 0.03	66.82 $\pm$ 0.21	5.97 $\pm$ 0.01	61.03 $\pm$ 0.06	7.20 $\pm$ 0.03	55.61 $\pm$ 0.22	8.49 $\pm$ 0.04
	RCML	74.76 $\pm$ 0.56	4.87 $\pm$ 0.02	65.74 $\pm$ 0.38	5.96 $\pm$ 0.02	62.39 $\pm$ 0.22	7.00 $\pm$ 0.01	57.62 $\pm$ 0.01	8.18 $\pm$ 0.02
	SAT-ML	80.79 $\pm$ 0.30	4.31 $\pm$ 0.02	72.08 $\pm$ 0.19	5.50 $\pm$ 0.01	65.84 $\pm$ 0.03	6.68 $\pm$ 0.04	60.33 $\pm$ 0.11	7.84 $\pm$ 0.01
	ELR-ML	75.54 $\pm$ 0.05	4.69 $\pm$ 0.03	66.96 $\pm$ 0.17	5.97 $\pm$ 0.01	61.30 $\pm$ 0.27	7.19 $\pm$ 0.03	55.59 $\pm$ 0.11	8.51 $\pm$ 0.05
	JoCoR-ML	78.56 $\pm$ 0.23	4.47 $\pm$ 0.02	69.28 $\pm$ 0.32	5.48 $\pm$ 0.01	60.66 $\pm$ 0.01	6.55 $\pm$ 0.00	53.96 $\pm$ 0.07	7.65 $\pm$ 0.02
	PAIR-Net (ours)	82.70$\pm$0.30	4.10$\pm$0.01	75.93$\pm$0.94	5.16$\pm$0.05	68.61$\pm$0.25	6.26$\pm$0.01	62.91$\pm$0.52	7.46$\pm$0.01

restricts sampling to selected training images. Distribution-aware noise samples labels class-wise according to positive-label counts. IDMN uses the CLIP score distribution of present labels to select low-score positives and high-score negatives. For fair comparison, we adjust the instance-independent methods to match DD-IDMN in the target quantities of additive and subtractive corruptions at each noise rate while retaining their sampling strategies. Since IDMN does not explicitly control its noise quantity, we report the actual number of corruptions.

Noise-Robust Learning Methods. Following IDMN [21], we compare PAIR-Net with standard BCE, label correlation regularization (LCR [18]), robust collaborative multilabel learning (RCML [20]), self-adaptive training for MLC (SAT-ML [19]), early learning regularization (ELR-ML [19]), and joint training with co-regularization (JoCoR-ML [19]).

E. Effect of Different Noise Types under DD-IDMN

We investigate the effects of different noise types under DD-IDMN. As illustrated in Fig. 5, mixed noise consistently yields the lowest mAP across all three datasets. It contains both additive noise and subtractive noise, making it the most complex and challenging noise setting.

Additionally, Fig. 5 shows a clear gap between additive and subtractive noise under the DD-IDMN noise setting, with additive noise yielding greater performance degradation than subtractive noise. In DD-IDMN, additive noise is produced by flipping visually confusing hard negative labels to positive ones. Once these labels are assigned as positive labels, the model is explicitly encouraged to associate image regions with incorrect categories, which directly introduces misleading semantic supervision during training. In contrast, subtractive

TABLE III

COMPARISON WITH DIFFERENT ROBUST LEARNING METHODS UNDER ADDITIVE AND SUBTRACTIVE DD-IDMN NOISE ON THE UCMERGED (UCM), AID, AND MLRSNET DATASETS. RESULTS ARE REPORTED AS THE MEAN $\pm$ STD OVER 3 RANDOM RUNS WITH SEEDS IN (0, 1, 2).

Datasets	Algorithms	Additive-20%		Additive-40%		Subtractive-20%		Subtractive-40%	
		mAP $\uparrow$ (%)	Ham. $\downarrow$ (%)	mAP $\uparrow$ (%)	Ham. $\downarrow$ (%)	mAP $\uparrow$ (%)	Ham. $\downarrow$ (%)	mAP $\uparrow$ (%)	Ham. $\downarrow$ (%)
UCM	BCE	85.82 $\pm$ 0.38	8.34 $\pm$ 0.07	74.07 $\pm$ 0.83	12.53 $\pm$ 0.21	91.56 $\pm$ 0.03	7.61 $\pm$ 0.09	85.44 $\pm$ 0.32	10.33 $\pm$ 0.30
	LCR	85.39 $\pm$ 0.66	8.32 $\pm$ 0.09	75.78 $\pm$ 0.52	12.68 $\pm$ 0.04	91.87 $\pm$ 0.20	7.39 $\pm$ 0.02	85.99 $\pm$ 0.56	9.80$\pm$0.03
	RCML	85.71 $\pm$ 0.93	8.22 $\pm$ 0.18	77.11 $\pm$ 1.07	11.80 $\pm$ 0.22	91.06 $\pm$ 0.59	7.89 $\pm$ 0.21	87.36 $\pm$ 0.61	10.47 $\pm$ 0.29
	SAT-ML	90.06 $\pm$ 0.26	7.65 $\pm$ 0.08	81.91 $\pm$ 0.16	11.08 $\pm$ 0.04	92.87 $\pm$ 0.05	7.63 $\pm$ 0.05	88.41 $\pm$ 0.06	10.44 $\pm$ 0.03
	ELR-ML	84.83 $\pm$ 0.44	8.67 $\pm$ 0.01	74.84 $\pm$ 0.04	13.05 $\pm$ 0.04	91.85 $\pm$ 0.15	7.45 $\pm$ 0.08	86.26 $\pm$ 0.22	9.81 $\pm$ 0.05
	JoCoR-ML	87.09 $\pm$ 0.56	7.38$\pm$0.10	77.47 $\pm$ 0.76	10.33 $\pm$ 0.28	89.11 $\pm$ 0.02	8.44 $\pm$ 0.22	80.51 $\pm$ 0.61	11.02 $\pm$ 0.08
	PAIR-Net (ours)	91.27$\pm$0.32	7.64 $\pm$ 0.29	87.37$\pm$0.50	9.67$\pm$0.89	93.23$\pm$0.29	7.01$\pm$0.58	89.87$\pm$0.14	10.09 $\pm$ 0.01
AID	BCE	73.95 $\pm$ 0.45	9.83 $\pm$ 0.03	65.07 $\pm$ 0.28	14.37 $\pm$ 0.10	81.27 $\pm$ 0.33	9.67 $\pm$ 0.40	78.15 $\pm$ 0.20	15.03 $\pm$ 0.19
	LCR	73.19 $\pm$ 0.22	9.64 $\pm$ 0.05	65.17 $\pm$ 0.24	14.24 $\pm$ 0.08	81.62 $\pm$ 0.07	9.91 $\pm$ 0.14	77.85 $\pm$ 0.10	14.98 $\pm$ 0.06
	RCML	75.51 $\pm$ 0.65	9.38 $\pm$ 0.04	65.99 $\pm$ 0.54	14.41 $\pm$ 0.08	81.11 $\pm$ 0.19	10.07 $\pm$ 0.08	77.08 $\pm$ 0.25	14.84 $\pm$ 0.24
	SAT-ML	74.84 $\pm$ 0.16	8.59 $\pm$ 0.06	68.15 $\pm$ 0.20	13.25 $\pm$ 0.14	82.31 $\pm$ 0.09	10.29 $\pm$ 0.44	78.75 $\pm$ 0.07	15.44 $\pm$ 0.09
	ELR-ML	73.91 $\pm$ 0.30	9.85 $\pm$ 0.10	65.05 $\pm$ 0.09	14.46 $\pm$ 0.10	81.49 $\pm$ 0.29	10.47 $\pm$ 0.09	77.75 $\pm$ 0.07	15.11 $\pm$ 0.02
	JoCoR-ML	75.08 $\pm$ 0.65	8.95 $\pm$ 0.02	66.19 $\pm$ 0.34	13.40 $\pm$ 0.19	80.89 $\pm$ 0.09	10.08 $\pm$ 0.07	76.90 $\pm$ 0.17	15.38 $\pm$ 0.08
	PAIR-Net (ours)	77.61$\pm$0.28	8.21$\pm$0.06	73.54$\pm$0.57	10.63$\pm$0.58	83.82$\pm$0.25	9.44$\pm$0.18	79.43$\pm$0.31	14.67$\pm$0.43
MLRSNet	BCE	75.78 $\pm$ 0.24	4.73 $\pm$ 0.03	68.87 $\pm$ 0.43	6.13 $\pm$ 0.10	79.69 $\pm$ 0.11	4.75 $\pm$ 0.01	73.21 $\pm$ 0.17	6.06 $\pm$ 0.04
	LCR	74.88 $\pm$ 0.25	4.60 $\pm$ 0.10	68.30 $\pm$ 0.14	6.00 $\pm$ 0.01	79.84 $\pm$ 0.02	4.73 $\pm$ 0.01	73.16 $\pm$ 0.05	6.12 $\pm$ 0.02
	RCML	74.30 $\pm$ 0.03	4.80 $\pm$ 0.02	68.68 $\pm$ 0.12	5.89 $\pm$ 0.02	76.58 $\pm$ 0.28	4.93 $\pm$ 0.01	71.14 $\pm$ 0.01	6.22 $\pm$ 0.01
	SAT-ML	81.43 $\pm$ 0.10	4.06 $\pm$ 0.01	75.89 $\pm$ 0.09	5.01 $\pm$ 0.02	85.40 $\pm$ 0.02	4.62 $\pm$ 0.02	79.92 $\pm$ 0.02	6.10 $\pm$ 0.01
	ELR-ML	75.33 $\pm$ 0.30	4.69 $\pm$ 0.01	68.14 $\pm$ 0.07	5.99 $\pm$ 0.03	79.72 $\pm$ 0.04	4.73 $\pm$ 0.01	73.13 $\pm$ 0.02	6.12 $\pm$ 0.01
	JoCoR-ML	78.19 $\pm$ 0.55	4.14 $\pm$ 0.03	71.32 $\pm$ 0.38	4.67 $\pm$ 0.07	75.70 $\pm$ 0.30	5.26 $\pm$ 0.03	64.62 $\pm$ 0.45	7.03 $\pm$ 0.04
	PAIR-Net (ours)	82.45$\pm$0.22	3.77$\pm$0.01	78.79$\pm$0.09	4.60$\pm$0.01	87.91$\pm$0.35	4.07$\pm$0.01	83.05$\pm$0.21	5.30$\pm$0.01

