

SURVEY

From Pixels to Images: A Structural Survey of Deep Learning Paradigms in Remote Sensing Image Semantic Segmentation

QUANWEI LIU^{ID 1,2}, (Student Member, IEEE), TAO HUANG^{ID 1,2}, (Senior Member, IEEE),
JIAQI YANG^{ID 3}, (Member, IEEE), AND WEI XIANG^{ID 4}, (Senior Member, IEEE)

¹College of Science and Engineering, James Cook University, Cairns, QLD 4878, Australia

²Center for AI and Data Science Innovation, James Cook University, Cairns, QLD 4878, Australia

³Department of Forest and Wildlife Ecology, University of Wisconsin–Madison, Madison, WI 53706, USA

⁴School of Computing, Engineering, and Mathematical Sciences, La Trobe University, Melbourne, VIC 3086, Australia

Corresponding author: Tao Huang (tao.huang1@jcu.edu.au)

This work was supported in part by the Australian Government through the Australian Research Council's Discovery Projects Funding Scheme under Project DP220101634.

ABSTRACT Remote sensing images (RSIs) capture both natural and human-induced changes on the Earth's surface, serving as essential data for environmental monitoring, urban planning, and resource management. Semantic segmentation (SS) of RSIs enables the fine-grained interpretation of surface features, making it a critical task in RS analysis. With the increasing diversity and volume of RSIs collected by sensors on various platforms, traditional processing methods struggle to maintain efficiency and accuracy. In response, deep learning (DL) has emerged as a transformative approach, enabling substantial advances in remote sensing image semantic segmentation (RSISS) by automating hierarchical feature extraction and high segmentation performance. As researchers continue to explore end-to-end SS, DL-based RSSIS has undergone a structural evolution from pixel-level and patch-based classification to tile-level and image-level segmentation. However, existing reviews often focus on individual components, such as supervision strategies or fusion stages, and lack a unified operational perspective aligned with segmentation granularity and the training/inference pipeline. This paper provides a comprehensive review by organizing DL-based RSSIS into a pixel–patch–tile–image hierarchy, covering early pixel-based methods, prevailing patch-based and tile-based techniques, and emerging image-based approaches. Specifically, the survey analyzes four supervision strategies, eleven feature extraction strategies, and six information fusion strategies, revealing the field's progression from local to global feature extraction, from traditional DL architectures to foundation models, and from unimodal to multimodal segmentation. This review offers a holistic and structured understanding of DL-based RSSIS, highlighting representative datasets, comparative insights, and open challenges related to data scale, model efficiency, domain robustness, and multimodal integration. Furthermore, to facilitate reproducible research, curated code collections are provided at: RSSIS.

INDEX TERMS Remote sensing semantic segmentation, deep learning, segmentation granularity, pixel–patch–tile–image taxonomy, feature extraction, multimodal fusion, vision foundation models, remote sensing datasets.

I. INTRODUCTION

Diverse perception modalities have emerged alongside advances in electromagnetic spectrum research, including

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamed Soufiane Jouini^{ID}.

remote sensing (RS) via satellites, aircraft, and autonomous aerial vehicles (UAVs) equipped with various sensors. These modalities exhibit distinct spatial, spectral, temporal, and radiometric resolution characteristics. For instance, hyperspectral images (HSIs) offer both spatial context and detailed spectral information across numerous bands, while

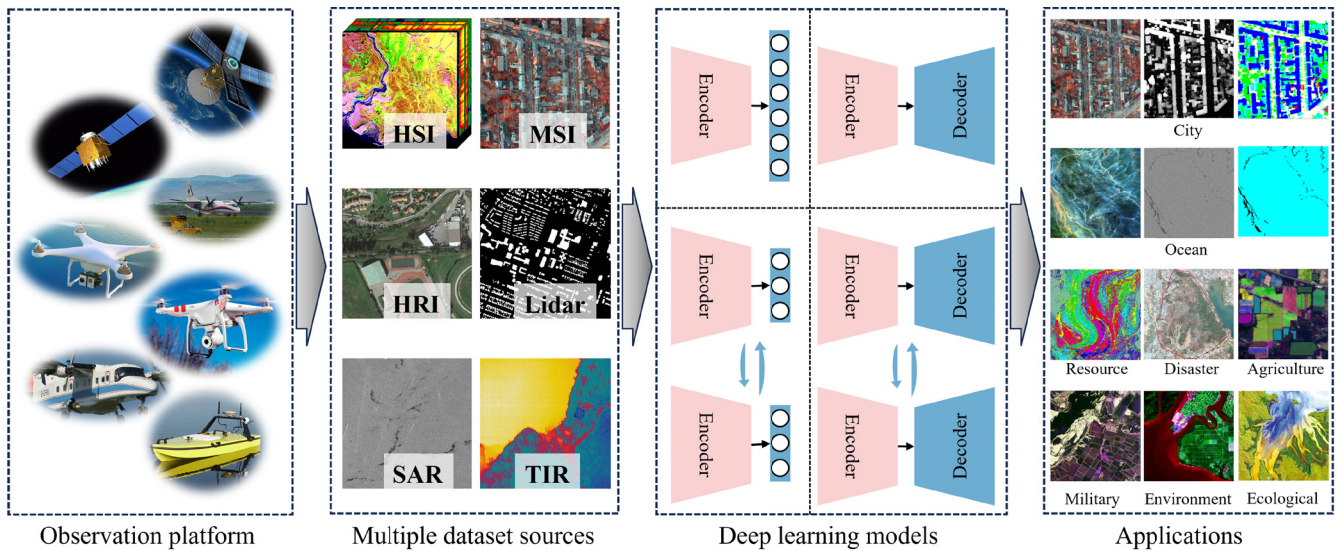


FIGURE 1. Processing flow for remote sensing image semantic segmentation. $\downarrow \uparrow$ denotes the feature interaction.

high-resolution imagery (HRI) provides finer textural detail and enhances the discrimination of small targets. Remote sensing image semantic segmentation (RSISS) seeks to classify RS imagery into distinct categories on a pixel-wise basis, enabling fine-grained interpretation of the Earth's surface. This capability plays a critical role in marine [1], urban [2], forest [3], arable [4], and disaster-related applications [5]. The semantic segmentation (SS) process is illustrated in Figure 1.

Early RSISS methods were primarily based on machine learning (ML) techniques relying on handcrafted feature extraction, including texture, structural, spectral, and scattering descriptors, combined with classifiers such as support vector machines (SVMs) and random forests. Although these approaches achieved notable progress, performance improvements often required increasingly complex feature engineering, limiting scalability and generalization [6]. The emergence of deep learning (DL) fundamentally transformed this paradigm by enabling automatic hierarchical feature extraction through data-driven architectures.

As researchers continue to explore end-to-end SS, DL-based RSISS has undergone a structural evolution in how segmentation is performed. Early DL models operated primarily at the pixel or patch level, where each pixel or local region was processed independently. In such settings, images were divided into numerous pixels or patches, and predictions were aggregated to form segmentation maps. Such methods are commonly referred to as classification models, specifically the pixel- or patch-level classification model. While pixel- and patch-based approaches require relatively small training data and model parameters, they suffer from repeated computations and limited global contextual awareness, which constrain generalization capability. As larger public datasets with pixel-level annotations became available, fully convolutional networks (FCNs) [7],

U-Net [8], vision foundation models (VFMs) [9] and computer cluster [10] enabled tile- and image-level semantic segmentation, significantly improving computational efficiency and contextual modeling.

This evolution from pixel-based classification to image-level segmentation reflects not only architectural innovation but also the co-development of data scale and feature extraction mechanisms. Across these regimes, advances in feature extraction, such as multi-scale modeling, global context learning, representation transfer, and domain-robust training, have played a central role in expanding receptive fields, improving robustness, and enabling cross-domain generalization. Accordingly, this paper organizes DL-based RSISS into a unified pixel-patch-tile-image hierarchy, explicitly linking segmentation granularity to training/inference pipelines and feature extraction strategies, as illustrated in Figure 3.

Beyond unimodal feature extraction, multimodal data fusion further enriches semantic representation. Different sensing modalities provide complementary information: for example, concrete pavements and roofs may share similar spectral signatures in HSI but differ in elevation captured by LiDAR [11], while certain vegetation types may be indistinguishable in LiDAR yet separable in HSI. Fusion of modalities such as HSI, LiDAR, multispectral imagery, and SAR enables more comprehensive and robust scene understanding. In this survey, multimodal fusion is examined as a complementary dimension that interacts with the proposed pixel-patch-tile-image framework.

Table 1 explicitly compares our proposed framework with existing surveys. Although numerous surveys have reviewed remote sensing image semantic segmentation from perspectives such as spectral and spatial feature extraction, supervision strategies, and fusion stages, most focus on individual components in isolation. These conventional taxonomies often detach algorithmic design from the realities

TABLE 1. A summary of the recently published surveys in remote sensing image semantic segmentation. The subscript in the first column indicates the year of publication. For example, Imani et al.₂₀ [12] indicate that this paper was published in 2020 by Imani et al.

Paper	Publication	Survey Topic	Pix	PatchU		TileU		PatchM		TileM				Data	
				TDL	FM	TDL	FM	Linear fusion	Nonlinear fusion	Linear fusion	Nonlinear fusion				
											Att	Gat	Re		Align
Imani et al. ₂₀ [12]	Inf. Fusion	Spectral and spatial fusion for hyperspectral image classification	✓	✓	×	×	×	×	×	×	×	×	×	×	
Rasti et al. ₂₀ [13]	IEEE GRSM	Feature extraction for hyperspectral imagery from Shallow to Deep	✓	✓	×	×	×	×	×	×	×	×	×	×	
Zang et al. ₂₁ [14]	IEEE JSTAR	Land-Use mapping for high-spatial resolution remote sensing image	✓	✓	×	✓	×	×	×	×	×	×	×	✓	
Yuan et al. ₂₁ [15]	Expert Syst. Appl.	Deep learning methods for semantic segmentation of remote sensing imagery	✓	✓	×	✓	×	×	×	✓	×	✓	×	✓	
Ahmad et al. ₂₂ [16]	IEEE JSTAR	Hyperspectral image classification—traditional to deep models	✓	✓	×	×	×	×	×	×	×	×	×	×	
Yu et al. ₂₃ [6]	Remote Sens.	Deep learning methods for semantic segmentation in remote sensing with small data	×	×	×	✓	×	✓	×	✓	×	✓	×	×	
Huang et al. ₂₄ [17]	IEEE JSTAR	Deep-learning-based semantic segmentation of remote sensing images	×	×	×	✓	✓	×	×	✓	✓	✓	✓	✓	
Liu et al. ₂₄ [18]	IEEE JSTAR	Optical and SAR image deep feature fusion in semantic segmentation	✓	✓	×	✓	✓	×	×	✓	✓	✓	✓	✓	
Kumar et al. ₂₄ [19]	Comput. Sci. Rev.	Deep learning for hyperspectral image classification	✓	✓	×	×	×	✓	×	×	×	×	×	✓	
Akewar et al. ₂₅ [20]	IEEE GRSM	An integration of natural language and hyperspectral imaging	✓	✓	✓	✓	✓	×	×	×	×	×	×	×	
Rehman et al. ₂₅ [21]	Artif. Intell. Rev.	Deep learning for HS-LiDAR imagery classification	✓	✓	×	×	×	✓	×	×	×	×	×	✓	
This Survey	–	Deep learning advances in remote sensing semantic segmentation	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	

Pix: Pixel Classification, PatchU: Patchwise Unimodality Classification, TileU: Tilewise Unimodal Segmentation, PatchM: Patchwise Multi-modality Classification, TileM: Tilewise Multi-modality Segmentation, TDL: Traditional DL Model, FM: Foundation Model, Att: Attention Mechanism Fusion, Gat: Gated Mechanism Fusion, Re: Reconstruction Mechanism Fusion, Align: Alignment Mechanism Fusion.
 Journal abbreviations: Inf. Fusion: *Information Fusion*; IEEE GRSM: *IEEE Geoscience and Remote Sensing Magazine*; IEEE JSTAR: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*; Remote Sens.: *Remote Sensing*; Expert Syst. Appl.: *Expert Systems with Applications*; Comput. Sci. Rev.: *Computer Science Review*; Artif. Intell. Rev.: *Artificial Intelligence Review*.

TABLE 2. Comparison of key features across different segmentation paradigms.

Granularity Paradigm	Representative Algorithms	Computational Complexity	Parameter Count	Dataset Type / Scale	Supervision Requirement	Generalization Capability
Pixel-level	MLP, SAE	Very Low	Minimal (<1M)	Hyperspectral / Small	Dense pixel labels	Weak (Prone to local shifts)
Patch-level	2D/3D-CNN	Low (Train) High (Infer)	Light (0.1M - 10M)	HSI, LiDAR / Medium	Central-pixel label	Limited (Lacks global context)
Tile-level	U-Net, DeepLab	Moderate to High	Heavy (10M - 200M)	Optical, SAR / Large	Dense pixel/Weak labels	Strong (Within domain)
Image-level	REST	Extremely High	Massive (>100M)	Global scale	Dense pixel/Weak labels	Strong (Within domain)

With the introduction of AlexNet, deep learning methods represented by convolutional neural networks (CNNs) officially entered the public consciousness [22]. In [27], Hu et al. were the first to employ CNNs with multiple layers in the spectral domain for HSI classification. In addition, RNNs are frequently used to capture temporal dependencies in image time series, improving classification accuracy and reducing model complexity in tasks such as crop classification, where seasonal variations are significant [28].

Pixel-based classification methods marked the introduction of DL into RSI-related computer vision tasks. Although these methods have been surpassed by more recent models in information extraction, training efficiency, and accuracy, they have opened a new direction for future research, highlighting the significant potential of DNNs in RSI analysis.

B. PATCH-BASED CLASSIFICATION

Patch-based classification extracts image features from patches and uses them to predict the central pixel of each

patch on a per-pixel basis. A patch refers to a small region surrounding the central pixel. The successful application of DL models such as CNN, Transformer, and Mamba in image classification has led to the development of various patch-based variants for RSIS [29]. CNNs were introduced into road extraction tasks in [30], where large image patches were used to provide contextual information for small-patch predictions.