noise removes hard positive labels that are difficult to recognize. Although it weakens the supervision for some true categories, it mainly leads to insufficient learning rather than forcing the model to fit incorrect category signals during training. These observations motivate PAIR-Net to focus its explicit noise filtering mechanism on the more destructive additive noise, thereby mitigating the dominant effect of label noise under the DD-IDMN noise setting.

F. Comparison with Different Noise Injection Strategies

We compare DD-IDMN with different RS multilabel noise generation methods using the BCE baseline method under the same experimental settings. As shown in Table I, DD-IDMN yields lower mAP and higher Hamming loss than conventional instance-independent methods, including uniform, partial uniform, and distribution-aware noise models, across different noise levels and datasets. These results indicate that our DD-IDMN provides a more challenging synthetic benchmark for evaluating model robustness.

More importantly, DD-IDMN provides more precise noise control than IDMN. As demonstrated in Fig. 6, the actual number of noisy labels generated by IDMN deviates from the target quantity, especially on the more challenging MLRSNet dataset. This observation is consistent with our analysis that the number of noisy labels generated by IDMN may deviate from the intended target, resulting in an unbalanced benchmark. In contrast, DD-IDMN explicitly decouples the generation of additive and subtractive noise, allowing it to match the additive and subtractive corruption quantities across datasets. This design enables a fairer and more systematic evaluation of model robustness under different types of multilabel noise.

G. Comparison with Different Noise-robust Learning Methods

To evaluate PAIR-Net under DD-IDMN, we compare it with representative noise-robust multilabel learning methods on UCMerced [56], AID [57], and MLRSNet [58]. As demonstrated in Table II, LCR and ELR-ML provide limited gains over BCE, suggesting that label correlation and early learning regularization are insufficient for the challenging mixed noise generated by DD-IDMN. Additionally, RCML and JoCoR-ML obtain lower Hamming loss in some cases, especially on UCMerced and AID, indicating that collaborative training can help reduce label-wise prediction errors. However, these gains are not consistently reflected in mAP, showing limited overall robustness to challenging instance-dependent noise.

In contrast, SAT-ML achieves more competitive performance, indicating that self-adaptive target refinement is beneficial for DD-IDMN. By updating training targets with an exponential moving average of model predictions, SAT-ML reduces the direct influence of unreliable labels. Nevertheless, PAIR-Net achieves the best mAP on all evaluated datasets and the best or second-best Hamming loss in most cases, especially at higher noise rates (*i.e.*, 30% and 40%). The results demonstrate that PAIR-Net effectively suppresses noisy labels and maintains reliable learning under mixed DD-IDMN.

We further evaluate all methods under additive and subtractive noise generated by DD-IDMN. As shown in Table III, PAIR-Net consistently achieves the best mAP under additive noise, confirming its ability to suppress destructive added labels. This result is consistent with the design motivation of PAIR-Net, which explicitly filters unreliable positive supervision caused by additive noise. Under subtractive noise, PAIR-Net also remains competitive, as our PAI strengthens category-

TABLE IV
GENERALIZATION ACROSS DIVERSE NOISE GENERATION STRATEGIES ON UCMERGED (UCM) AND AID. Mix- η DENOTES THE TARGET NOISE QUANTITY K_η FOR BOTH ADDITIVE AND SUBTRACTIVE CORRUPTION. RESULTS ARE MEAN $\pm$ STD OVER 3 RUNS WITH SEEDS (0, 1, 2).