To further enhance patch-based classification capability, there are two main directions. The first is to mine more discriminative features from existing data, and the second is to integrate richer information via fusion. To fully exploit the information contained in the data, feature extraction strategies such as multi-scale modeling, joint feature learning, image augmentation, and domain adaptation (DA) have been developed. These methods explore feature representations from different perspectives and improve classification performance across diverse scenarios. To fuse more features, multimodal fusion strategies, such as attention and gating

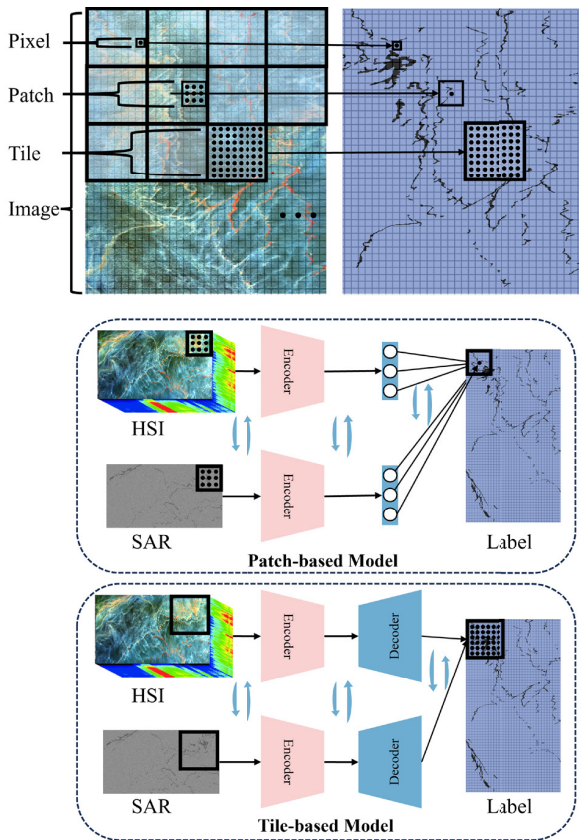


FIGURE 3. Illustration of the proposed pixel-patch-tile-image taxonomy for RIS classification and semantic segmentation. Pixel-based methods perform per-pixel classification, predicting a class label for each pixel independently. Patch-based classification assigns one label to each patch, where a local neighbourhood centred at a target pixel is used as input to predict a single class, and the prediction is then mapped back to the patch centre. Tile-based semantic segmentation processes a larger image tile in one forward pass and outputs a dense segmentation map for all pixels within the tile simultaneously. Image-based segmentation generalizes tile-based inference to arbitrary image sizes and resolutions, producing consistent dense predictions for large-area RSIs.

mechanisms, are employed to regularize and align features across modalities. They aim to obtain more complete object representations and enhance the classification ability for specific targets.

Although these models outperformed traditional approaches, patch-based classification only predicts one label per patch. This leads to two problems: i) the inference process is inefficient. To perform global predictions, each pixel is computed multiple times; ii) the generalization capacity is weak. The limited contextual information in a small patch restricts model performance and weakens generalization capacity. As a result, this approach is particularly suitable for scenarios involving small datasets where individual pixels carry substantial discriminative information, such as hyperspectral or multispectral tasks.

C. TILE-BASED SEMANTIC SEGMENTATION

Tile-based segmentation aims to extract image features by tile and utilize the extracted features to predict the segmentation map end-to-end. Tile-based RSIS methods, primarily built

upon FCN and U-Net architectures, have gained momentum due to the increasing availability of publicly labelled datasets. These models achieve a balance between efficiency and generalization by reducing the number of trainable parameters without compromising segmentation performance [31], [32], [33].

Unlike patch-based methods, which infer pixel labels using patches, tile-based segmentation performs scale-invariant, pixel-level prediction. In this approach, an input tile generates a corresponding output tile, substantially improving computational efficiency and making tile-based segmentation the foundational framework for modern SS methods. Despite these advantages, this paradigm still faces several challenges, including increased computational burden as tile size grows and limited perceptual capability for fine-grained targets. To overcome these challenges and improve model performance, knowledge distillation, fine-tuning, domain generalization, and multimodal data fusion methods have been advanced. Each of these approaches is discussed in the following sections, with technical details and insights, to clarify their contributions to improving tile-based SS.

D. IMAGE-BASED SEMANTIC SEGMENTATION

Image-based segmentation extracts image features from the entire scene concurrently and utilizes these global representations to predict the semantic map of the whole image in a truly end-to-end manner. While tile-based methods dominate current applications, they inherently rely on cropping large-scale RSIs into smaller grids due to GPU memory limitations. This artificial physical cropping disrupts genuine geographical topology, causing edge artifacts at tile seams and forcing the truncation of large-scale continuous targets.

To overcome the suboptimal strategies of cropping and fusion, image-based segmentation has recently evolved toward holistic learning paradigms. Recent breakthroughs, such as REST architecture [10], introduce novel spatial parallel interaction mechanisms. These frameworks are specifically designed to handle whole-scene RSIs directly without any cropping operations. By supporting diverse encoders and decoders in a plug-and-play fashion, such holistic learning architectures enable seamless integration with mainstream segmentation methods and advanced foundation models while bypassing traditional memory bottlenecks. This paradigm shift resolves the performance degradation associated with local cropping, ensuring that the global contextual dependencies of complex geographical landscapes are fully preserved.

Image-based segmentation represents the ultimate progression in the spatial context modeling hierarchy, shifting the field from local fitting to global cognition. Although it provides a comprehensive and seamless understanding of massive scenes, this holistic nature places extreme demands on the scalability of computing architectures.

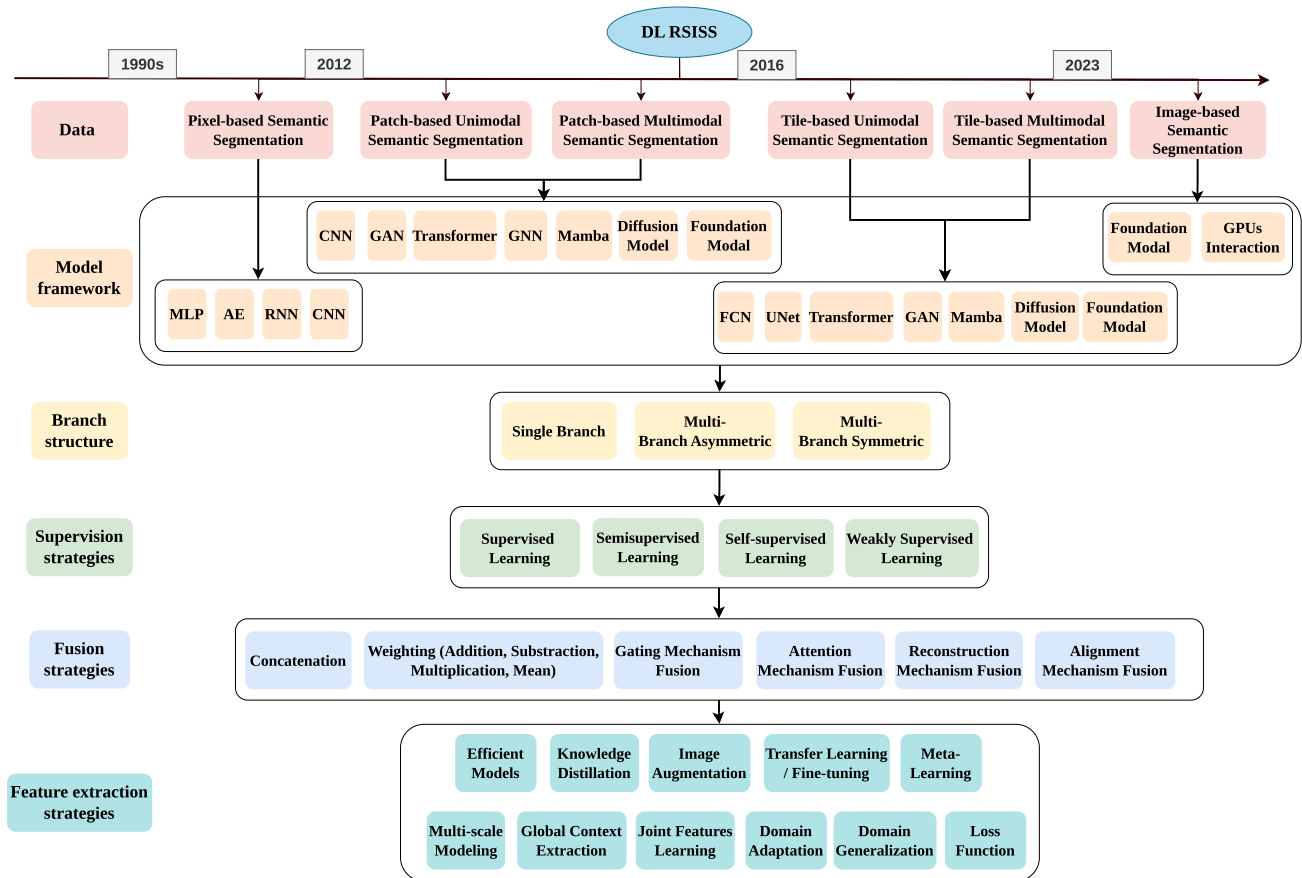


FIGURE 4. Illustration of the data processing, model framework, branch structure, supervision, fusion, and feature extraction strategies used for remote sensing image semantic segmentation.

Moving forward, this approach requires deeper exploration into parameter-efficient fine-tuning (PEFT) and the construction of multimodal universal interfaces, enabling these holistic models to efficiently process ultra-large-scale and heterogeneous imagery without degrading back to tile-wise processing.

To comprehensively evaluate these four structural paradigms, Table 2 summarizes their fundamental characteristics. It explicitly compares the representative algorithms, computational complexity, parameter count, suitable dataset types, supervision requirements, and general generalization capabilities across the pixel, patch, tile, and image hierarchies. This macro-level comparison provides an immediate and intuitive methodological trade-off guide for researchers and engineers.

Furthermore, while this qualitative summary outlines the general theoretical boundaries, comparing absolute quantitative metrics across completely different benchmarks can be inherently misleading if forced into a single dense manuscript table. To provide rigorous empirical evidence without compromising manuscript readability, we have established a continuously updated open-source performance evaluation database on our GitHub repository at RSISS. This repository hosts exhaustive quantitative benchmarks that detail specific accuracy metrics, FLOPs, parameter counts, and processing

speeds for representative models discussed throughout this survey.

III. FEATURE EXTRACTION TECHNIQUES

Over the past decade, RSISS has undergone significant evolution, establishing a comprehensive spectrum of methodologies that range from pixel-level classification to holistic image-level segmentation. Despite the structural differences across these processing granularities, they fundamentally share a core set of underlying feature extraction strategies. The dynamic permutations and combinations of these strategies constitute the bedrock of modern remote sensing image semantic segmentation. Table 3 summarizes the primary feature extraction strategies explored in this field. In this section, we systematically detail the application of these strategies across pixel, patch, tile, and image data processing paradigms. By articulating these shared mechanisms across different hierarchical levels, we aim to inspire future researchers to architect novel feature extraction paradigms and further elevate segmentation performance. Because these strategies are predominantly implemented in complex combinations within the existing literature, our subsequent subsections deliberately highlight only the most representative studies to elucidate the core mechanisms of each specific strategy.

TABLE 3. An overview of feature extraction strategies in remote sensing image semantic segmentation. Based on the surveyed literature, existing feature extraction techniques are further subdivided and organized into several subclasses.

Feature extraction strategies	Patch-based feature extraction references	Tile-based feature extraction references
Multi-scale modeling	Image-level: [34], [35], [36], [37], [38] Feature-level: [39], [40], [41]	Image-level: [42] Feature-level: [43], [44], [45], [46], [47], [48]
Global context extraction	[49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59]	[60], [61], [62], [63], [64], [65]
Joint feature learning and loss function	[38], [49], [51], [57], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76]	[33], [77], [78], [79], [80], [81], [82]
Efficient models	[69], [83], [84], [85], [86]	[47], [61], [65], [77], [87]
Knowledge distillation	Response-based: [88], [89] Feature-based: [90], [91], [92]	Response-based: [93], [94] Feature-based: [95], [96] Relationship-based: [97], [98]
Image augmentation	Transformation-based: [99], [67], [100], [101], [102], [103], [104]	Transformation-based: [105], [106], [107], [108] Generation-based: [109], [110], [111], [112]
Transfer learning and fine-tuning	[113], [114], [115], [116]	[117], [118], [9], [119], [120], [121], [64]
Meta-learning	[122], [123], [124], [125], [126], [127], [128], [129]	[130], [131], [132], [133], [134], [135], [136], [137]
Domain adaptation	Feature-level: [126], [138], [139], [140], [141], [142], [143], [144], [145], [146], [147], [148], [149] Output-level: [34]	Input-level: [109], [110], [111], [112], [150], [151], [152] Feature-level: [60], [108], [153], [154], [155] Output-level: [156], [157], [158], [159], [160], [161], [162], [163]
Domain generalization	Data manipulation: [164], [165], [166], [167] Domain-invariant feature representation: [168], [169], [170], [171], [172] General learning strategy: [173], [174]	Data manipulation: [175], [176] General learning strategy: [63], [177], [178], [179], [180], [181]
Self-supervised learning	Self-predictive generative learning: [173], [182], [183], [184], [185], [186] Discriminative contrastive learning: [102], [103], [104], [186], [187], [188], [189]	Self-predictive generative learning: [190], [191], [192], [193], [62], [194] Discriminative contrastive learning: [105], [106], [190], [194], [195], [196]
Semi-supervised learning	Graph-based: [57], [197] Pseudo-labelling/Self-training: [198], [199] Joint optimization: [67], [200], [201], [202]	Pseudo-labelling/Self-training: [203], [204], [205] Consistency regularization: [3], [79], [107], [205], [206], [207], [208], [209], [210], [211], [212], [213], [214], [215]
Weakly-supervised learning	[216]	[217], [218], [219], [220], [221], [222], [223], [224], [225]

A. MULTI-SCALE MODELING

Object scales vary dramatically in RSIs, and constructing robust multi-scale representations is particularly important. Various DL models strive to extract valuable information for target perception from limited receptive fields. CNNs extract local spatial features, and only through residual networks can the receptive field be gradually expanded to learn more complex spatial patterns [226], [227]. The capsule network (CapsNet) [228] captures positional and directional dependencies between features, thereby providing more detailed feature representations. Additionally, GNNs [57] can capture spatial dependencies from non-Euclidean data

structures, such as graphs, offering an efficient alternative for spatial feature extraction.

However, clear scale differences among various objects in RSIs. A more effective strategy for enhancing feature extraction is to capture spatial dependencies from multi-scale contexts, which enables the model to gather information at various levels, which is particularly important when handling objects of different scales. Common techniques for achieving multi-scale information include pyramid-based feature extraction and multi-branch feature fusion methods, which can be reduced to the image-level [34], [35], [36], [37], [38], [42] and feature-level [39], [40],

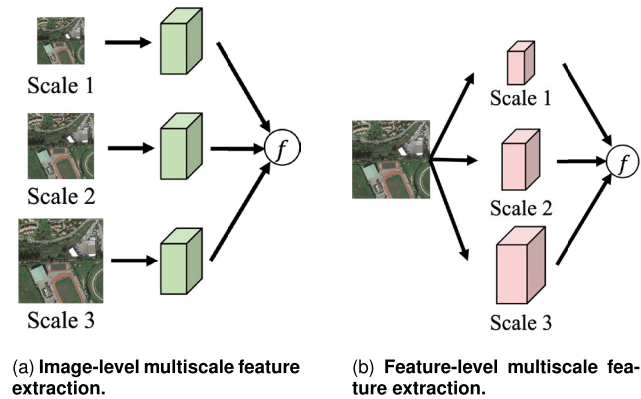


FIGURE 5. Two representative multiscale feature extraction paradigms. (a) Image-level multiscale feature extraction, where the input image is rescaled into multiple resolutions and processed independently before feature aggregation. (b) Feature-level multiscale feature extraction, where multiscale representations are generated within the network from a single-resolution input and subsequently fused.

[41], [43], [44], [45], [46], [47], [48] methods, as depicted in Figure 5.

For patch-wise classification, the relatively small input patch size has led many studies to adopt multi-scale image cropping before feeding data into the model. For instance, Zhao and Du [35] introduce a multiscale CNN with image pyramids and spectral-spatial feature fusion via logistic regression, improving urban classification performance. Gao et al. [36] propose a lightweight multiscale residual network using mixed depthwise separable convolutions and high-level shortcuts to extract spectral-spatial features.

Meanwhile, multi-scale feature fusion is widely used in tile-wise SS. UNet-like models merge features from different depths, which can also be regarded as a form of multi-scale contextual representations [43], [44], [47], [48]. For example, DeepResUnet [48] follows U-Net's encoder-decoder, extracting multiscale features via cascaded downsampling and skip-concatenation, while residual blocks deepen each scale for robust building segmentation. MAREsU-Net [44] replaces plain skip connections in U-Net with multi-stage linear-attention blocks that refine multiscale encoder-decoder features and inject global context during stage-wise fusion in decoding.

Many RSIS methods are directly adapted from computer vision architectures to obtain multi-scale features. For example, D-LinkNet, based on the LinkNet architecture [229], incorporates dilated convolutions to improve road extraction performance [45]. Kemker et al. [46] adapted the SharpMask and RefineNet models to process MSI data. The atrous spatial pyramid pooling (ASPP) has also been adopted to capture RSI features at multiple spatial scales [42].

In summary, multi-scale modeling has emerged as a simple yet effective feature extraction strategy and is now regarded as a fundamental design principle in modern RSIS architectures. Many networks implicitly incorporate multi-scale feature extraction in their design, even when it is not explicitly highlighted. However, multi-scale designs often

introduce substantial computational overhead. Consequently, developing lightweight multi-scale feature extraction mechanisms while maintaining strong segmentation performance remains a key direction for future research.

B. GLOBAL CONTEXT EXTRACTION

For humans, objects are categorized based on their shape, texture, and other macroscopic features, where contextual information plays a crucial role in determining object categories. In RSIS models, global context is typically captured through (i) pooling-based modules and (ii) attention mechanisms.

While global pooling offers a simple and efficient solution, it often leads to the loss of small targets due to aggressive downsampling. In contrast, attention mechanisms provide a more flexible and effective means to capture long-range dependencies by assigning adaptive weights across the image space. It enables the model to focus selectively on relevant parts of the input, thereby improving the capture of dependencies and contextual relationships. The most widely used attention strategies include channel attention, spatial attention, and self-attention mechanisms [49], [50], [60], [61]. Figure 6 presents an overview of the channel and spatial attention mechanisms, while Figure 7 illustrates the self-attention mechanism framework.

For high-dimensional data, increasing studies have been given to integrate both spectral and spatial attention for feature extraction [51], [52], [53], [55], [56]. Mei et al. [51] employed RNNs with attention to capture intrinsic spectral correlations. Ma et al. [52] proposed a dual-branch structure to extract spectral and spatial features separately, applying distinct attention mechanisms in each branch. This parallel configuration allows for the independent optimization of complementary feature sets prior to fusion [54]. In [55], spatial and spectral attention modules are arranged in a cascaded structure within a residual block framework, a design well-suited for refining features progressively across network layers [56].

In high-resolution data, global feature extraction is crucial for capturing long-range spatial dependencies that CNNs alone often miss. Wang et al. [61] introduced a bilateral design, where a CNN branch preserves fine-grained local details and a Transformer branch models long-range interactions to handle large appearance variations in very fine-resolution urban scenes. In a related direction, MBATA-GAN [60] adopted mutually boosted attention to strengthen cross-domain global interactions over high-level features. More recently, RSIS increasingly relies on Transformer-centric pipelines and foundation models to enhance global context modeling. GLOTS [62] leverages a pretrained ViT encoder and a global-local attention decoder to inject global cues into multi-scale representations, while FreeMix [63] builds on a pretrained foundation backbone and a Mask2Former-style Transformer decoder to propagate global semantics to dense masks under domain shifts. MFNet [64] further demonstrates that fine-tuning

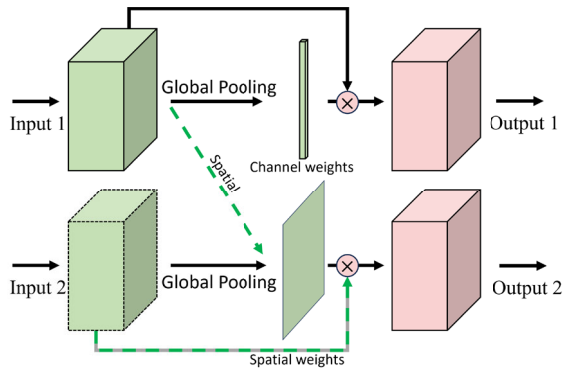


FIGURE 6. Illustration of attention mechanisms in feature enhancement. The top branch shows channel attention, where global pooling generates channel-wise weights to recalibrate feature responses. The bottom branch shows spatial attention, where spatial importance maps are computed to emphasize informative regions. The green dashed lines indicate cross-attention interactions between channel and spatial representations.

SAM-style ViT encoders and coupling them with explicit global channel aggregation can provide strong global priors for high-resolution multimodal segmentation.

In addition, the Mamba network has been adopted in various computer vision tasks for its efficiency in integrating both global and local contextual information [58], [59]. It achieves this without incurring the high computational cost typically associated with traditional self-attention mechanisms [65]. Mamba belongs to the family of selective state space models, which model long-range dependencies via recurrent state updates and input-dependent filters. In principle, each token's state summarizes information from all previous tokens, leading to a global receptive field in the sequence dimension.

C. JOINT FEATURE LEARNING AND LOSS FUNCTION

By designing specific neural network architectures and incorporating customized loss functions, networks can be optimized to extract multiple types of information in a targeted manner, such as spatial, spectral, local, and global information. Joint feature learning has been shown to significantly enhance segmentation performance.

At the data level, HSIs provide both spatial and rich spectral features [68], [69], [71], while polarimetric SAR imagery enables decomposition into polarization and spatial components [70]. Zhong et al. [71] incorporate CNN structures to extract spectral and spatial features from HSIs. Gao et al. [70] constructed a dual-branch deep CNN where polarization features were extracted from a six-channel real matrix and spatial features from a Pauli RGB image. More generally, heterogeneous network architectures are often employed to extract joint features from separate pathways [49], [51], [72]. For example, Zhang et al. [66] used a 1-D CNN to extract spectral features and a 2-D CNN to extract spatial features. SSUN [38] utilised LSTM networks for spectral feature extraction and a multi-scale CNN for spatial features.

At the feature level, CNNs are effective in capturing local patterns, while attention mechanisms extract global

context [75]. Jia et al. [73] proposed 3-D Gabor-modulated kernels to characterize the internal spatial-spectral structure of HSI data from various perspectives. Roy et al. [74] proposed a morphFormer that implements a learnable spectral and spatial morphological network. The morphological convolution operations enhanced the interaction between shape and structural information. In addition, semantic labels convey category-level information, and edge labels provide spatial boundary details [76]. The integration of these features helps to delineate semantic regions more precisely, improving segmentation accuracy. Song et al. [76] applied a Laplacian filter to extract edge features as an explicit supervisory signal, incorporating a segmentation head and edge decoder into the network to jointly learn semantic and boundary features and enhance generalisation.

The loss function measures the discrepancy between the model's predicted output and the ground truth. Modifying the loss function can guide the model's optimization, enabling it to learn specific features that aid segmentation. Commonly used loss functions in SS include cross-entropy loss, dice loss, focal loss, and infoNCE loss [33], [77], [78], and their weighted summation [57], [67]. Specialized loss functions are integral to constructing semi-supervised learning (SeL), self-supervised learning (SSL), and multi-task learning frameworks, enabling the model to extract meaningful information from data. In these cases, losses such as adversarial loss, cycle consistency loss, perceptual loss, local consistency loss, and global diversity metric loss are commonly employed [79], [80]. Additionally, edge loss is frequently used as an auxiliary objective to enhance boundary delineation [81], [82]. This strategy aids in detecting small objects and distinguishes between classes with similar shapes.

D. EFFICIENT MODELS

Due to differences in learning styles, such as recurrent, transductive, and inductive learning [230], DL architectures vary significantly in time or space requirements. Technically, RNNs, which follow a recurrent learning style, are generally slow during training due to their loop structure but offer fast inference once trained [83]. SeL GNNs operate under a transductive learning paradigm. Their shallow architecture enables fast training and inference, though they require considerable memory for processing [57]. An efficient GNN such as FDGC [85] used inductive style training, achieving a balance between speed and efficiency, but lost the ability to extract global non-Euclidean distance relationships. CNNs, Transformers, and Mamba models are usually used as inductive models. Although these models require more time per training unit, they can achieve convergence using only tens or hundreds of samples per class in patchwise classification tasks, often resulting in overall faster training [84].

In pixel- and patch-level classification, overlapping patches from neighbouring pixels introduce significant redundancy in computation, thereby inevitably extending test time [69], [86], [231]. In tilewise segmentation, the

scale invariance of inputs and outputs addresses the issue of redundant computations during the inference phase [7]. However, recent advanced DL models for segmentation frequently involve complex architectures and require large training datasets, which result in substantial computational demands. As a result, computational efficiency in tile-based segmentation primarily concerns the number of parameters and training speed.

To construct efficient models, lightweight modules such as depthwise convolution, pointwise convolution [231], linear attention [47], bilateral segmentation networks [87], and Mamba modules [65] are widely adopted. These components help reduce computational cost while having only a minor impact on segmentation accuracy or none at all. For example, linear attention mechanisms have been integrated into lightweight bilateral contextual networks to improve computational efficiency [47], [61], [77]. CM-UNet [65] combines a CNN-based encoder with a Mamba-based decoder to efficiently extract and integrate both local and global features for RSIS. This design improves segmentation performance while maintaining a low computational footprint.

E. KNOWLEDGE DISTILLATION (KD)

Among the efficient building blocks for deep models, KD has been proposed as an effective solution for transferring and refining the knowledge of large models to small models [232]. KD typically enables lightweight student networks to mimic the behaviour of larger, well-trained teacher models, achieving competitive or even superior performance with lower complexity. These methods offer practical pathways to reduce model complexity while maintaining accuracy and lay the foundation for more robust models in scenarios with limited annotations or dynamic domain shifts. KD has evolved from basic logit matching to adaptive, interpretable, and lifelong knowledge transfer, including response-based [88], [89], [93], [94], feature-based [90], [91], [92], [95], [96], and relationship-based [97], [98] approaches based on the different knowledge categories of teacher models.

1) RESPONSE-BASED KD

One major line of work focuses on response-based KD, in which soft output logits from the teacher guide the student [88], [89]. For example, Chi et al. [88] proposed SSKD, a self-supervised framework that generates soft labels for unlabeled HSI patches using spectral-spatial similarity, effectively leveraging large-scale unlabeled data for supervision. In DSCT, Dong et al. [93] proposed a hybrid CNN-Transformer distillation framework with a novel target-nontarget KD strategy that explicitly guides decision boundary refinement. Likewise, MSTNet-KD [94] employs multilevel output alignment to bridge deep decoder layers between student and teacher networks.

2) FEATURE-BASED KD

Feature-based KD captures structural representations from internal layers of the teacher model. Shi et al. [90]

proposed an explainable scale distillation network that transfers multi-scale information from a complex teacher to a single-scale student. The distilled knowledge preserves both the discriminative power and interpretability of multi-scale representations while reducing computational cost. Zhao et al. [91] further extended KD to lifelong learning by proposing a continual spectral-spatial feature distillation strategy that maintains knowledge across sequential HSI tasks without catastrophic forgetting. Complementary to this, Li et al. [92] developed HyperKD, which combines exemplar replay with cross-spectral-spatial KD, allowing the student model to inherit not only output predictions but also spectral and spatial distributions from previous tasks. In STONet-S, Zhou et al. [95] introduced frequency-aware KD using discrete cosine transforms to decompose and transfer high- and low-frequency components, preserving both edge and semantic information. Additionally, Sun et al. [96] addressed the robustness of KD under weak supervision. The proposed BAKD framework combines boundary-aware and uncertainty-weighted distillation to reduce the impact of noisy annotations, especially near semantic edges.

3) RELATIONSHIP-BASED KD

Relationship-based KD has also gained traction. Graph-aware methods such as GAGNet-S transfer topological context by encoding cross-scale and inter-pixel dependencies [97]. Meanwhile, AKD [98] strategies in transformer-based dual-path networks enhance student learning by minimizing the angle between teacher-student feature vectors across channels. Overall, KD research in RSIS demonstrates a clear trend toward structure-aware, frequency-guided, and uncertainty-adaptive strategies to enhance generalization while reducing model complexity.

F. IMAGE AUGMENTATION

The primary function of image augmentation is to artificially increase the size and diversity of training datasets, thereby improving model generalization and robustness. Since collecting and annotating large-scale datasets is both costly and time-consuming, image augmentation generates additional samples by perturbing existing images while preserving their semantic labels. This helps segmentation models cope with variations in environment, illumination, sensor characteristics, and viewpoint, without sacrificing pixel-level detail.