Datasets	Noise Model	Algorithm	Mix-10%		Mix-20%		Mix-30%		Mix-40%	
			mAP $\uparrow$ (%)	Ham. $\downarrow$ (%)	mAP $\uparrow$ (%)	Ham. $\downarrow$ (%)	mAP $\uparrow$ (%)	Ham. $\downarrow$ (%)	mAP $\uparrow$ (%)	Ham. $\downarrow$ (%)
UCM	Uniform	BCE	91.97 $\pm$ 0.80	6.68 $\pm$ 0.19	88.44 $\pm$ 0.37	7.96 $\pm$ 0.03	82.57 $\pm$ 1.61	9.80 $\pm$ 0.22	75.05 $\pm$ 0.11	11.66 $\pm$ 0.25
		RCML	91.70 $\pm$ 0.76	6.89 $\pm$ 0.47	89.35 $\pm$ 0.96	7.92 $\pm$ 0.47	84.54 $\pm$ 1.13	9.80 $\pm$ 0.25	76.62 $\pm$ 2.12	11.39 $\pm$ 0.71
		SAT-ML	93.15 $\pm$ 0.36	6.75 $\pm$ 0.25	91.31 $\pm$ 0.64	7.94 $\pm$ 0.23	88.47 $\pm$ 1.17	10.28 $\pm$ 0.11	83.73 $\pm$ 0.35	12.47 $\pm$ 0.51
		JoCoR-ML	93.49 $\pm$ 0.28	6.28$\pm$0.27	91.02 $\pm$ 0.72	7.83$\pm$0.47	87.66 $\pm$ 0.75	9.58 $\pm$ 0.24	80.93 $\pm$ 0.86	11.62 $\pm$ 0.32
		PAIR-Net (ours)	93.81$\pm$0.64	6.44 $\pm$ 0.80	91.96$\pm$0.38	7.88 $\pm$ 0.49	88.97$\pm$0.94	9.35$\pm$0.29	84.42$\pm$0.61	10.24$\pm$0.16
	Partial uniform	BCE	91.93 $\pm$ 0.97	5.82 $\pm$ 0.25	90.38 $\pm$ 0.85	6.37 $\pm$ 0.01	87.17 $\pm$ 1.29	7.57 $\pm$ 0.50	85.26 $\pm$ 0.58	8.54 $\pm$ 0.19
		RCML	93.41 $\pm$ 0.43	5.57 $\pm$ 0.21	92.09 $\pm$ 0.29	5.93 $\pm$ 0.16	88.05 $\pm$ 1.48	7.24 $\pm$ 0.47	85.48 $\pm$ 1.62	8.59 $\pm$ 0.39
		SAT-ML	92.24 $\pm$ 0.88	6.23 $\pm$ 0.31	90.74 $\pm$ 0.61	6.83 $\pm$ 0.14	88.49 $\pm$ 0.54	7.13 $\pm$ 0.26	85.26 $\pm$ 0.31	7.53$\pm$0.13
		JoCoR-ML	93.13 $\pm$ 0.72	5.64 $\pm$ 0.22	91.41 $\pm$ 0.80	6.23 $\pm$ 0.29	87.69 $\pm$ 0.58	7.63 $\pm$ 0.41	83.08 $\pm$ 1.04	9.56 $\pm$ 0.55
		PAIR-Net (ours)	94.09$\pm$0.28	5.33$\pm$0.78	92.30$\pm$0.51	5.32$\pm$0.33	89.08$\pm$0.76	6.91$\pm$0.10	87.89$\pm$1.43	8.40 $\pm$ 0.43
	Distribution-aware	BCE	92.35 $\pm$ 0.32	7.02 $\pm$ 0.30	89.14 $\pm$ 1.01	9.21 $\pm$ 0.41	85.01 $\pm$ 0.88	11.74 $\pm$ 0.43	77.68 $\pm$ 0.11	15.44 $\pm$ 0.30
		RCML	91.90 $\pm$ 0.35	6.93 $\pm$ 0.29	89.29 $\pm$ 0.40	9.15 $\pm$ 0.52	85.24 $\pm$ 0.81	11.72 $\pm$ 0.39	78.92 $\pm$ 0.68	15.18 $\pm$ 0.25
		SAT-ML	93.16 $\pm$ 0.20	6.61 $\pm$ 0.28	92.03 $\pm$ 0.46	8.22 $\pm$ 0.09	88.12 $\pm$ 0.55	10.27 $\pm$ 0.37	83.70 $\pm$ 0.55	14.59 $\pm$ 0.47
		JoCoR-ML	93.34 $\pm$ 0.36	6.09 $\pm$ 0.29	91.42 $\pm$ 0.45	8.12 $\pm$ 0.37	87.48 $\pm$ 0.72	10.87 $\pm$ 0.42	80.11 $\pm$ 0.50	15.55 $\pm$ 0.60
		PAIR-Net (ours)	93.71$\pm$0.75	6.29$\pm$0.54	92.80$\pm$0.40	9.09$\pm$0.24	88.92$\pm$0.63	10.16$\pm$0.69	84.86$\pm$0.27	12.11$\pm$0.71
	IDMN	BCE	90.80 $\pm$ 0.38	7.61 $\pm$ 0.33	83.66 $\pm$ 0.79	10.17 $\pm$ 0.13	70.60 $\pm$ 0.88	13.64 $\pm$ 0.36	58.74 $\pm$ 1.33	18.06 $\pm$ 0.49
		RCML	89.99 $\pm$ 0.27	7.80 $\pm$ 0.25	85.20 $\pm$ 0.16	10.36 $\pm$ 0.33	72.39 $\pm$ 0.43	12.96 $\pm$ 0.18	60.94 $\pm$ 0.13	17.71 $\pm$ 0.06
		SAT-ML	91.83 $\pm$ 0.15	7.69 $\pm$ 0.26	87.08 $\pm$ 0.70	10.35 $\pm$ 0.15	76.63 $\pm$ 0.38	12.75 $\pm$ 0.04	63.90 $\pm$ 0.22	15.31 $\pm$ 0.22
		JoCoR-ML	90.54 $\pm$ 0.41	7.36 $\pm$ 0.05	82.24 $\pm$ 0.47	10.44 $\pm$ 0.14	69.43 $\pm$ 0.33	12.92 $\pm$ 0.30	59.71 $\pm$ 0.48	17.68 $\pm$ 0.14
		PAIR-Net (ours)	93.43$\pm$0.22	7.28$\pm$0.22	88.76$\pm$0.61	9.93$\pm$0.19	81.34$\pm$0.29	11.90$\pm$0.20	72.95$\pm$0.31	14.93$\pm$0.21
AID	Uniform	BCE	80.19 $\pm$ 0.28	6.81 $\pm$ 0.10	75.10 $\pm$ 0.30	8.54 $\pm$ 0.13	66.77 $\pm$ 0.96	11.60 $\pm$ 0.27	55.91 $\pm$ 0.71	16.36 $\pm$ 0.52
		RCML	79.94 $\pm$ 0.14	6.59 $\pm$ 0.12	75.38 $\pm$ 0.55	8.10 $\pm$ 0.10	67.40 $\pm$ 0.32	11.10 $\pm$ 0.15	57.66 $\pm$ 1.62	15.54 $\pm$ 0.68
		SAT-ML	81.60 $\pm$ 0.30	6.65 $\pm$ 0.19	77.99 $\pm$ 0.52	7.12$\pm$0.10	74.21 $\pm$ 0.50	9.31$\pm$0.20	69.45 $\pm$ 0.24	14.59 $\pm$ 0.51
		JoCoR-ML	81.50 $\pm$ 0.31	6.66 $\pm$ 0.05	76.74 $\pm$ 0.87	8.63 $\pm$ 0.18	70.02 $\pm$ 0.58	11.88 $\pm$ 0.24	59.74 $\pm$ 0.56	16.43 $\pm$ 0.19
		PAIR-Net (ours)	82.53$\pm$0.65	6.48$\pm$0.17	78.93$\pm$0.65	7.57 $\pm$ 0.49	75.36$\pm$0.41	10.10 $\pm$ 0.42	71.64$\pm$0.58	14.54$\pm$0.15
	Partial uniform	BCE	80.58 $\pm$ 0.57	6.19 $\pm$ 0.04	77.74 $\pm$ 0.93	6.45 $\pm$ 0.16	75.08 $\pm$ 0.71	7.32 $\pm$ 0.11	72.39 $\pm$ 0.52	8.25 $\pm$ 0.46
		RCML	82.02 $\pm$ 0.33	5.92$\pm$0.12	79.03 $\pm$ 0.87	6.33 $\pm$ 0.15	76.13 $\pm$ 0.38	7.05 $\pm$ 0.10	73.93 $\pm$ 0.36	7.80 $\pm$ 0.62
		SAT-ML	80.85 $\pm$ 0.52	6.29 $\pm$ 0.07	78.34 $\pm$ 0.64	6.48 $\pm$ 0.14	75.66 $\pm$ 0.27	6.65$\pm$0.13	73.86 $\pm$ 0.25	6.90$\pm$0.12
		JoCoR-ML	80.08 $\pm$ 0.41	6.12 $\pm$ 0.12	78.14 $\pm$ 0.63	6.36 $\pm$ 0.21	74.94 $\pm$ 0.52	7.36 $\pm$ 0.12	71.76 $\pm$ 0.64	8.47 $\pm$ 0.61
		PAIR-Net (ours)	82.64$\pm$0.53	6.04 $\pm$ 0.42	79.82$\pm$1.35	6.27$\pm$0.23	77.64$\pm$0.63	6.91 $\pm$ 0.21	74.01$\pm$0.63	7.41 $\pm$ 0.22
	Distribution-aware	BCE	80.43 $\pm$ 0.61	8.23 $\pm$ 0.37	73.63 $\pm$ 0.38	11.79 $\pm$ 0.26	65.03 $\pm$ 0.20	16.08 $\pm$ 0.47	57.88 $\pm$ 0.86	20.59 $\pm$ 0.16
		RCML	79.38 $\pm$ 0.50	8.08 $\pm$ 0.17	73.34 $\pm$ 0.35	11.70 $\pm$ 0.07	65.57 $\pm$ 0.57	15.86 $\pm$ 0.20	57.82 $\pm$ 1.47	19.93 $\pm$ 0.52
		SAT-ML	81.92 $\pm$ 0.14	7.63$\pm$0.11	78.22 $\pm$ 0.50	11.38 $\pm$ 0.12	67.16 $\pm$ 0.23	14.20 $\pm$ 0.09	61.82 $\pm$ 0.33	16.07 $\pm$ 0.03
		JoCoR-ML	80.41 $\pm$ 0.44	7.70 $\pm$ 0.24	73.55 $\pm$ 0.13	11.54 $\pm$ 0.24	66.11 $\pm$ 0.28	16.45 $\pm$ 0.29	60.69 $\pm$ 0.74	20.52 $\pm$ 0.54
		PAIR-Net (ours)	82.38$\pm$0.44	7.69 $\pm$ 0.50	79.18$\pm$0.30	10.57$\pm$0.49	68.78$\pm$0.38	14.06$\pm$0.46	62.75$\pm$0.62	16.05$\pm$0.36
	IDMN	BCE	76.34 $\pm$ 0.49	9.30 $\pm$ 0.04	69.88 $\pm$ 0.07	13.34 $\pm$ 0.05	63.90 $\pm$ 0.36	17.96 $\pm$ 0.17	55.94 $\pm$ 0.26	23.53 $\pm$ 0.06
		RCML	76.49 $\pm$ 0.35	8.71 $\pm$ 0.16	70.84 $\pm$ 0.27	12.58 $\pm$ 0.14	65.16 $\pm$ 0.33	17.48 $\pm$ 0.24	56.14 $\pm$ 0.48	23.40 $\pm$ 0.13
		SAT-ML	76.99 $\pm$ 0.55	8.59 $\pm$ 0.07	69.97 $\pm$ 0.11	12.15 $\pm$ 0.12	63.09 $\pm$ 0.37	16.75 $\pm$ 0.07	55.34 $\pm$ 0.47	22.22 $\pm$ 0.08
		JoCoR-ML	76.42 $\pm$ 0.17	8.90 $\pm$ 0.20	68.90 $\pm$ 0.26	12.72 $\pm$ 0.03	63.56 $\pm$ 0.30	17.76 $\pm$ 0.16	54.43 $\pm$ 0.32	23.39 $\pm$ 0.39
		PAIR-Net (ours)	79.82$\pm$0.62	8.48$\pm$0.45	76.13$\pm$0.80	11.80$\pm$0.14	70.52$\pm$1.05	15.27$\pm$0.61	64.06$\pm$0.61	19.31$\pm$0.10

specific representations through textual semantic guidance and helps infer visually present yet unlabeled categories. Overall, the results under additive, subtractive, and mixed noise demonstrate the robustness of PAIR-Net under DD-IDMN.

H. Generalization Across Diverse Noise Generation Strategies

To further validate the generalization capability of PAIR-Net beyond DD-IDMN, we conduct experiments on UCMerced and AID datasets under diverse mixed noise scenarios. The evaluated noise generation methods include instance-independent noise models (*i.e.*, Uniform, Partial uniform, and Distribution-aware noise) and the instance-dependent IDMN method, thereby covering corruption guided by both random sampling and image content. For each strategy, we compare

PAIR-Net with representative noise-robust multilabel learning methods under different mixed noise rates. As demonstrated in Table IV, PAIR-Net maintains robust performance across different noise generation strategies. Under instance-independent noise settings, PAIR-Net achieves competitive or superior performance in most cases, indicating that the proposed prompt-guided RS feature representation and activation-aware filtering strategy can also reduce the influence of randomly corrupted label entries. Furthermore, under IDMN, PAIR-Net obtains consistent gains on both UCMerced and AID, especially at higher noise rates such as Mix-40%. These results indicate that PAIR-Net is not specialized to our DD-IDMN. Instead, it provides a general robust learning framework for RS multilabel classification under diverse noisy supervision patterns.

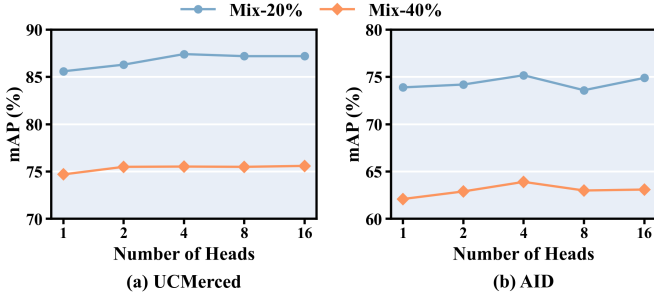


Fig. 7. Classification performance with different numbers of attention heads on the UCMerced and AID datasets under mixed noise settings.

TABLE V
IMPACT OF DIFFERENT COMPONENTS IN PAIR-Net ON THE UCMERGED (UCM) AND AID DATASETS UNDER MIXED NOISE SETTINGS.

Dataset	Method	Mix-20%		Mix-40%	
		mAP $\uparrow$	Ham. $\downarrow$	mAP $\uparrow$	Ham. $\downarrow$
UCM	BCE	78.31 $\pm$ 0.98	11.46 $\pm$ 0.12	53.96 $\pm$ 0.70	17.65 $\pm$ 0.18
	+ PAI	82.15 $\pm$ 0.72	11.36 $\pm$ 0.28	58.36 $\pm$ 0.73	16.41 $\pm$ 0.17
	+ GradCAM	85.29 $\pm$ 0.20	10.98 $\pm$ 0.35	71.33 $\pm$ 2.67	14.57 $\pm$ 0.32
	+ ADF w/o $\mathcal{P}$	86.36 $\pm$ 0.32	10.47 $\pm$ 0.26	73.94 $\pm$ 2.43	14.72 $\pm$ 0.35
	PAIR-Net w/o PAI	83.84 $\pm$ 0.51	10.95 $\pm$ 0.36	68.91 $\pm$ 2.64	15.42 $\pm$ 0.33
AID	BCE	68.96 $\pm$ 0.09	14.44 $\pm$ 0.28	54.74 $\pm$ 0.43	24.37 $\pm$ 0.09
	+ PAI	70.81 $\pm$ 0.46	13.97 $\pm$ 0.16	58.77 $\pm$ 0.96	23.40 $\pm$ 0.03
	+ GradCAM	71.11 $\pm$ 0.72	12.48 $\pm$ 0.12	61.21 $\pm$ 0.32	19.88 $\pm$ 0.12
	+ ADF w/o $\mathcal{P}$	72.59 $\pm$ 0.78	12.52 $\pm$ 0.16	62.75 $\pm$ 0.54	19.82 $\pm$ 0.10
	PAIR-Net w/o PAI	73.95 $\pm$ 0.27	12.53 $\pm$ 0.04	62.02 $\pm$ 0.21	20.27 $\pm$ 0.45
	PAIR-Net	87.42$\pm$0.48	10.16$\pm$0.09	75.53$\pm$0.60	14.41$\pm$0.06
	PAIR-Net	75.18$\pm$0.82	12.46$\pm$0.12	63.91$\pm$0.44	19.45$\pm$0.07

I. Ablation Study

In all main experiments, PAIR-Net adopts the same configuration across datasets and noise rates. To analyze each component and configuration, we conduct ablation studies on UCMerced and AID under mixed noise settings. Specifically, we evaluate PAIR-Net components and parameters, including the text encoder, learnable context tokens, attention heads, filtering threshold, warm-up duration, and filtering duration.