1) TRANSFORMATION-BASED IMAGE AUGMENTATION

From a transformation-based perspective, augmentation applies predefined geometric or radiometric operations to existing images. Typical choices include random flipping, cropping, rotation, jittering, and random erasing, which introduce local variations but keep the global scene structure unchanged [99], [100], [101]. For modality-specific data such as HSI and LiDAR, specialized operations have been designed, for example, adding band-wise noise, simulating virtual samples via random spectral scaling [99], or selecting

unlabeled hyperpixels around labelled seeds to densify training sets [101]. In self-supervised learning, such transformations are further organized into pairs or groups to construct positive and negative views for contrastive objectives, making augmentation an integral part of representation learning rather than a mere pre-processing step [67], [94], [102], [103], [104]. In consistency regularization, SeL and KD, pairwise images generated through data augmentation can also guide models to learn invariant feature representations and complementary knowledge [93], [107], [108]. For example, ECAE [107] uses color jitter and Gaussian blur perturbations to enforce consistency when learning from unlabeled data, improving pseudo-labels and boundaries. DSCT [93] trains distilled CNN/transformer students with random resizing, cropping, flipping, and photometric distortions, boosting robustness and generalization across datasets.

2) GENERATION-BASED IMAGE AUGMENTATION

Beyond these handcrafted transformations, generation-based augmentation leverages generative models to synthesize new training images or re-style existing ones. In RSIS, this is often coupled with unsupervised DA: the generator maps labeled source images into target-style images while preserving semantic structure, effectively creating additional supervision for the target domain [109], [110], [111], [112]. Benjdira et al. [109] demonstrated this idea by training a CycleGAN-based image-to-image translation network. Tasar et al. [110] further simplified the generative component with ColorMapGAN, which learns a global color mapping between source and target images. This design produces “fake” training images that are semantically identical to the originals yet spectrally aligned with the test domain. Ji et al. [111] generalized this line of work with a full-space DA network, in which adversarial learning is performed jointly across the image, feature, and output spaces.

Overall, transformation-based and generation-based augmentation are complementary: the former enriches local appearance and geometric diversity around a fixed data distribution, whereas the latter synthesizes target-style or domain-bridging samples that explicitly reshape the training distribution. For RSIS, especially in cross-city or cross-sensor scenarios, GAN-driven, generation-based augmentation is a powerful tool for narrowing domain gaps while reusing existing pixel-wise annotations, thereby enhancing the practicality of deep segmentation models in real-world deployments.

G. TRANSFER LEARNING AND FINE-TUNING

Transfer learning refers to the reuse or adaptation of a pre-trained model for new tasks. Rather than training a model from scratch, transfer learning enables the application of learned knowledge, significantly reducing training time and resource demands, particularly when limited data are available for the target task [113], [114], [115], [116]. For instance, Zhong et al. [116] integrated transfer learning with DA to reduce distributional discrepancies across scenes

and sensor types, proposing a cross-scene transfer learning approach that performs well even with a small number of labelled samples. One issue in transfer learning is the domain shift between the original data and fine-tuning data. To overcome this, He et al. [115] introduce a mapping layer that projects the HSI to 3-D features and then uses a pretrained CNN to initialize the network.

The rapid emergence of large-scale VFMs, such as ViT-based backbones and the Segment Anything Model (SAM) [233], has fundamentally changed how transfer learning is applied in RSIS. Directly fine-tuning all parameters of these models on relatively small RS datasets is often inefficient and prone to overfitting, due to the large number of parameters, high memory cost, and substantial training time. As a result, PEFT has become increasingly favored over conventional full-model fine-tuning. Instead of updating the entire backbone, PEFT freezes most pre-trained weights and inserts lightweight, task-specific modules, such as adapters, low-rank adapter branches, or multi-level feature adaptation blocks, thereby improving both accuracy and efficiency [9], [64], [117], [118], [119], [120], [121].

Recent SAM-based RSIS methods exemplify this shift. Multi-LoRA [118] fine-tuned SAM frameworks for urban man-made object extraction, which introduces multiple LoRA modules into the SAM encoder and decoder, enabling accurate road and building extraction while updating only a small fraction of parameters. Other works propose multi-level feature adaptation of SAM for high-resolution land use/land cover classification [121], or design unified multimodal fine-tuning frameworks that plug multimodal fusion heads on top of frozen VFMs to jointly exploit optical, SAR, and LiDAR data [64].

Overall, the role of transfer learning in RSIS is evolving from classical “pretrain on natural images + full fine-tuning” toward adapter- and PEFT-driven fine-tuning of foundation models. Traditional transfer learning remains effective for moderate-sized CNNs, but for large VFMs, it is increasingly replaced or complemented by PEFT strategies that offer a better trade-off between accuracy, efficiency, and generalization across diverse RS scenes and modalities.

H. META-LEARNING

Meta-learning enables models to adapt quickly to new tasks by training on a variety of related tasks. The objective is to learn a generalizable strategy that enables the model to perform well on new tasks with minimal data and few training iterations [234].

In HSI classification, meta-learning has become a central tool for tackling small-sample and cross-domain settings. Early works employed residual 3-D CNNs and episodic training to learn a discriminative metric space in which query pixels are classified by nearest neighbors to class prototypes [122]. To further reduce dependence on annotated sources, unsupervised meta-learning with multiview constraints has been proposed, which constructs tasks

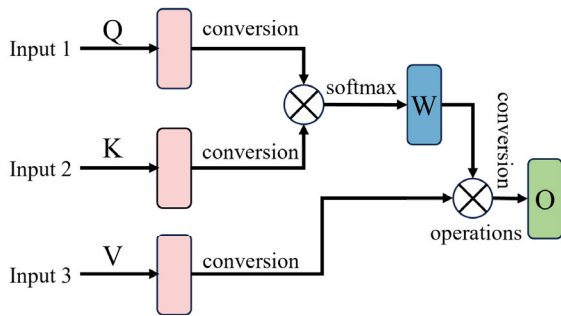


FIGURE 7. Self-attention and cross-attention mechanisms based on the query-key-value (Q-K-V) framework. Query (Q), key (K), and value (V) features are first projected through linear transformations. Attention weights are computed via similarity between Q and K followed by a softmax operation, and are then applied to V to produce the output representation. In self-attention, Q, K, and V originate from the same feature source, whereas in cross-attention, they are derived from different feature sources or modalities.

from unlabeled HSIs and trains relation networks to learn transferable metrics [123]. Building on this foundation, a series of cross-domain few-shot HSI classification methods explicitly integrate DA into the meta-learning pipeline, including deep cross-domain metric learning with adversarial alignment [124], contrastive few-shot learning that enhances inter-class separation in highly confounded spectral spaces [125], graph-based DA that leverages graph convolutions over spectral-spatial prototypes [126], and multi-stage relation networks with dual metrics to improve class representation under extremely scarce labels [127]. More recent transformer-based and distillation-based approaches further push this line of research by injecting spectral coordinate priors into asymmetric encoder-decoder architectures and by decoupling meta-knowledge distillation across domains [128], [129], yielding strong cross-domain few-shot HSI classification performance.

In parallel, meta-learning has been extended from patch-level classification to tile-level SS in RSIs, particularly in few-shot settings. A first step was to formulate RSIS as a prototype-based few-shot segmentation approach, in which deep metric learning is used to extract pixel embeddings, construct class prototypes from support masks, and segment query images via nearest-prototype matching in the embedding space [130]. Subsequent works refine this metric-based meta-learning paradigm to better accommodate the unique characteristics of high-resolution aerial and satellite images [131], [132], [133].

Overall, existing meta-learning studies in RS are better viewed as several complementary research streams rather than a single linear trajectory. One stream focuses on task-level episodic meta-learning for HSI, where models are trained to rapidly adapt to new spectral classes or scenes under scarce labels. A second stream emphasizes metric- and optimization-based few-shot learning with explicit cross-domain robustness, in which adversarial alignment, graph modeling, transformers, and knowledge distillation are integrated into the meta-learning pipeline to handle shifts across sensors, regions, and acquisition conditions.

A third stream extends these ideas from classification to segmentation, yielding few-shot segmentation frameworks tailored to high-resolution RGB and multispectral RSIs. Despite the diversity of streams, these works share a common core meta-learning principle: by simulating many support-query episodes during training, the model acquires an inductive bias that enables rapid adaptation to new categories, domains, and geographic regions from only a handful of annotated examples.

I. DOMAIN ADAPTATION (DA)

Due to sensor nonlinearities, seasonal variation, or weather differences, it is common for training and test sets to be collected under different conditions. The discrepancy between the data distribution of the source domain (SD) and the target domain (TD) is known as a distribution shift or domain gap. DA, typically in the form of unsupervised domain adaptation (UDA), aims to leverage data from both the SD and TD during training to learn shared knowledge across the input, feature, and output spaces, reducing the domain gap and improving model performance on the TD [109], [110], [138], [139]. As illustrated in Figure 8, DA strategies for SS can be roughly categorized into three levels: input-level, feature-level, and output-level adaptation. Input-level adaptation directly reduces distribution gaps in the image space, feature-level adaptation encourages learned representations to be domain-invariant, and output-level adaptation aligns prediction maps or pseudo-labels between the SD and TD.

1) INPUT-LEVEL ADAPTATION

Input-level DA typically relies on style transfer or image translation to make the source and target images more similar. Many GAN-based approaches have been proposed to transform images from either the SD or TD [109], [110], [111], [112], [150], [151], [152]. Among them, CycleGAN is one of the most widely used methods. It provides a bidirectional image-to-image translation framework that facilitates the transfer of knowledge from a labeled source domain to an unlabeled target domain [109], [111]. Tasar et al. [150] further explored realistic scenarios with multiple source domains exhibiting distinct distributions, and subsequently proposed ColorMapGAN, a simplified model that only transfers color information from training to test images to better preserve the geometric structure of RSIs [110]. In addition to DL-based techniques, classical methods such as the Wallis filter have also been employed for radiometric alignment at the input level [152]. However, due to the limited number of training samples and the small spatial extent of image patches, input-level adaptation has received relatively little attention in patchwise RSIS, where learning stable translation models remains challenging.

2) FEATURE-LEVEL ADAPTATION

Feature-level DA aims to learn domain-invariant representations by encouraging the alignment of feature distributions

between the SD and TD. This is typically achieved through metric-based losses combined with backpropagation and can be implemented via adversarial learning, explicit metric alignment, or reconstruction- and disentanglement-based strategies [138], [139], [140]. Commonly used metrics include adversarial loss, covariance and parameter losses, maximum mean discrepancy (MMD), contrastive domain discrepancy, Wasserstein distance, and cosine distance [60], [108], [153], [154].

Adversarial-based DA leverages a domain discriminator to encourage the model to extract features that are indistinguishable between the SD and TD, thereby learning domain-invariant and discriminative representations. Typical adversarial DA networks consist of a feature-extraction subnetwork and a domain-discriminator subnetwork. For example, the two-branch attention adversarial DA network proposed by Huang et al. [141] employs a dual-branch attention feature extractor to capture spectral-spatial attention features, together with a discriminator containing two classifiers. Xu et al. [126] further introduced a graph-guided DA few-shot learning method, which combines graph neural networks and adversarial training after feature extraction to construct a graph-structured domain discriminator.

Metric alignment explicitly constrains the feature distributions of the SD and TD using distance-based objectives. Liu et al. [143] align the conditional distribution of each class by combining a class-wise domain adversarial neural network with MMD, using pseudo labels of target data during optimization. PCDM-UDA proposed a multi-view unsupervised DA method that integrates pseudo-class distribution-guided label correction, phase-based domain-invariant features, and trusted prediction to enhance cross-scene HSI classification [144]. Optimal transport-based methods learn a transport matrix to map the source distribution to the target distribution and achieve feature alignment by minimizing the transport cost [145]. For instance, Zhang et al. [138] proposed TsTNet, a topological structure and semantic information transfer network that incorporates topological priors for cross-scene classification.

Reconstruction-based and feature-disentanglement-based DA reduce domain discrepancy by decomposing representations into domain-shared and domain-specific components. GANs and their variants are widely applied in HSI DA to implicitly enforce reconstruction constraints [146]. Ye et al. [147] form a cross-domain mapping chain by connecting multiple CycleGANs in series; mapping errors are accumulated and backpropagated in each cycle, significantly improving the accuracy of cross-sensor HSI mapping. Another line of work focuses explicitly on disentangling domain information. The FDDAN model proposed by Li et al. [148] extracts domain-shared and domain-specific features through a feature disentanglement network and reconstructs the input to constrain the effectiveness of the separation. Similarly, the S4DL framework by Chen et al. designs a spectral-spatial channel mask to separate domain information, ensuring that alignment between the SD and TD is conducted only on

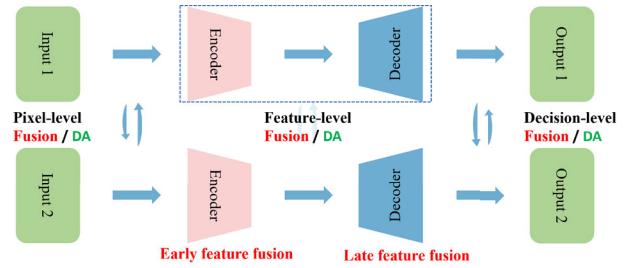


FIGURE 8. Unified remote sensing image semantic segmentation framework incorporating fusion and domain adaptation (DA) at multiple levels. Fusion can be performed at the pixel level (input fusion), feature level (early or late feature fusion within encoder-decoder architectures), or decision level (output fusion). DA can be applied at the pixel, feature, or decision stages to enhance cross-domain robustness.

channels that are free from domain bias [149]. Beyond HSI, Zhang et al. [155] introduced a layer alignment strategy using covariance and parameter losses to mitigate domain shift in high-resolution RSIs. Wu and Wang [108], [154] proposed a two-stage UDA framework that combines adversarial learning and self-training for optical RSIs. Lu et al. [153] presented an end-to-end global-local alignment mechanism with dynamically adjusted adversarial weights, and Ma et al. [60] incorporated attention mechanisms into GAN-based UDA frameworks to capture long-range dependencies between high-level features from different domains.

3) OUTPUT-LEVEL ADAPTATION

Output-level DA focuses on aligning prediction maps or pseudo labels to reduce domain shift. This is typically achieved through adversarial learning [156], [157] or self-training [158], [159]. In adversarial learning, a domain discriminator is applied to the prediction space to encourage domain-invariant and confident outputs. Zheng et al. [156] and Chen et al. [157] employed entropy-guided adversarial models that emphasize low-confidence regions in the TD during adaptation. Self-training methods iteratively refine predictions in the TD using a model trained on the SD. Inspired by DAFormer [160], Li et al. [158] proposed a Transformer-based self-training framework that incorporates gradual class weighting and local dynamic quality estimation to enhance UDA. In [159], self-training is used to iteratively update pseudo labels and improve cross-domain segmentation performance. Combining adversarial learning with self-training can further improve DA performance by leveraging high-quality pseudo labels [161], [162], [163]. For example, Ma et al. [162] introduced a strategy based on local consistency and global diversity metrics to strengthen adaptation in RSIs. Tong et al. [34] proposed an output-level adaptation method based on pseudo-labelling and a retrieval-based sample selection strategy.