Impact of Different Components. We evaluate each component in PAIR-Net under mixed noise rates. As detailed in Table V, introducing the Prompt-guided Attention Interaction (PAI) module consistently improves performance over BCE. This result indicates that cross-modal semantic priors enhance RS class discrimination under noisy supervision. Although GradCAM-guided filtering brings further gains by identifying unreliable positive labels, replacing it with refined activation maps $\hat{\mathbf{G}}_{i,c}$ in our dynamic filtering (ADF) module improves performance. This improvement demonstrates that refined activation maps provide reliable spatial evidence for filtering additive noise. Furthermore, we isolate the effect of ADF without the textual prior from PAI, denoted as PAIR-Net w/o PAI. This variant still improves over BCE on both datasets, showing that ADF can enhance noise robustness without cross-modal interaction. Overall, PAIR-Net achieves the best performance, indicating that PAI and ADF provide complementary benefits under noisy multilabel supervision.

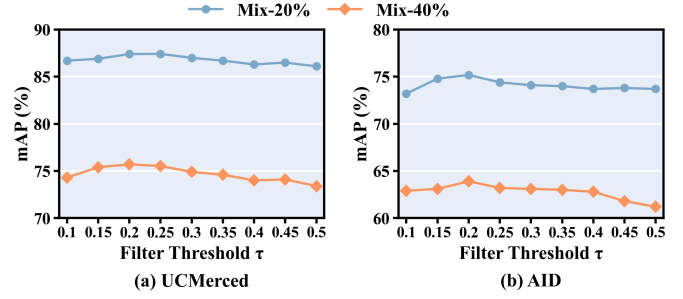


Fig. 8. Classification performance with different filtering thresholds τ on the UCMerced and AID datasets under mixed noise settings.

Impact of the Number of Learnable Context Tokens. We evaluate the effect of the prompt context length M on PAIR-Net. As shown in Table VI, learnable prompts generally outperform the fixed hand-crafted prompt (*i.e.*, "an aerial picture of [CLASS]" with $M = 0$). This result indicates that learnable context tokens improve the context modeling capacity of text prompts, helping the model capture domain-specific semantics for RS categories. However, increasing M to 20 does not bring consistent further improvement. This phenomenon can be attributed to the fact that excessive context tokens may introduce redundant prompt parameters and weaken the general semantic prior of CLIP under noisy supervision. To balance noise robustness and effective representation learning, we employ $M = 16$ for PAIR-Net in all experiments.

Impact of Different Text Encoders. We investigate the influence of CLIP text encoders in PAIR-Net on UCMerced and AID under mixed noise settings. The results in Table VII demonstrate that incorporating textual semantic information improves performance over the BCE baseline, indicating that text semantic priors are beneficial for robust RS multilabel classification under noisy supervision. Among the evaluated encoders, ViT-L/14 achieves the best overall performance across both datasets and noise rates. These results suggest that a stronger text encoder provides more discriminative semantic representations, enabling PAIR-Net to learn robust class-specific features under mixed noise settings. Thus, the text encoder of CLIP ViT-L/14 is employed in PAIR-Net.

Impact of Different Numbers of Attention Heads. We investigate the influence of the number of attention heads in the PAI module on PAIR-Net. As illustrated in Fig. 7, increasing the number of heads from 1 to 4 generally enhances classification performance on both UCMerced and AID datasets. This trend suggests that multiple attention heads capture diverse cross-modal interaction patterns and strengthen RS feature discrimination under noisy supervision. However, further increasing the number of heads to 8 or 16 yields no consistent gains and even causes slight degradation in some cases. This phenomenon can be attributed to the fact that excessive attention heads introduce redundant interaction patterns and make the model more sensitive to noisy semantic correlations. Moreover, a larger number of heads inevitably increases the computational overhead. Thus, we set the number of attention heads to 4 for PAIR-Net in all experiments to balance classification performance and model efficiency.

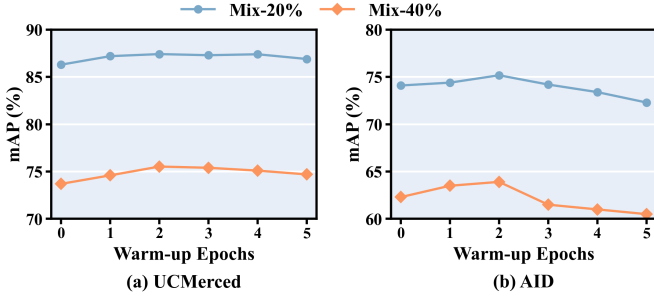


Fig. 9. Classification performance with different numbers of warm-up epochs on the UCMerced and AID datasets under mixed noise settings.

TABLE VI
IMPACT OF LEARNABLE CONTEXT TOKENS M ON THE UCMERGED (UCM) AND AID DATASETS UNDER MIXED NOISE SETTINGS.

Dataset	M	Mix-20%		Mix-40%	
		mAP $\uparrow$	Ham. $\downarrow$	mAP $\uparrow$	Ham. $\downarrow$
UCM	0	86.50 $\pm$ 0.42	10.65 $\pm$ 0.16	68.93 $\pm$ 1.04	14.60 $\pm$ 0.29
	4	86.15 $\pm$ 1.40	10.59 $\pm$ 0.31	75.30 $\pm$ 0.71	14.60 $\pm$ 0.28
	8	87.01 $\pm$ 0.81	10.17 $\pm$ 0.26	75.41 $\pm$ 0.89	14.53 $\pm$ 0.21
	16	87.42 $\pm$ 0.48	10.16 $\pm$ 0.09	75.53 $\pm$ 0.60	14.41 $\pm$ 0.06
	20	87.65 $\pm$ 0.75	10.24 $\pm$ 0.32	74.53 $\pm$ 1.27	14.71 $\pm$ 0.30
AID	0	72.57 $\pm$ 1.32	12.58 $\pm$ 0.14	59.22 $\pm$ 0.11	19.72 $\pm$ 0.48
	4	74.14 $\pm$ 1.03	12.52 $\pm$ 0.11	62.05 $\pm$ 1.07	19.69 $\pm$ 0.37
	8	74.46 $\pm$ 0.99	12.50 $\pm$ 0.12	63.13 $\pm$ 1.97	19.53 $\pm$ 0.32
	16	75.18 $\pm$ 0.82	12.46 $\pm$ 0.12	63.91 $\pm$ 0.44	19.45 $\pm$ 0.07
	20	74.56 $\pm$ 1.21	12.26 $\pm$ 0.53	62.82 $\pm$ 1.26	19.57 $\pm$ 0.29

Impact of Different Filter Thresholds. We further investigate the influence of the filter threshold τ within our activation-aware dynamic filtering module. The parameter τ controls the tolerance to cross-class semantic interference, where a positive label is filtered if its Spatial Containment Score (SCS) exceeds this threshold. Specifically, a smaller τ indicates a more aggressive filtering strategy, meaning that the model faces a greater risk of erroneously discarding genuine positive labels. Conversely, a larger τ implies conservative filtering, which struggles to detect and clean false positive labels. As illustrated in Fig. 8, the performance shows a consistent favorable range across datasets and noise conditions. Specifically, mAP generally reaches its best value at $\tau = 0.2$ on both UCMerced and AID datasets. When τ is overly low (*i.e.*, 0.1), the performance becomes suboptimal, reflecting the unintended removal of hard but clean present labels. Conversely, as τ increases beyond 0.3, the performance experiences noticeable degradation. This result demonstrates that an excessively high threshold fails to effectively suppress deceptive additive noise labels. Thus, to strike an optimal balance between preserving clean targets and eliminating noise labels, we set $\tau = 0.2$ as the filter threshold for all experiments.