J. DOMAIN GENERALIZATION (DG)

Unlike DA, which typically assumes access to labelled SD data and unlabeled TD data, DG tackles the more challenging setting in which models are trained only on one or multiple

SD datasets and must generalize to unseen TD data. DG, as one of the ultimate goals of RS feature extraction, existing methods can be broadly grouped into three categories: (i) data manipulation strategies that generate diverse auxiliary domains or styles [164], [165], [166], [167], [175], [178], (ii) representation learning methods that explicitly enforce domain-invariant feature representations [168], [169], [170], [171], [172], and (iii) general learning strategies that design robust optimization objectives or training pipelines to improve out-of-domain performance [63], [173], [177], [178], [179], [180], [181].

1) DATA MANIPULATION

In patch-based classification, most existing work focuses on data manipulation. Data manipulation denotes strategies that perturb or extend the SD data distribution to approximate potential domain shifts. This can be implemented through various data augmentation techniques, such as texture or style randomization and frequency-based style mixing, or by learnable generators that map SD images to one or more extended domains. In contrast to DA, where image translation is guided by real TD data, data manipulation in DG is performed without TD supervision and instead exposes the model to a range of synthetic domain variations to facilitate the learning of domain-invariant representations. For example, Zhang et al. [164] designed a morph encoder to learn domain-invariant morphological knowledge during domain expansion, highlighting that DG can be more practical than traditional DA when annotated TD data are unavailable. Zhao et al. [165] proposed a symmetric encoder-decoder to construct extended domains, combined with adversarial contrastive learning to obtain domain-invariant features. S^2 ECNet [166] simulates spectral and spatial deviations from the TD via two independent branches, and further introduces a causal alignment module with contrastive learning to extract causal invariant features and mitigate the bias caused by direct feature alignment. Recent work on dynamic style manipulation further enriches this line by generating structurally constrained auxiliary domains for cross-scene HSI classification [167].

Given the persistent shortage of labelled data, data manipulation again emerges as an efficient and straightforward strategy to improve model generalization in tile-based segmentation. Specifically, Iizuka et al. [175] proposed FOSMix, a frequency-based augmentation technique that blends image styles in the frequency domain while preserving semantic content, thereby enhancing domain diversity. Liang et al. [176] introduced CCDR, an SDG method that employs randomized texture and style transformations together with a class-aware consistency constraint, achieving strong generalization with a simple training pipeline.

2) DOMAIN-INVARIANT FEATURE REPRESENTATION

A second line of work targets DG by explicitly learning domain-invariant feature representations. In HSI

classification, language-aware DG networks such as LDGNet [168] leverage large-scale image-text models with dual-stream encoders, aligning visual and linguistic features in a shared semantic space via supervised contrastive learning to obtain cross-domain-invariant representations. Semantic-aware co-parallel architectures such as SCPNet [169] employ dual-branch spatial-spectral and multiscale modules, along with an optimized semantic space, to collaboratively extract domain-invariant features. Frequency- and geometry-based models such as FDGNet [170] further disentangle frequency components and exploit data geometry on Riemannian manifolds to suppress domain-specific artefacts while preserving discriminative structure, whereas invariant semantic DG shuffle networks refine the semantic space by shuffling and aligning semantic components across domains [171]. At the multi-source level, MS-CDG [172] combines data-aware adversarial augmentation with model-aware multi-level diversification and distribution-consistency alignment to jointly encourage domain-invariant representation learning across heterogeneous sources. Apart from HSI classification, similar principles of learning style-robust and semantically consistent representations are also adopted implicitly in RGB and multispectral images, where they are usually integrated into the backbone design or loss formulation and are therefore discussed together with general DG training strategies below rather than as a separate category.

3) GENERAL LEARNING STRATEGY

Several works focus on general learning strategies that couple data- and feature-level DG in a unified training pipeline [63], [173], [174], [177], [178], [179], [180], [181]. StyleMap [177] regularizes both feature-space geometry and style diversity by combining style mapping with a prototypical contrastive loss, providing a theory-driven DG framework. A second group of works builds on vision or geospatial VFMs: GeoSA+BaSA [178] and sensor-agnostic DG [179] couple PEFT with batch-wise style augmentation or generative pre-training to obtain sensor- and scene-agnostic representations, while CrossEarth [181] targets remote-sensing DG by combining an Earth-Style Injection augmentation pipeline with multi-task training that jointly learns semantic segmentation and masked image modeling on a shared DINOv2 backbone, and it validates this design on an RSDG benchmark covering 28 cross-domain settings, showing strong cross-scene generalization. Beyond generic land-cover mapping, CRLMDG [180] introduces causal representation learning and multi-target consistency for fine-grained landslide mapping under complex domain shifts. HyperSIGMA [173] and SpectralX [174] share a Transformer core and self-supervised masked reconstruction pretraining on large spectral corpora to learn transferable spatial-spectral representations. They explicitly address HSI redundancy via specialized attention/tokenisation and add lightweight adaptation components to support downstream classification and generalize to unseen regions, seasons, and tasks. FreeMix [63] extends DG to an open-vocabulary setting

by combining entity-mask mixing with RS-specific self-supervised initialization to jointly handle unseen domains and unseen classes.

K. SELF-SUPERVISED LEARNING (SSL)

SSL, a powerful model pre-training method, enables models to extract useful knowledge from unlabeled data by designing pretext tasks whose supervision signals are automatically derived from the data itself [235]. After training on large-scale unlabeled RSIs, the learned representations can be adapted to downstream classification or segmentation tasks using only a small number of labelled samples. In RS, SSL methods for SS can be broadly grouped into two primary sub-directions: (i) self-predictive generative learning methods, which reconstruct or predict masked/perturbed content [62], [173], [182], [183], [184], [185], [186], [190], [191], [192], [193], [194], and (ii) discriminative contrastive learning methods, which learn to distinguish similar from dissimilar representations [102], [103], [104], [105], [106], [186], [187], [188], [189], [190], [194], [195], [196].

1) SELF-PREDICTIVE GENERATIVE LEARNING

Early SSL for patchwise SS mainly relied on AE frameworks. AE models were extended to CNN architectures, where Conv–Deconv networks learn spectral–spatial features in an unsupervised manner and then require only a small number of labels to achieve strong supervised classification performance [182], [183]. However, the compressed features obtained by AEs are not always optimal for downstream tasks, motivating richer self-predictive objectives. Subsequent works designed pretext tasks that explicitly exploit the inherent relationships within RSIs, such as image inpainting, transformation prediction, and solving jigsaw puzzles [190]. These innate relationship prediction tasks encourage the model to understand local structural coherence, but often focus on narrow and specific cues, limiting their generality.

Masking-based learning has therefore gained increasing popularity [236]. In this paradigm, parts of the input are intentionally hidden, and the model is trained to reconstruct the missing content, which helps capture both local and global dependencies and uncover contextual and structural relationships in the data. Originally proposed in denoising AEs [237], masking-based SSL evolved through BEiT [238] and matured in masked autoencoders (MAE) [236]. In patchwise classification, MAEST [184] was the first to integrate the masking–reconstruction strategy of MAE with spectral–spatial feature extraction. HyperSIGMA [173] pre-trained with MAE on the large HyperGlobal-450K dataset. It separately pre-trains spatial and spectral subnetworks, enabling strong cross-scene transfer across diverse high- and low-level HSI tasks. To further explore the potential of SSL for HSI analysis, MSST [185] constructed a large EnMAP-based dataset and used masked image reconstruction to pretrain Transformer models, yielding robust spectral–spatial representations. For tilewise SS,

self-predictive SSL is also widely used. Methods such as RingMo and SatMAE design RS-specific masking patterns and reconstruction targets, enabling ViT-like backbones to learn generalizable features from large-scale multi-sensor RSIs [191], [192]. Recent consistency-regularized masking strategies further improve robustness by enforcing prediction consistency under different masking configurations [62], [193].

2) DISCRIMINATIVE CONTRASTIVE LEARNING

Discriminative SSL aims to learn the similarities and differences between data samples, typically by constructing positive and negative pairs and optimizing a contrastive objective. In patchwise classification, contrastive learning has become a mainstream approach for unsupervised feature extraction. Data augmentation is the most common means for creating positive/negative pairs [187], [188]. Liu et al. [103] employed a multi-view approach by grouping HSI bands to form sample pairs. Cao et al. [186] used two separate AEs to generate sample pairs and built a ProtoNCE-based contrastive learning model. XDCL [189] introduced a cross-domain discrimination task that leverages both spatial and spectral information to extract shared representations. Graph-based methods further exploit the relational structure of HSI data: graph augmentation modules generate paired graphs for contrastive learning in GCNs, leading to domain-invariant and topology-aware features [102], [104].

In tilewise SS, contrastive SSL has also been widely adopted. Earlier works [190], [195] trained encoders using classical instance-discrimination frameworks, whereas later methods proposed multi-granularity contrastive objectives to capture information at different levels. Li et al. [105] jointly extracted local and global representations, Muhtar et al. [106] focused on both pixel-level and image-level representations, and Dong et al. [196] combined instance-level and semantic-level contrast. Despite their effectiveness, contrastive methods generally rely on complex data augmentation pipelines and meticulous design of positive and negative pairs, which can make them less flexible and harder to tune than purely masking-based approaches.

To exploit the complementary strengths of generative and discriminative SSL, hybrid strategies have been proposed. Cao et al. [194] introduced a Transformer-based framework that integrates contrastive and masked learning, enabling both pixel-level feature learning and global spectral–spatial representations. This hybrid design consistently outperforms models that rely on a single learning paradigm.

L. SEMI-SUPERVISED LEARNING (SeL)

SeL provides a balanced and effective alternative to purely supervised or purely unsupervised schemes by integrating task-specific knowledge from supervised learning and task-agnostic knowledge from unsupervised learning. SeL is particularly attractive in RS because labeled RSIs are scarce and expensive to acquire, whereas unlabeled data

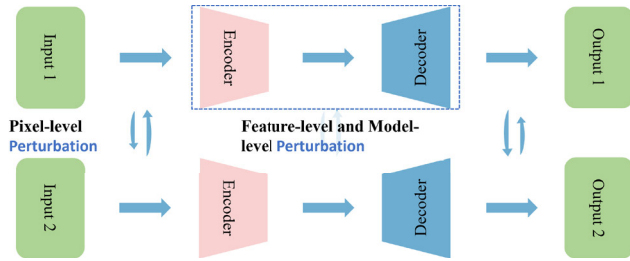


FIGURE 9. Unified remote sensing image semantic segmentation framework incorporating perturbation at multiple levels.

are abundant [239]. Existing SeL methods for SS can be broadly organized into three strategy families: (i) self-training and graph-based pseudo-label propagation [57], [197], [198], [199], [203], [204], [205], (ii) joint optimization of supervised and unsupervised losses [67], [200], [201], [202], and (iii) consistency regularization [3], [79], [107], [205], [206], [207], [208], [209], [210], [211], [212], [213], [214], [215].

1) SELF-TRAINING AND GRAPH-BASED PSEUDO-LABEL PROPAGATION

Self-training, also referred to as pseudo-labelling, is one of the most widely used SeL strategies. An initial classifier predicts labels for unlabeled samples, and these predictions are iteratively refined and reused as pseudo labels for further training [198], [199], [203], [204], [205]. Self-training is often coupled with adaptive thresholding to select high-confidence pseudo labels. For example, ICNet [204] introduces an iterative contrastive network that alternates updates between paired networks to enhance pseudo-label quality and segmentation performance. In graph-based patchwise classification, works [57], [197] extend this idea by using superpixels or graph nodes to propagate labels, which can also be treated as pseudo-labels. Combining pseudo-labels with real labels helps mitigate the problem of limited annotated samples, but overall performance is highly sensitive to the quality of pseudo-labels, making their refinement an ongoing challenge.

2) JOINT OPTIMIZATION OF SUPERVISED AND UNSUPERVISED LOSSES

A flexible framework for constructing SeL models is to combine supervised and unsupervised loss functions in a unified objective. Such frameworks allow models to learn both task-specific and task-agnostic knowledge, effectively leveraging unlabeled data to improve decision-making [67], [200], [201], [202]. For HSI classification, Liu et al. [201] proposed a dual-task model in which a generator and discriminator engage in unsupervised adversarial training across modalities, while the encoder of the generator, equipped with a classification head, is trained in a supervised manner. Huang et al. [202] further advanced this idea by combining contrastive and supervised losses into a reconstruction loss via additional branches, encouraging the

network to learn representations that are simultaneously reconstructive, discriminative, and robust. Related universal semi-supervised strategies couple supervised objectives with unsupervised contrastive losses, striking a favourable balance between training speed and effectiveness across multiple HSI datasets.

3) CONSISTENCY REGULARIZATION

Consistency regularization constitutes a third line of SeL research and is particularly prevalent in tilewise SS. The central idea is to enforce prediction consistency under multiple perturbations of the same input, thereby improving generalization and reducing reliance on large labelled datasets [206]. Perturbations can be applied at the input, feature, and model levels, as illustrated in Figure 9. Most existing methods focus on pixel-level perturbations and employ random image augmentations such as color jittering [208], random paste [209], cutout, edge enhancement, grayscale conversion, and blurring [205], [207]. More advanced strategies further incorporate pseudo-label selection [240], CutMix-based mixing [210], or combinations of multiple transformations to promote the learning of domain-invariant features [211]. In contrast, relatively few studies consider feature- or model-level perturbations in isolation. Chen et al. [212] introduced random drop and noise in the feature-map domain, while Li et al. [206] proposed several types of random feature perturbations within a GAN framework to optimize a consistency loss. Model-level perturbation techniques, such as mean teacher and cross pseudo supervision, introduce variations in model weights or structure to improve robustness. In practice, applying perturbations at multiple levels simultaneously has proved more effective for learning invariant representations [79], [107], [213]. For instance, PiCoCo [214] and ClassHyPer [215] incorporate both feature- and sample-level consistency regularization, while Semi-FCMNet [3] combines data augmentation with MT across input, model, and feature levels to maintain consistency and amplify salient features.