Impact of the Number of Warm-up Epochs. The warm-up phase allows the model to develop sufficient RS visual representations before our dynamic filtering (ADF) is activated. As demonstrated in Fig. 9, using no warm-up stage or only one warm-up epoch results in suboptimal performance. This result indicates that the model has not learned reliable activation maps without a sufficient warm-up stage, causing our ADF to

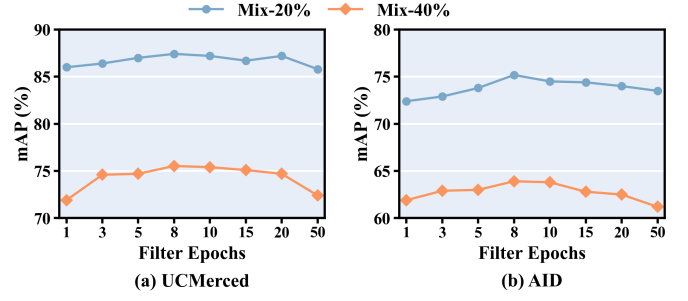


Fig. 10. Classification performance with different numbers of filtering epochs on the UCMerced and AID datasets under mixed noise settings.

TABLE VII
IMPACT OF DIFFERENT TEXT ENCODERS IN PAIR-NET ON UCMERGED (UCM) AND AID UNDER MIXED NOISE SETTINGS.

Dataset	Text Encoder	Mix-20%		Mix-40%	
		mAP $\uparrow$	Ham. $\downarrow$	mAP $\uparrow$	Ham. $\downarrow$
UCM	BCE	78.31 $\pm$ 0.98	11.46 $\pm$ 0.12	53.96 $\pm$ 0.70	17.65 $\pm$ 0.18
	ViT-B/16	86.62 $\pm$ 1.06	10.62 $\pm$ 0.15	72.53 $\pm$ 0.60	14.93 $\pm$ 0.12
	ViT-B/32	87.04 $\pm$ 0.58	10.73 $\pm$ 0.19	73.90 $\pm$ 0.62	14.86 $\pm$ 0.19
	ViT-L/14	87.42 $\pm$ 0.48	10.16 $\pm$ 0.09	75.53 $\pm$ 0.60	14.41 $\pm$ 0.06
AID	BCE	68.96 $\pm$ 0.09	14.44 $\pm$ 0.28	54.74 $\pm$ 0.43	24.37 $\pm$ 0.09
	ViT-B/16	73.53 $\pm$ 0.40	12.60 $\pm$ 0.13	61.95 $\pm$ 0.62	19.91 $\pm$ 0.29
	ViT-B/32	74.18 $\pm$ 0.85	12.53 $\pm$ 0.06	63.41 $\pm$ 0.13	19.71 $\pm$ 0.38
	ViT-L/14	75.18 $\pm$ 0.82	12.46 $\pm$ 0.12	63.91 $\pm$ 0.44	19.45 $\pm$ 0.07

mistakenly discard clean labels. The performance reaches its best value when the warm-up duration is set to two epochs, suggesting that the network has learned adequate preliminary features to generate reasonably reliable activations for filtering at this early training stage. However, further increasing the number of warm-up epochs degrades performance. This phenomenon can be attributed to the gradual memorization of noisy labels, which corrupts the activation maps and weakens the reliability of ADF. Thus, we adopt a two-epoch warm-up stage for PAIR-Net in all experiments to balance activation reliability against the risk of noise memorization.

Impact of Different Filtering Durations. We investigate the influence of the filtering epochs, which determine the duration of the dynamic noise filtering phase. As shown in Fig. 10, an insufficient number of filtering epochs, *i.e.*, from 1 to 5 epochs, yields suboptimal performance. A short filtering phase fails to sufficiently purify the training set, leaving many noisy labels that continue to corrupt the learned representations. The performance peaks at 8 filtering epochs, indicating that the filtering mechanism can effectively remove deceptive noisy labels within this duration while avoiding excessive correction of clean positive labels. However, further extending the filtering duration, *i.e.*, from 15 to 50 epochs, slightly degrades performance. This phenomenon occurs because the model may gradually capture noise-specific patterns in later training stages, which leads to less reliable spatial activation maps and increases the risk of accidentally filtering clean samples. Thus, we set the number of filtering epochs to 8 to balance noise purification and training stability.

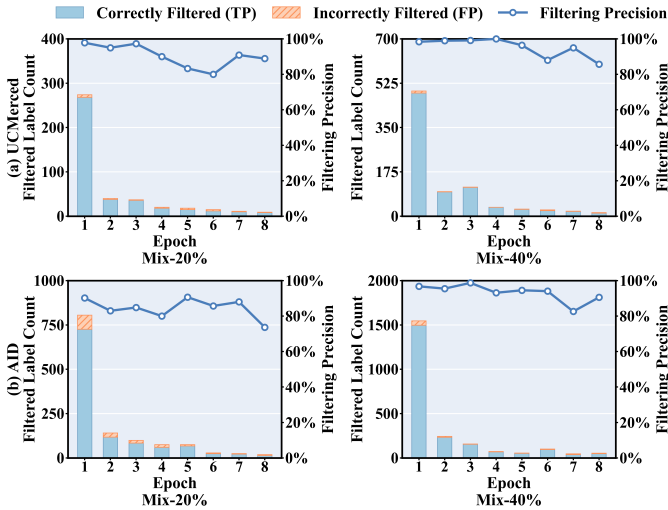


Fig. 11. Filtered label quantities and precision of PAIR-Net during the entire filtering stage under DD-IDMN on the UCMerced and AID datasets. The stacked bars report the numbers of correctly filtered additive noisy labels (TP) and incorrectly filtered genuine positive labels (FP), while the blue curves denote the filtering precision, *i.e.*, $TP/(TP + FP)$.

TABLE VIII

COMPUTATIONAL COST COMPARISON ON THE MLRSNET DATASET WITH 60 OUTPUT CATEGORIES USING A SINGLE NVIDIA RTX 3090 GPU.

Method	Params (M)	FLOPs (G)	Inference (ms/img)
BCE	23.63	4.13	0.70
PAIR-Net	32.17	4.63	0.78

Filtering Reliability Analysis. To analyze PAIR-Net’s filtering reliability, we record the number of filtered labels and filtering precision on UCMerced and AID. As shown in Fig. 11, PAIR-Net maintains high filtering precision, averaging 94.79% across evaluated settings. Moreover, most noisy labels are removed in early filtering epochs, while only a few are filtered in later epochs. This phenomenon suggests that PAIR-Net mainly removes highly suspicious positives at the early stage and avoids aggressively discarding clean positive labels.

To further examine filtering behavior under a more severe noise setting, we present a noise identification case study at Additive-60% in Fig. 12. This visualization further supports PAIR-Net’s ability to reliably reduce additive noise under more challenging and diverse noisy supervision.

J. Computational Efficiency

As summarized in Table VIII, we report the computational cost on MLRSNet with 60 output categories, including trainable parameters, FLOPs, and inference time. Compared with BCE, PAIR-Net incurs minor inference overhead, since the filtering module is used only during training. Thus, PAIR-Net improves noise robustness with limited inference latency.

V. LIMITATION

In our current formulation, PAIR-Net is primarily designed to mitigate additive noise rather than jointly correct additive and subtractive noise. This design is motivated by empirical analysis under DD-IDMN, where additive noise yields more



Fig. 12. Case study of noise identification under the Additive-60% DD-IDMN setting, where red labels indicate additive noisy labels. The filtered labels obtained by PAIR-Net show a clear reduction in additive noisy labels.

severe degradation than subtractive noise. Accordingly, PAIR-Net focuses on identifying and rectifying unreliable positive labels caused by additive corruption, without explicitly recovering missing positive labels. Despite this limitation, PAIR-Net remains effective across noise settings. Specifically, our prompt-guided representation learning exploits textual semantics to enhance RS visual discriminability under noisy supervision, while ADF suppresses false positive supervision. In future work, we will explore unified label rectification to simultaneously suppress false positive labels and recover missed true labels for improved noise robustness.