Overall, patchwise classification has primarily used self-training and joint optimization, whereas tilewise segmentation has driven the development of richly regularized frameworks, offering promising directions for future patchwise SeL designs.

M. WEAKLY SUPERVISED LEARNING (WSL)

Coarse annotations, such as image-level labels, bounding boxes, point annotations, and scribble annotations, are significantly easier to obtain than pixel-level labels. Weakly supervised learning semantic segmentation (WSLSS) methods offer a practical solution for performing SS under weak supervision, mitigating the challenges posed by the lack of dense pixel annotations. The general framework of WSLSS involves two steps: generating pseudo-masks from coarse annotations and then training the model using standard SS techniques [241].

Among all WSL types, image-level WSL is the most widely studied and challenging. The foundational idea originates from the work of Zhou et al. [242], who observed that CNNs exhibit strong localization capabilities even when trained solely on image-level labels. Their class activation mapping method became a standard technique for generating pseudo-label seeds in WSSS [217]. Subsequent studies addressed issues related to the noise and incompleteness of initial pseudo labels and spatial context [216], [218], [219], [220], [221], [222]. For instance, Yang et al. [216] proposed ITER to bridge image-level tags and pixelwise HSI prediction via a two-stage pipeline. CAM-based pseudo-labels are progressively refined using spectral/spatial activation and spectral-spatial alignment losses, after which a highly enhanced Transformer with high-frequency-aware self-attention learns boundary-detailed dense maps. Javed et al. [222] proposed a weakly supervised DA approach for built-up region segmentation, incorporating a detection network that leverages image-level labels to support DA.

Other forms of weak supervision have also been applied to various RSIS tasks. NFANet [223] proposed a neighbour sampler to utilize point-level labels for water body extraction. Wei et al. [224] introduced a scribble-based weak-supervision method for road surface extraction that combines road label propagation with holistically nested edge detection to generate training masks. Li et al. [225] leveraged low-resolution land-cover products as weak supervision signals and proposed a low-to-high framework for large-scale, high-resolution land-cover mapping.

In summary, while WSSS methods show promise in reducing annotation burdens, further developments are needed to adapt them to the unique characteristics of RSIs. Future work should prioritize improving boundary detection, addressing scale and class imbalance, and integrating modern techniques such as multimodal fusion and SSL to enhance the generalizability and accuracy of WSL in RSIS.

IV. INFORMATION FUSION TECHNIQUES

With the increasing availability of multimodal data, attention has been shifting toward multi-source RSI analysis. Beyond the feature extraction strategies discussed in the previous section, the fusion module is a critical component for capturing complementary features from multimodal images. Accordingly, a DL-based fusion taxonomy is introduced that categorizes fusion techniques into simple linear-operation fusion and complex nonlinear-interaction fusion. Linear operation fusion integrates multimodal information through concatenation or linear weighting, whereas nonlinear interaction fusion achieves specific fusion objectives via more complex operations. This taxonomy provides a more nuanced study direction for future data fusion research. It is worth noting that these strategies are not only essential for multimodal data but also widely applicable to multi-branch fusion settings in unimodal SS. Meanwhile, these strategies cover a spectrum ranging from pixel-level classification to holistic image-level segmentation. Table 4 summarises the

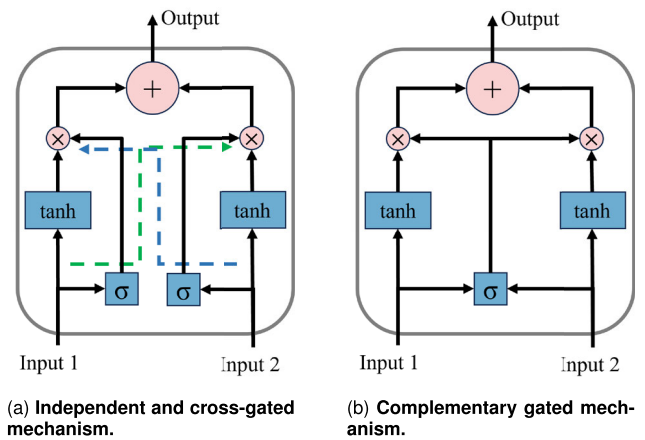


FIGURE 10. Comparison of gated fusion mechanisms for feature modulation. (a) Independent and cross-gated mechanism, where each input generates gating weights via sigmoid activation to modulate either its own features or those of the other branch. (b) Complementary gated mechanism, where a shared gating signal regulates both inputs to encourage complementary feature integration. The $\tanh$ blocks denote nonlinear feature transformation, and σ denotes sigmoid-based gating.

main patch-based and tile-based feature fusion for semantic segmentation models.

A. LINEAR OPERATION FUSION

Linear operation fusion is a simple yet effective method widely used in multimodal fusion approaches, which consists of elementary arithmetic and concatenation. As illustrated in Figure 8, the fusion operation can be implemented at different stages of the model pipeline, namely: pixel-level fusion [31], [245], feature-level fusion [2], [264], and decision-level fusion [248], [263]. Moreover, the fusion operation can be performed once or multiple times, namely: single-point fusion and multi-point fusion. As the name suggests, single-point fusion only performs one fusion operation, and multi-point fusion conducts multiple fusions during forward propagation.

In the early stages of multimodal fusion, raw data is typically channel-wise concatenated directly to maintain compatibility with traditional unimodal models [31], [243], [244], [245], [246], [247], [260], [261], [262]. For instance, Yang et al. [245] stack HSI and LiDAR channels at the input and rely on a single-stream backbone to capture cross-modal interactions through standard convolutional feature extraction realized with group convolutions. Dong et al. [247] produce a single fused representation from GF-5 and Sentinel-2B at the pixel level and then perform downstream lithology classification using a ViT encoder followed by graph-based modeling in ViT-DGCN. Kampffmeyer et al. [31] first demonstrated the feasibility of tilewise RSIS using a vanilla FCN architecture, in which pixel-level concatenation of the true orthophoto, digital surface model (DSM), and normalized DSM was performed. Yue et al. [260] append DSM-derived depth cues to the optical channels and feed the resulting stacked tensor into a U-Net-style encoder-decoder composed of conventional convolutional blocks for SS.

TABLE 4. An overview of information fusion strategies for remote sensing image semantic segmentation. Based on the surveyed literature, existing feature fusion techniques are further classified into several subclasses.

Feature fusion techniques	Patch-based feature fusion reference	Tile-based feature fusion reference
Linear operation fusion		
Single-point fusion	Single-branch structure: [243], [244], [245], [246], [247], [248] Symmetric structure: [2], [249], [250], [251], [252], [253], [254], [255], [256] Asymmetric structure: [257], [258], [259]	Single-branch structure: [31], [260], [261], [262], [263] Symmetric structure: [264], [265] Asymmetric structure: [32]
Multi-point fusion	Symmetric structure: [266], [267], [268], [269], [270], [271] Asymmetric structure: [272], [273], [274], [275]	Symmetric structure: [276], [277], [278], [279]
Nonlinear interaction fusion		
Attention mechanism fusion	Single cross-attention: [280], [281], [282], [283], [284], [285], [286], [287], [288], [289] Mutual cross-attention: [11], [41], [290], [291], [292], [293], [294], [295], [296], [297], [298], [299]	Single cross-attention: [300] Mutual cross-attention: [301], [302], [303], [304], [305], [306], [307]
Gated mechanism fusion	Independent gated mechanism: [297] Complementary gated mechanism: [308], [309], [310]	Independent gated mechanism: [311], [312] Complementary gated mechanism: [313], [314], [315], [316] Cross-gated mechanism: [317]
Reconstruction mechanism fusion	Self-reconstruction: [318], [319], [320], [321], [322] Cross-reconstruction: [323], [324], [325]	Self-reconstruction: [326], [327]
Alignment mechanism fusion	[328], [329]	[326], [330], [331], [332], [333], [334]

Xie et al. [262] stacking heterogeneous geospatial sources as the network input, and they learn fused representations using classical CNN segmentation backbones such as SegNet with VGG16-style feature extraction, complemented by DeepLab-inspired components in their hybrid architecture.

In a single-branch formulation, the model keeps one primary predictor, while the auxiliary modality is introduced only as an external constraint or prior during post-processing, rather than through an explicit parallel feature-extraction branch. In this spirit, Liu et al. [263] use an FCN as the sole dense predictor on optical imagery and inject LiDAR-derived cues as additional unary information within a higher-order CRF to refine the final labelling. Likewise, Zhao et al. [248] treat the CNN output as the main decision signal and incorporate DSM-driven pixel affinities in a hierarchical random-walk layer to regularize predictions and improve spatial consistency. This design is attractive because it preserves a compact, unimodal-compatible backbone and typically reduces parameter growth by avoiding a second deep branch, while still leveraging complementary modalities through plug-in structured inference, which can enhance boundary coherence and suppress noise without redesigning the main network.

However, the network must learn cross-modal interactions from raw stacked channels, making it more sensitive to dynamic-range mismatch, modality-specific noise, and

residual misregistration, and it does not explicitly model nonlinear inter-modal relationships. Feature-level fusion integrates extracted features from different modalities to form a more comprehensive representation, enabling deeper fusion. The branches used to extract multimodal features can either share the same structure or be customized to the data types, i.e., symmetric or asymmetric. Single-point symmetric fusion is the most common form of patchwise multimodal fusion [249], [250], [251], [252], [253], [254], [255]. In [252] and [253], a shared encoder was employed during training, where contrastive and clustering losses from each branch were aggregated. Feature interactions between branches were learned using backpropagation. Du et al. [244] trained a multimodal graph and CNN framework using two unsupervised loss functions and applied an SVM classifier to the unified fused feature representation. In addition, CapViT [246] integrated CapsNet with a Transformer encoder to enable long-range global feature fusion from multi-scale patches. FrIT [255] employed a fractional Fourier image transformer to capture global contextual and sequential information through a linear fusion strategy.

Due to increased model complexity, only a few works focus on leveraging single-point fusion in tile-based segmentation [264], [265]. Multimodal information fusion at a single location often requires a large amount of training data to achieve optimal performance. To address

this limitation, many studies have adopted multi-point linear fusion strategies to enhance modality interaction and improve the efficiency of complementary information extraction [266], [267], [268], [269], [270], [271], [276], [277], [278], [279]. In patchwise classification, Zhang et al. [268] aligned the source and target domains before applying a classification network in combination with multi-point linear fusion, achieving strong cross-scene performance. Zhao et al. [269] introduced Octave convolution and fractional Gabor convolution to preserve multisource, multiscale, and multi-directional features during feature extraction. Zhang et al. [270] performed concatenation and addition operations on features derived from a Transformer encoder to achieve effective fusion. In tilewise SS, FuseNet [276] demonstrated that early summation fusion improves classification accuracy by leveraging the complementary nature of multimodal data through joint feature learning. Other works followed this early feature-fusion strategy, incorporating additional enhancements. Peng et al. [277] and He et al. [279] combined early feature fusion with dense connections, atrous convolution, and attention mechanisms. More recently, MultiSenseSeg [278] introduced modality-specific experts to extract consistent features and reduce inter-modal differences before concatenation for downstream encoding and decoding.

The above approaches employ symmetric network architectures for all modalities. Designing non-symmetric networks tailored to different modalities is considered more effective for handling heterogeneous multimodal data [257], [274]. Guo et al. [257] emphasized the unique characteristics of spectral and spatial modalities, advocating for distinct feature extraction strategies. Du et al. [274] concatenated pixel-level and feature-level representations to extract both local and global features from HSI and LiDAR data. Several works [32], [258], [272], [273], [275] employed two different network structures to separately extract spectral features from HSIs and spatial features from LiDAR data under both single-point and multi-point linear fusion frameworks. Sherrah et al. [32] fused a pretrained CNN and FCN within the encoder to generate full-resolution labelling without downsampling. AMSSE-Net [272] implemented an involution operation for spectral feature characterization and applied five linear operation combinations to fuse multimodal features. MHST [275] processed fused multimodal features through CNNs and Transformer blocks to extract global spectral and local spatial information, respectively.

B. NONLINEAR INTERACTION FUSION

Multimodal feature fusion provides a more comprehensive feature representation. However, simple linear fusion of individually extracted features can lead to redundancy and increase the risk of overfitting [280]. To address redundant information generated by modality imbalance and rigid feature stacking, various nonlinear interaction fusion methods have been proposed. These methods incorporate attention, gating, reconstruction, and alignment mechanisms

to facilitate deeper feature interactions and enhance SS performance.

1) ATTENTION MECHANISM FUSION

Multimodal SS approaches frequently use attention mechanisms to assign weights to different spatial or spectral regions, enabling the model to focus on informative components across modalities. As shown in Figures 6 and 7, when attention weights are derived from other modalities, a cross-attention module is constructed. Based on the structure of fusion models, attention mechanism fusion can be categorised into single cross-attention and mutual cross-attention approaches. This design reflects the varying importance and contribution of each modality to the overall task.

For the single cross-attention model, weights are generated from a single modality and applied to guide feature extraction in another modality. This approach is based on the understanding that different modalities carry varying levels of information. Exploiting privileged information from the dominant modality can improve model performance, whereas the other modality may introduce noise or redundancy. [280], [281], [282], [283], [284], [285], [286], [287], [288], [289], [300]. CNN remains a widely used backbone in cross-attention modules due to its efficiency in capturing local spatial structures and contextual information. FusAt-Net [280] was the first to use a cross-attention mechanism in HSI classification where the attention map derived from LiDAR data was used to emphasise the spatial features of HSI. Dong et al. [281] also generate an attention mask from the LiDAR features and then reuse this mask to modulate both streams, so that elevation cues are injected while spatial structures in HSI are explicitly enhanced. Ren et al. [300] propose a dual-stream HRNet where SAR and optical images are encoded separately, and SAR features are injected into the optical stream at the end of each HRNet stage via cross-modal fusion, providing complementary, all-weather cues to improve dense land-cover labelling. Subsequent works [282], [283], [287] adopted similar strategies to reinforce HSI spatial representations.

Recent studies employed pure Transformer encoders [284], [285], [288], [335] or CNN-Transformer hybrid architecture [286] for multimodal feature extraction. The key distinction among these implementations lies in how the query, key, and value components are drawn from different modalities to establish cross-attention. For example, ExViT [285] uses cross-modality attention and all-token fusion, enabling information to flow across modalities at multiple levels and yielding more discriminative, synergistic multimodal representations than single-modality baselines. Zhao et al. [286] propose a hierarchical CNN-Transformer fusion framework for HSI-LiDAR joint classification. CNNs extract spatial-spectral and elevation features, transformers model long-range dependencies via tokenization, and a cross-token attention fusion encoder integrates both modalities to boost land-cover recognition.

TABLE 5. Summary of unimodal RS datasets used for SS, where SemanticSeg is an abbreviation of semantic segmentation.

Type	Datasets	Image size	GSD(m)	Classes	Area(km^2)	Labels	Task	Year
HSI	Indian Pines (IP)	$1 \times 145 \times 145 \times 224(200)$	20	16	8.41	10,249	Classification	1992
	Washington DC	$1 \times 1208 \times 307 \times 210(191)$	1.5-3.0	7	1.48	26,332	Classification	1995
	Kennedy Space Center (KSC)	$1 \times 512 \times 614 \times 224(176)$	18	13	101.86	4,756	Classification	1996
	Cuprite	$1 \times 250 \times 191 \times 224$	20	30	19.1	47,750	Classification	1997
	Salinas Valley (SV)	$1 \times 512 \times 217 \times 224(204)$	3.7	16	1.52	54,129	Classification	1998
	University of Pavia (UP)	$1 \times 610 \times 340 \times 115(103)$	1.3	9	0.63	42,776	Classification	2001
	Center of Pavia	$1 \times 1096 \times 715 \times 115(102)$	1.3	9	2.03	148,152	Classification	2001
	HOSD [336]	$18 \times \text{Variable} \times 224$	3.2-8.1	2	-	14.84M	SemanticSeg	2010
	Hyrank Dioni	$1 \times 250 \times 1376 \times 176$	30	12	309.6	20,024	Classification	2017
	Hyrank Loukia	$1 \times 249 \times 945 \times 176$	30	14	211.78	13,503	Classification	2017
	Matiwan Village	$1 \times 3750 \times 1580 \times 250$	0.5	20	1.48	5.925M	Classification	2017
	WHU-Hi HanChuan [337]	$1 \times 1217 \times 303 \times 270$	0.109	16	0.0044	368,751	Classification	2018
	WHU-Hi HongHu [337]	$1 \times 940 \times 475 \times 270$	0.043	22	0.00083	446,500	Classification	2018
	WHU-Hi LongKou [337]	$1 \times 550 \times 400 \times 270$	0.463	9	0.047	220,000	Classification	2018
	Xiongan [338]	$1 \times 3750 \times 1580 \times 250$	0.5	19	1.481	2,941,881	Classification	2020
MSI	AeroRIT [339]	$1 \times 1973 \times 3975 \times 372$	0.4	5	1.25	7.843M	SemanticSeg	2020
	WHU-OHS [340]	$7795 \times 512 \times 512 \times 32$	10	24	26.21	90M	SemanticSeg	2024
	Zurich Summer	$20 \times 1000 \times 1150 \times 4$	0.61	8	8.56	23M	SemanticSeg	2015
	RIT-18 [46]	$3 \times \text{Variable} \times 6$	0.047	18	0.46	209 M	SemanticSeg	2017
SAR	LandCoverNet [341]	$8941 \times 256 \times 256 \times 10$	10	7	58596	585.96M	SemanticSeg	2020
	MADOS [1]	$6754 \times 240 \times 240 \times 11$	10	15	38903	389.03M	SemanticSeg	2024
	OSI [342]	$1112 \times 1250 \times 650 \times 1$	10	5	90350	903.2M	SemanticSeg	2019
	SOS-G [343]	$3877 \times 256 \times 256 \times 1$	12.5	2	39700.48	254.08M	SemanticSeg	2022
HRI	SOS-P [343]	$4193 \times 256 \times 256 \times 1$	5×20	2	27479.24	274.79M	SemanticSeg	2022
	SpaceNet1	$6000 \times \text{Variable} \times 3$	0.5	2	2544	-	SemanticSeg	2017
	SpaceNet2	$24586 \times 650 \times 650 \times 3$	0.3	2	3011	10.39B	SemanticSeg	2017
	INRIA [344]	$360 \times 1500 \times 1500 \times 3$	0.3	2	810	810M	SemanticSeg	2017
	DeepGlobe [345]	$1146 \times 2448 \times 2448 \times 3$	0.5	7	1716.9	6.87B	SemanticSeg	2018
	Zeebruges	$7 \times 1000 \times 1000 \times 3$	0.05	8	1.75	7M	SemanticSeg	2018
	GID [34]	$150 \times 6800 \times 7200 \times 3$	4	5	506	7.34B	SemanticSeg	2020
	SkyScapes [346]	$16 \times 5616 \times 3744 \times 3$	0.13	31	5.69	0.34M	SemanticSeg	2019
	GID-Fine [34]	$30000 \times 56 \times 56 \times 3$	4	15	506	94.08M	SemanticSeg	2020
	UAVid [347]	$300 \times 1.5 \times \text{Variable} \times 3$	-	8	-	2.5B	SemanticSeg	2020
	LandCover.ai [348]	$33 \times 9000 \times 9500 \times 3$ $8 \times 4200 \times 4700 \times 3$	0.25-0.5	4	216.27	2.98B	SemanticSeg	2021
	LoveDA [349]	$5987 \times 1024 \times 1024 \times 3$	0.3	7	536.15	12B	SemanticSeg	2021
	FloodNet [350]	$2343 \times 4000 \times 3000 \times 3$	0.015	9	6.3	28B	SemanticSeg	2021

In mutual cross-attention methods, attention weights are derived from all participating modalities. This approach assumes that each modality carries important discriminative information that collaboratively enhances the overall model performance. The discriminative information is regarded as prior in each modality to assist feature extraction in the other modality. Such cross-attention methods are widely adopted in both patchwise classification [11], [41], [290], [291], [292], [293], [294], [295], [296], [297], [298], [299] and tilewise segmentation [301], [302], [303], [304], [305], [306], [307]. For example, in patchwise classification, a spectral-spatial enhancement module can simultaneously strengthen spatial features of HSI using LiDAR data and reinforce LiDAR features using spectral information of HSI [41], [291]. S²ENet [292] introduces S²EM with two complementary cross-modal modules: SAEM computes global spatial affinities from LiDAR to reweight HSI features, while SEEM computes channel-wise relations from HSI to recalibrate LiDAR channels, enabling separate extraction of spatial and channel information before fusion.

Another approach involves extracting features from each modality separately to enhance the fused modality or using

the fused feature to improve the individual modalities. [296], [305]. For instance, TSDN [290] concatenates HSI spatial cues and LiDAR elevation and applies a 3D cross-attention block to strengthen complementary mixed-modality representations. Yang et al. [302] first obtain the fused feature by concatenating the IRRG and DSM features. Then, they concatenate the two weighted fused features using the spatial and channel weights to obtain the final fused feature. DSHFNet [295] builds a spatial self-attention map from one modality, then uses it to reweight the value features of the other two modalities. The final aggregation feature produces enhanced mixed-modality fusion representations. Studies [301], [303] perform attention-aware multimodal fusion by globally pooling each modality to obtain channel statistics, learning modality-specific channel weights via a gating network, reweighting each modality's feature channels, and summing the recalibrated features to output a fused representation. One variant approach is to extract prior features from each modality separately to enhance the principal features, as proposed in [11].

To fully leverage the autonomous learning capabilities of neural networks, many studies also employ pure symmetric

network architectures to extract distinct features from each modality. In Zhang et al. [293], a symmetric mutual-guidance attention module generates attention maps in each branch, enabling bidirectional cross-modal supervision that enhances informative structures and suppresses irrelevant responses before multilevel fusion and classification. MAHiDFNet [294] applies identical spatial self-attention on both the HSI spatial and LiDAR branches to refine modality-specific cues before fusion. The same idea can also be seen in MACN [297], CMFAEN [298], MSFMamba [299], CMFNet [304], MCANet [306] and FTransUNet [307].

2) GATED MECHANISM FUSION

While attention-based fusion techniques are widely used, they typically generate attention weights based on similarity calculations, which incurs excessively high training costs, increasing the computational burden and reducing model efficiency. Gated mechanism fusion addresses this issue by generating a learnable gating coefficient directly via the network. The gating coefficient controls the contributions of different features across modalities and determines the optimal intermediate representation. Based on the operating conditions, gated mechanism fusion methods primarily include independent-gated, complementary-gated, and cross-gated mechanisms, as depicted in Fig. 10.

The independent gate plays an important role in both screening and reconstructing informative features, which uses the independent gate module to control each modality's contribution to the final segmentation. GRRNet [311] introduced a gated feature selection and fusion system that adaptively integrates both low-level and high-level encoder features into a unified feature map to guide decoding. Li et al. [297] introduced MACN for multisource RS classification, where the adaptive CNN encoder gates a dual-branch CNN and selects suitable 2D/3D-style convolutions based on input shape, better preserving HSI spectral-spatial structure while extracting LiDAR elevation cues for later fusion. BCLNet [312] proposed a gated-attention fusion framework that combines the strengths of gating and attention mechanisms to extract common features and minimize modality discrepancies across heterogeneous RSIs. However, this isolated fusion strategy lacks information sharing, making it difficult to obtain meaningful complementary features.

The cross-gated mechanism comprises two gate modules, each employing multimodal information transmission between features to achieve pure information exchange. Kang et al. [317] proposed a cross-gate fusion module that learns SE-based gates from each modality and cross-controls the other, enabling bidirectional information flow to filter redundant features and exploit complementarity for higher accuracy. However, the cross-gated fusion methods only perform cross-control of multimodal features at the final stage of the fusion module, making it difficult to adequately filter redundant information. In contrast, the complementary gate first fuses multimodal images at the input, and then, through a learnable gating function, selects and

combines complementary characteristics from both modalities. It achieves a balance between the independent gate and the cross gate, realizing a seamless integration of features [308], [309], [310], [313], [314], [315], [316].

The implementation of the complementary gate fusion falls into two categories: the first involves computing a gating coefficient g after merging multimodal inputs, then using g and $1 - g$ to calculate the contribution of each modality to the final fusion model [314], [315]. For example, CHGFNet [308] employed this type of gate for land cover classification, taking into account the varying influence of each modality when classifying different land categories. Subsequently, Li et al. [310] implemented a similar gating mechanism where gate weights were derived from multispectral, SAR and slope data. This improves the reliability of impervious surface mapping in complex topographic regions. Zhang et al. [309] extract features from original and superpixel-smoothed images in two branches for SAR image classification, then use a gated channel attention module to generate channel-wise gates and adaptively fuse feature maps, balancing misclassification smoothing and detail preservation.

The second approach feeds the fused features into multiple independent gated modules to obtain gating coefficients that control branch modalities. At each fusion stage, the concatenated features in CEGFNet [313] are fed into two independent gating modules to produce modality-wise coefficients, which reweight each branch before conventional convolutional encoding-decoding in order to learn complementary representations. MGFNet [316] performs optical-SAR SS using an MLP-Gate fusion module to learn complementary gating coefficients that reweight each modality before feature concatenation, which shows the ability to extract more complex, nonlinear cross-modal dependencies and better suppress redundant or noisy channels.

3) RECONSTRUCTION MECHANISM FUSION

Multimodal data show very high heterogeneity. To reduce the redundant information in the fused feature, a feasible solution is to learn a more compact intermediate representation directly. Inspired by the AE, the reconstruction mechanism fusion is proposed, which constrains the model by enforcing a reconstruction process during training. The reconstruction process enhances completeness, reduces redundancy, and improves the discriminative capacity of multimodal middle-feature representations. They can be roughly divided into self-reconstruction fusion [318], [319], [320], [321], [322], [327] and cross-reconstruction fusion [323], [324], [325].

Self-reconstruction fusion learns compact feature representation by minimizing the discrepancy between input and reconstructed data. EndNet [319] used an AE structure to reconstruct raw patch images directly. GLT Net [318] and MIViT [322] focus on reconstructing encoded multi-scale local spatial features using CNNs with the aid of Transformer modules. In particular, MIViT employs information aggregation and distribution flows to generate non-redundant,

complementary features for classification. Lu et al. [321] incorporated adversarial learning into a multimodal context, where generators and discriminators engage in an adversarial process. This approach preserves fine details and complementary information across modalities, allowing for the extraction of high-order semantic features.

The other form of reconstruction approach, cross-reconstruction fusion, transforms one modality into another rather than reconstructing the original input [323], [324], [325]. At the data level, Zhang et al. [324] reconstructed LiDAR data from HSI within an unsupervised feature extraction framework. The hidden representation generated during translation preserves feature quality, even when training with limited labeled samples. At the feature level, CCR Net [323] first extracts modality-specific features using CNNs and then fuses them through a cross-channel reconstruction module. Shivam [325] designed a self-looping CNN that first extracts pre-fusion features and then reconstructs one modality's representation from another, facilitating more discriminative feature learning.

Reconstruction-mechanism fusion is relatively rare in tilewise SS. A few works adopt the reconstruction mechanism fusion to maintain performance under missing inputs in small datasets [326] or modality-incomplete inference [327]. The main reason is that dense SS pipelines are already compute- and memory-intensive, and adding extra decoders, plus reconstruction objectives, substantially increases optimisation cost and implementation complexity.

4) ALIGNMENT MECHANISM FUSION

Alignment mechanism fusion aims to reduce the modality gap before aggregation so that fused representations remain semantically consistent. In practice, alignment is less standardized than attention or gated fusion, and it often appears as parameter-sharing constraints, adversarial regularization, or explicit alignment objectives [326], [328], [329], [330], [331], [332], [333], [334]. For example, Hong et al. [328] propose a cross-fusion design that reuses unimodal weights, enabling cross-modal transfer while keeping the backbone compact. Hang et al. [329] learn a shared embedding for hyperspectral and LiDAR data via a shared projection and a cross-modality contrastive objective that pulls together representations from the same spatial location. For optical-SAR fusion, Li et al. [331] introduce a semantic distribution alignment loss based on MMD to map high-level features into a shared latent space. Then, MMFNet [330] develops a progressive fusion framework that extracts modal-invariant components such as phase features to bridge residual semantic gaps. Complementarily, Hong et al. [326] integrate self-GANs and mutual-GANs to learn perturbation-insensitive representations and explicitly reduce inter-modality discrepancy, yielding more robust cross-modal segmentation.

Recent work further broadens the landscape of alignment mechanism fusion, especially for optical-SAR data where geometric distortions and scale inconsistency make

“align-first, fuse-later” increasingly common. Chen et al. [333] propose MDCA-Net, introducing a cross-modal multi-directional alignment module to mitigate modality discrepancies induced by different imaging mechanisms before subsequent context aggregation and fusion. Li et al. [334] develop SACFNet for multimodal SS and explicitly position alignment as a prerequisite to fusion by designing a multi-scale, multi-directional alignment module to rectify scale/geometric mismatches between optical and SAR features prior to contextual fusion and dynamic integration. Complementarily, MoSAiC [332] demonstrates an embedding-level alignment route, where intra- and inter-modality contrastive objectives exploit the natural co-registration of multi-sensor patches to enforce semantic consistency across modalities, yielding more transferable multimodal representations under limited labels.