VI. CONCLUSION

In this study, we revisit instance-dependent multilabel noise generation for remote sensing multilabel scene classification. We identify the cross-distribution thresholding bias in IDMN, where the score distribution of present labels is utilized to determine both additive and subtractive noise. To address this limitation, we propose DD-IDMN, which decouples the score distributions of present and absent labels for precise noise control. Through experimental analysis under DD-IDMN, we observe that additive noise yields more severe performance degradation than subtractive noise. Guided by this observation, we propose PAIR-Net from an activation-aware label rectification perspective. Specifically, it combines prompt-guided RS representation learning with dynamic filtering of unreliable positive labels. By analyzing activation differences between true positive and false positive labels, we design the Spatial Containment Score to identify additive noisy labels whose activations are largely contained within those of other positive categories. Comprehensive experiments demonstrate that PAIR-Net outperforms representative noise-robust multilabel methods across various noise generation methods.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 62372155 and 62302149, in part by Basic Research Program of Jiangsu under Grant BK20250188, in part by the Research Funds of Jiangsu Hydraulic Research Institute under Grant 2025z065, in part by the Major Science and Technology Program of the Ministry of Water Resources of China under Grant SKS-2022072 and in part by the China Postdoctoral Science Foundation under Grant 2025M771578.

REFERENCES

- [1] J. Guo, S. Jiao, H. Sun, B. Song, and Y. Chi, "Cross-modal compositional learning for multilabel remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 5810–5823, 2025.
- [2] R. Du, X. Tang, J. Ma, X. Zhang, and L. Jiao, "Mlamba: A mamba-based efficient network for multi-label remote sensing scene classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 7, pp. 6245–6258, 2025.
- [3] R. Du, X. Tang, J. Ma, X. Zhang, F. Liu, and L. Jiao, "Semantic-assisted feature integration network for multilabel remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–15, 2025.
- [4] S. Pandey, P. Echuri, V. M. Vemulapalli, and S. Chakraborty, "Multi-label classification in remote sensing: Leveraging high-resolution patches for low-resolution tasks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026, pp. 1512–1520.
- [5] D. Yu, S. Ji, S. Wei, and K. Khoshelham, "3-d building instance extraction from high-resolution remote sensing images and dsm with an end-to-end deep neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–19, 2024.
- [6] K. Karalas, G. Tsagkatakis, M. Zervakis, and P. Tsakalides, "Land classification using remotely sensed data: Going multilabel," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3548–3563, 2016.
- [7] C.-H. Lin, J.-T. Lin, P.-Y. Chiu, S.-P. Chen, and C. C. Lin, "Quantum-driven multihead inland waterbody detection with transformer-encoded cygnss delay-doppler map data," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–21, 2025.
- [8] G. Astruc, N. Gonthier, C. Mallet, and L. Landrieu, "Omnisat: Self-supervised modality fusion for earth observation," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 409–427.
- [9] Q. Zheng, R. Wang, G. Han, W. Li, X. Wang, Q. Shao, X. Wu, L. Cao, G. Zhou, and S. Hu, "A spatiotemporal multiscale deep learning model for subseasonal prediction of arctic sea ice," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–22, 2024.
- [10] S. Z. Zahra, N. Saher, K. Sattar, M. M. S. Missen, R. N. Bashir, M. Faheem, and A. Rehman, "Deep spatio-temporal learning for multi-hazard events: A congru multi-label classification approach," *Geoinformatica*, vol. 30, no. 1, p. 8, 2026.
- [11] Y. Li, P. Matgen, and M. Chini, "Quantifying and communicating uncertainty in sar-based flood mapping via density-aware neural networks and conformal risk control," *IEEE Trans. Geosci. Remote Sens.*, vol. 64, pp. 1–20, 2026.
- [12] D. Hong, C. Li, X. Li, G. Camps-Valls, and J. Chanussot, "Foundation models in remote sensing: Evolving from unimodality to multimodality," *IEEE Geosci. Remote Sens. Mag.*, 2026.
- [13] J. Lin, T. Yu, and Z. J. Wang, "Rethinking crowdsourcing annotation: partial annotation with salient labels for multilabel aerial image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.
- [14] G. Sumbul, A. De Wall, T. Kreuziger, F. Marcelino, H. Costa, P. Benevides, M. Caetano, B. Demir, and V. Markl, "Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 3, pp. 174–180, 2021.
- [15] Z. Sun, Y. Yao, T. Liu, Z. Li, F. Shen, and J. Tang, "Jo-snc: Combating noisy labels through fostering self-and neighbor-consistency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 48, no. 4, pp. 4708–4725, 2026.
- [16] M. Sheng, Z. Sun, T. Zhou, X. Shu, J. Pan, and Y. Yao, "Ca2c: A prior-knowledge-free approach for robust label noise learning via asymmetric co-learning and co-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2025, pp. 901–911.
- [17] H. Yang, Q. Yao, B. Han, and J. T. Kwok, "Searching to exploit memorization effect in deep learning with noisy labels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 7833–7849, 2024.
- [18] Y. Hua, S. Lobry, L. Mou, D. Tuia, and X. X. Zhu, "Learning multi-label aerial image classification under label noise: A regularization approach using word embeddings," in *IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 525–528.
- [19] T. Burgert, M. Ravanbakhsh, and B. Demir, "On the effects of different types of label noise in multi-label remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.
- [20] A. K. Aksoy, M. Ravanbakhsh, and B. Demir, "Multi-label noise robust collaborative learning for remote sensing image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 5, pp. 6438–6451, 2024.
- [21] Y. Kim, S. Kim, Y. Ro, and J. Lee, "Instance-dependent multilabel noise generation for multilabel remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 17087–17098, 2024.
- [22] S. Yang, S. Wu, E. Yang, B. Han, Y. Liu, M. Xu, G. Niu, and T. Liu, "A parametrical model for instance-dependent label noise," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14 055–14 068, 2023.
- [23] D. Cheng, T. Liu, Y. Ning, N. Wang, B. Han, G. Niu, X. Gao, and M. Sugiyama, "Instance-dependent label-noise learning with manifold-regularized transition matrix estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16 609–16 618.
- [24] X. Xia, T. Liu, B. Han, N. Wang, M. Gong, H. Liu, G. Niu, D. Tao, and M. Sugiyama, "Part-dependent label noise: Towards instance-dependent label noise," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 7597–7610, 2020.
- [25] D.-B. Wang, Y. Wen, L. Pan, and M.-L. Zhang, "Learning from noisy labels with complementary loss functions," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 11, 2021, pp. 10 111–10 119.
- [26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
- [27] L. Pang, J. Yao, K. Li, J. Zhou, D. Meng, and X. Cao, "Special: zero-shot hyperspectral image classification with clip," *IEEE Trans. Geosci. Remote Sens.*, 2026.
- [28] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou, "Remotclip: A vision language foundation model for remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024.
- [29] OpenAI, "GPT-5.5 System Card," 2026, [Online]. Available: <https://openai.com/index/gpt-5-5-system-card/>
- [30] T. Song, S. Bai, F. Yang, C. Gao, H. Chen, and J. Li, "Exploring hybrid contrastive learning and scene-to-label information for multilabel remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2024.
- [31] B. Ma, F. Wu, T. Hu, L. Fathollahi, X. Sui, Y. Liu, and B. Gantumur, "Label-driven graph convolutional network for multilabel remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2245–2255, 2024.
- [32] W. Zhang, M. Cai, Y. Ning, T. Zhang, Y. Zhuang, S. Lu, H. Chen, J. Li, and X. Mao, "Earthgpt-x: A spatial mllm for multilevel multisource remote sensing imagery understanding with visual prompting," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–21, 2025.
- [33] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 24 824–24 837, 2022.
- [34] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda, "Early-learning regularization prevents memorization of noisy labels," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 20 331–20 342, 2020.
- [35] Q. Huang, F. Jiang, and C. Huang, "Remote sensing scene classification with relation-aware dynamic graph neural networks," *Eng. Appl. Artif. Intell.*, vol. 150, p. 110513, 2025.
- [36] J. Ji, X. Lu, T. Zhang, Y. Guo, and G. Yang, "Remote sensing scene classification via pseudo-category-relation and orthogonal feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–14, 2025.
- [37] Y. Zhao, Y. Chen, S. Xiong, X. Lu, X. X. Zhu, and L. Mou, "Co-enhanced global-part integration for remote-sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2024.
- [38] J. Ma, W. Jiang, X. Tang, X. Zhang, F. Liu, and L. Jiao, "Multiscale sparse cross-attention network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–16, 2025.
- [39] K. Yang, N. Yang, and H. Tang, "Dynamic inference for on-orbit scene classification with the scale boosting model," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 138, p. 104447, 2025.
- [40] Y. Liu, Y. Zhong, S. Shi, and L. Zhang, "Scale-aware deep reinforcement learning for high resolution remote sensing imagery classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 209, pp. 296–311, 2024.
- [41] K. Wu, Y. Zhang, L. Ru, B. Dang, J. Lao, L. Yu, J. Luo, Z. Zhu, Y. Sun, J. Zhang *et al.*, "A semantic-enhanced multi-modal remote sensing foundation model for earth observation," *Nat. Mach. Intell.*, vol. 7, no. 8, pp. 1235–1249, 2025.
- [42] J. Wen, Z. Guo, Z. Zeng, and T. Wang, "Ip-scenenet: A multimodal network for scene classification with remote sensing images and lidar point clouds," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–18, 2025.
- [43] X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, L. Ru, L. Zhong, Z. Huang, K. Wu, D. Hu *et al.*, "Skysense: A multi-modal remote sensing

foundation model towards universal interpretation for earth observation imagery,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 27 662–27 673.

- [44] K. Chen, W. Li, J. Chen, Z. Zou, and Z. Shi, “Resolution-agnostic remote sensing scene classification with implicit neural representations,” *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
- [45] K. Chen, B. Chen, C. Liu, W. Li, Z. Zou, and Z. Shi, “Rsmamba: Remote sensing image classification with state space model,” *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.
- [46] Z. Geng, J. Wang, and R. Wang, “Tri-comamba: A tri-complementary mamba framework for multisource remote sensing image classification,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 19, pp. 7170–7187, 2026.
- [47] L. Peng, T. Wei, X. Chen, X. Chen, R. Sun, L. Wan, J. Chen, and X. Zhu, “Human-annotated label noise and their impact on convnets for remote sensing image scene classification,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 1500–1514, 2025.
- [48] Z. Zhu, T. Liu, and Y. Liu, “A second-order approach to learning with instance-dependent label noise,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10 108–10 118.
- [49] J. Cheng, T. Liu, K. Ramamohanarao, and D. Tao, “Learning with bounded instance and label-dependent label noise,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1789–1799.
- [50] L. Huang, C. Zhang, and H. Zhang, “Self-adaptive training: beyond empirical risk minimization,” *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 19 365–19 376, 2020.
- [51] H. Wei, L. Feng, X. Chen, and B. An, “Combating noisy labels by agreement: A joint training method with co-regularization,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13 726–13 735.
- [52] C. Liu, G. Perantoni, L. Bruzzone, and X. X. Zhu, “Adaptive gradient calibration for single-positive multi-label learning in remote sensing image scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 64, pp. 1–15, 2026.
- [53] T. Burgert, J. Henkel, and B. Demir, “Noise-adaptive regularization for robust multi-label remote sensing image classification,” *IEEE Trans. Geosci. Remote Sens.*, 2026.
- [54] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
- [55] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, 1979.
- [56] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in *Proc. 18th ACM SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.*, 2010, pp. 270–279.
- [57] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, “Aid: A benchmark data set for performance evaluation of aerial scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, 2017.
- [58] X. Qi, P. Zhu, Y. Wang, L. Zhang, J. Peng, M. Wu, J. Chen, X. Zhao, N. Zang, and P. T. Mathiopoulos, “Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding,” *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 337–350, 2020.
- [59] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, “Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, 2017.
- [60] Y. Hua, L. Mou, and X. X. Zhu, “Relation network for multilabel aerial image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4558–4572, 2020.
- [61] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv:1412.6980*, 2014.



Xinlei Zhang received the B.Eng. degree in Software Engineering from Chongqing Normal University (CNU), Chongqing, China, in 2023. She has been pursuing the Ph.D. degree in Computer Science and Technology from Hohai University (HHU), Nanjing, since 2024. Her research interests include noisy label learning, multilabel classification, and remote sensing object detection.



Fan Liu is a professor at Hohai University (HHU). He received B.S. and Ph.D. degrees from Nanjing University of Science and Technology (NJUST) in 2009 and 2015. From September 2008 to December 2008, he studied at Ajou University in South Korea. From February 2014 to May 2014, he worked at Microsoft Research Asia. His research interests include computer vision, pattern recognition, and machine learning. He serves as an Associate Editor for *Pattern Recognition* and *Remote Sensing*.



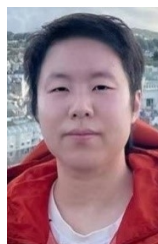
Chuanyi Zhang is an assistant professor at the College of Artificial Intelligence and Automation, Hohai University (HHU). He received a Ph.D. degree in 2023 from Nanjing University of Science and Technology (NJUST). His research interests include image processing, multi-modal learning, and machine learning.



Xiaoying Ji received the B.Eng. degree in Intelligent Science and Technology from Hohai University (HHU), Nanjing, China, in 2025. She is pursuing the M.S. degree in Computer Science and Technology at Hohai University. Her research interests include cross-modal retrieval and noisy correspondence learning.



Wenhui Wang received the B.Eng. degree in Artificial Intelligence from Shandong Agricultural University, Tai'an, China, in 2026. She is pursuing the M.S. degree in Computer Science and Technology at Hohai University. Her research interests include cross-modal retrieval, vision-language learning, and noisy correspondence learning.



research interests span multimedia computing, perceptual image processing, and computational vision.

Wei Zhou is an Assistant Professor at Cardiff University, United Kingdom. Previously, Wei studied and worked at other institutions such as the University of Waterloo (Canada), the National Institute of Informatics (Japan), the University of Science and Technology of China, Intel, Microsoft Research, and Alibaba Group. Dr. Zhou is now an Associate Editor of *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, and *Pattern Recognition*. Wei's



Jun Zhou (Fellow, IEEE) received a B.S. degree in computer science and a B.E. degree in international business from the Nanjing University of Science and Technology, Nanjing, China, in 1996 and 1998, respectively, the M.S. degree in computer science from Concordia University, Montreal, QC, Canada, in 2002, and the Ph.D. degree from the University of Alberta, Edmonton, AB, Canada, in 2006. In June 2012, he joined the School of Information and Communication Technology, Griffith University, Nathan, QLD, Australia, where he is currently a professor. Before this appointment, he was a Research Fellow with the Research School of Computer Science, Australian National University, Canberra, ACT, Australia, and a Researcher with the Canberra Research Laboratory, NICTA, Canberra. His research interests include pattern recognition, computer vision, and spectral imaging with their applications in remote sensing and environmental informatics. Dr. Zhou is an Associate Editor of *IEEE Transactions on Geoscience and Remote Sensing* and *Pattern Recognition*.