V. DATASETS FOR RSIS

As one of the most advanced forms of data-driven modeling, DL models, particularly those based on neural networks, require large volumes of data to automatically learn patterns, features, and relationships without manual feature engineering. The development and performance of these models are highly dependent on both the quality and quantity of training data.

With the increasing availability of airborne and spaceborne sensors, high-quality RSIs are now accessible on a daily basis. To facilitate efficient identification of relevant datasets, a list of the most representative datasets for RSIS is compiled, covering a wide range of thematic areas. As depicted in Tables 5 and 6, these RS datasets support the advancement of RSIS by enabling more comprehensive training and evaluation across diverse tasks. RSIs differ from conventional street-view imagery. Captured from high altitudes, this “bird’s-eye view” allows for broad area observation but introduces unique challenges in data processing [326]. In addition, dataset selection for the four training paradigms relies on empirical guidelines dictated by data characteristics. Massive spatial datasets require models at the tile- or image-level to accommodate memory limits, whereas smaller or highly spectral datasets suit models at the patch-level to extract local signatures without overfitting. We list the most common tasks for these datasets.

A. HIGH-RESOLUTION DATASETS

High-resolution images (HRIs) typically operate within the visible wavelength range, aligning with human visual perception and thereby simplifying the labeling process. With spatial resolutions at the metre or sub-metre level, HRIs provide rich spatial detail. And, they usually contain only three spectral bands, resulting in smaller data volumes that are easier to store and process. Due to these advantages, HRIs constitute the largest proportion of labelled RS data. For example, the SpaceNet initiative released a large-scale dataset comprising over 3000 km² of coverage and more than 10 billion labelled objects. Similarly, LoveDA [349] offers

TABLE 6. Summary of multimodal RS datasets used for SS, where ReferringSeg and ReasoningSeg are abbreviations of referring segmentation and reasoning segmentation, respectively.

Datasets	Type	Image size	GSD(m)	Classes	Area(km^2)	Labels	Task	Year
Trento	HSI LiDAR	$1 \times 166 \times 600 \times 63$ $1 \times 166 \times 600 \times 2$	1	6	0.1	30,414	Classification	2007
Berlin	HSI SAR	$1 \times 1723 \times 476 \times 224$ $1 \times 1723 \times 476 \times 4$	30	8	738.13	464,671	Classification	2009
MUUFLL Gulfport	HSI LiDAR	$1 \times 325 \times 220 \times 64(72)$ $1 \times 325 \times 220 \times 2$	1	11	0.07	53,687	Classification	2010
DFC 2013	HSI LiDAR	$1 \times 1095 \times 349 \times 144$ $1 \times 1095 \times 349 \times 1$	2.5	15	2.39	15,029	Classification	2012
ISPRS Vaihingen	RGB LiDAR DSM	$33 \times \text{Variable} \times 3$ $33 \times \text{Variable} \times 1$ $33 \times \text{Variable} \times 1$	0.09	6	1.34	168M	SemanticSeg	2013
ISPRS Potsdam	MSI LiDAR DSM	$38 \times 6000 \times 6000 \times 4$ $38 \times 6000 \times 6000 \times 1$ $38 \times 6000 \times 6000 \times 1$	0.05	6	3.42	1.37B	SemanticSeg	2013
DFC 2018	HSI LiDAR RGB	$1 \times 601 \times 2384 \times 48$ $1 \times 1202 \times 4768 \times 3$ $1 \times 1202 \times 4768 \times 3$	1 0.5	20	1.43	2.02M	Classification	2017
Augsburg [351]	HSI SAR LiDAR	$1 \times 332 \times 485 \times 180$ $1 \times 332 \times 485 \times 4$ $1 \times 332 \times 485 \times 1$	30	7	144.92	78,293	Classification	2021
LCZ [328]	MSI SAR	$1 \times 626 \times 643 \times 10$ $1 \times 626 \times 643 \times 4$	100	10	4025.18	30,087	Classification	2021
C2Seg-AB [352]	HSI MSI SAR	$1 \times 2465 \times 811 \times 242$ $1 \times 886 \times 1360 \times 242$ $1 \times 2465 \times 811 \times 4$ $1 \times 886 \times 1360 \times 4$ $1 \times 2465 \times 811 \times 2$ $1 \times 886 \times 1360 \times 2$	10	13	20 12.05	2M 1.21M	SemanticSeg	2023
C2Seg-BW [352]	HSI MSI SAR	$1 \times 13474 \times 8706 \times 116(330)$ $1 \times 6225 \times 8670 \times 116(330)$ $1 \times 13474 \times 8706 \times 4$ $1 \times 6225 \times 8670 \times 4$ $1 \times 13474 \times 8706 \times 2$ $1 \times 6225 \times 8670 \times 2$	10	13	1173.05 539.71	117.31M 5397M	SemanticSeg	2023
MDAS [353]	SAR LiDAR MSI HSI	$1 \times 888 \times 1371 \times 2$ $1 \times 29600 \times 45700 \times 1$ $1 \times 888 \times 1371 \times 12$ $1 \times 4036 \times 6232 \times 368$	10 0.25 10 2.2	16	121.75	-	SemanticSeg	2023
Ticino [354]	RGB PAN HSI VNIR HSI SWIR DTM	$1502 \times 256 \times 362 \times 3$ $1502 \times 96 \times 192 \times 1$ $1502 \times 96 \times 192 \times 60(63)$ $1502 \times 96 \times 192 \times 122(171)$ $1502 \times 101 \times 203 \times 1$	1.86-2.64 5	8/10	1331.72	-	SemanticSeg	2024
SZUTreeData-R1 [355]	RGB HSI LiDAR	$1 \times 6170 \times 4810 \times 3$ $1 \times 3085 \times 2405 \times 112$ $1 \times 3085 \times 2405 \times 1$	0.05 0.1	20	0.07	492,631	Classification	2025
SZUTreeData-R2 [355]	RGB HSI LiDAR	$1 \times 8080 \times 4888 \times 3$ $1 \times 4040 \times 2444 \times 112$ $1 \times 4040 \times 2444 \times 1$	0.05 0.1	21	0.1	696,620	Classification	2025
RefSegRS [356]	Text RGB	4420 $4420 \times 512 \times 512 \times 3$	0.13	14	19.58	-	ReferringSeg	2024
RRSIS-D [357]	Text RGB	17402 $17402 \times 800 \times 800 \times 3$	0.5-30	20	-	-	ReferringSeg	2024
RISBench [358]	Text RGB	52472 $52472 \times 512 \times 512 \times 3$	0.1-30	26	-	-	ReferringSeg	2025
EarthReason [359]	Text RGB	5434 $5434 \times \text{Variable} \times 3$	0.5-153	28	-	-	ReasoningSeg	2025

over 12 billion labelled instances across three cities. Such large-scale benchmark datasets enable the development of models with strong generalisation capability.

B. HYPERSPECTRAL DATASETS

HSIs capture information across a broad range of wavelengths, extending far beyond the visible spectrum. Each pixel in an HSI corresponds to a spectral vector that describes the interaction of light with materials at multiple wavelengths, enabling precise material identification and signal recognition [360]. Since the 1990s, datasets such as Indian Pines and Washington DC have been widely used in land cover and land use studies. However, the high dimensionality of HSI leads to large file sizes and complex processing, limiting the quantity and diversity of publicly available hyperspectral datasets [361]. Recently, WHU-OHS [340] was introduced as the first large-scale, publicly available hyperspectral dataset. It contains approximately 90 million manually labelled samples, with extensive geographic distribution and broad spatial coverage, offering new opportunities for robust model development and benchmarking.

C. MULTISPECTRAL DATASETS

Multispectral images (MSIs) provide a balance between HSIs and HRIs by offering high-quality spatial and spectral information at a smaller data scale. This makes them more efficient to process while still capturing essential spectral features for RS applications. The Sentinel-2 mission has markedly contributed to the availability of free MSIs for the RS community. These data support a wide range of applications, including agricultural monitoring, emergency management, water quality assessment, and LULC analysis. For instance, LandCoverNet [341] offers a global benchmark dataset for land cover classification using Sentinel-2 imagery at 10 m spatial resolution.

D. SAR DATASETS

As an active RS technique, SAR operates independently of lighting and most weather conditions, enabling continuous day-and-night imaging. SAR is highly sensitive to changes in surface roughness, making it particularly effective for detecting phenomena such as wind speeds, wave heights, ocean eddies, and surface composition. Due to these capabilities, SAR systems have become essential tools for ocean observation. Zhu et al. introduced a manually labeled dataset focused on oil spill detection in the Gulf of Mexico and Persian Gulf regions. Marios et al. [342] compiled oil spill data from Sentinel-1 observations across Europe between 2015 and 2017, offering a valuable benchmark for future SAR-based oil spill detection research.

E. LiDAR DATASETS

LiDAR provides precise 3-D spatial information, making it critical for applications requiring detailed topographic and structural analysis. Using laser pulses to measure distances,

LiDAR systems generate dense point clouds that represent both ground surfaces and above-ground features, such as buildings and vegetation. A key advantage of LiDAR is its ability to penetrate vegetation canopies, enabling accurate ground elevation mapping. However, LiDAR typically involves large file sizes and requires complicated preprocessing, including filtering and interpolation, to produce derived products such as digital terrain models (DTMs) and canopy height models. Despite these challenges, LiDAR datasets remain indispensable in urban planning, forestry monitoring, and disaster response, offering high-resolution, scalable insights across diverse application domains.

F. MULTIMODAL DATASETS

Multimodal RS combines data from different sensor types to extract complementary information, resulting in more comprehensive and accurate surface understanding and improved performance in downstream tasks. Early datasets such as Trento and Berlin were used to evaluate multimodal fusion performance, but were limited to two modalities and contained minimal annotation. With the introduction of MDAS [353] and Ticino [354], multimodal RS benchmark datasets have become more diverse and extensive. Ticino includes five modalities: RGB, DTMs, panchromatic, HSI in the visual-near-infrared (VNIR) range, and HSI in the short-wave infrared (SWIR) range. This dataset offers a robust benchmark for evaluating multimodal fusion algorithms across complex and heterogeneous inputs. More recently, natural language text has been introduced as a novel modality to align visual features with semantic concepts, enabling advanced vision-language tasks in the remote sensing domain. Emerging large-scale datasets, such as RRSIS-D [357], EarthReason [359], pair HRIs with rich textual descriptions to support complex multimodal applications like referring segmentation and reasoning segmentation.

VI. CRITICAL THINKING AND FUTURE DEVELOPMENTS

In previous sections, we examined the evolution of RSIS through a structural lens: the “pixel-patch-tile-image” hierarchy. This granularity-based taxonomy not only reveals progress in end-to-end SS but also exposes the core engineering bottlenecks inherent in each scale. While conventional reviews often attribute future breakthroughs to “deeper networks” or “larger datasets,” this survey posits that the future of RSIS depends fundamentally on how researchers navigate the trade-offs among segmentation granularities and, ultimately, on how they dismantle the boundaries that separate them. Accordingly, we propose critical perspectives and future directions across four key dimensions.

A. ENDURING VALUE OF PIXEL/PATCH-LEVEL METHODS IN LOW-LABEL AND HIGH-SPECTRAL REGIMES

With the rapid rise of tile-level and image-level models, pixel- and patch-based methods are frequently criticized for their restricted receptive fields and computational redundancy, often being dismissed as obsolete paradigms. However,

this perspective overlooks the complex multimodal nature of RS data. When processing HSI, the exceedingly high spectral dimensionality makes large-scale spatial modeling susceptible to the curse of dimensionality and severe overfitting. In such scenarios, the “local spatial-spectral focus” maintained by patch-level methods remains one of the most robust feature extraction strategies. Furthermore, in specialized domains such as environmental and ecological studies, annotation costs are prohibitively high. Researchers often only have access to sparse, pixel-wise labels, rendering data-hungry tile or image-level models impractical for direct application.

Future research at the pixel and patch level should pivot away from only chasing state-of-the-art metrics on standard optical benchmarks and instead focus on extreme operational conditions: First, researchers should explore lightweight patch models for multimodal data. This involves investigating how multiple data sources can be fused with hyperspectral patches at a localized level, maximizing local discriminative power while maintaining a highly efficient minimal receptive field. Second, investigating active learning in extreme few-shot scenarios is crucial. This requires studying how patch-level uncertainty quantification can guide active learning frameworks to achieve high-precision semantic extraction with minimal human annotation.

B. BOTTLENECKS AND EVOLUTION OF TILE-LEVEL PARADIGMS FOR EFFICIENT DENSE MAPPING

The tile-level paradigm is currently the dominant compromise in both academia and industry for processing HRIs, striking a viable balance between local detail retention and GPU memory constraints. However, its fundamental contradiction lies in the fact that artificial physical cropping disrupts genuine geographical topology. Edge artifacts at tile seams and the forced truncation of large-scale continuous entities remain the inherent vulnerability of tile-level models, particularly problematic in tasks like large-scale impervious surface extraction. While many studies attempt to mitigate this via overlapping inference, this introduces multiplied computational redundancy, directly contradicting the overarching goal of highly efficient, large-scale dense mapping.

To realize true large-scale RS information extraction, future tile-level architectures must transcend the limitations of conventional sliding windows. First, there is a need for seamless cross-tile contextual interaction. This involves developing architectures that can share boundary features or cache global context in real time during sequential tile inference, eliminating receptive-field truncation at the algorithmic level without relying on physical overlap. Second, optimizing efficient dense mapping in resource-constrained environments is essential. For edge computing devices such as airborne platforms, enhancing the inference efficiency of tile-level models is critical to enable high-precision, seamless mapping while avoiding redundant inference.

C. GENERALIZATION LIMITS AND UNIVERSAL INTERFACES OF IMAGE-LEVEL FOUNDATION MODELS

The holistic learning architecture introduces spatial parallel interaction mechanisms, breaking the previous physical limitations of having to crop large images into tiles, and achieves true end-to-end image-level segmentation. This paradigm resolves the performance degradation typical of traditional cropping and fusion processes. Nevertheless, this holistic nature places extreme demands on the scalability of the computing architecture. When facing ultra-large-scale, ultra-high-resolution, and multimodal heterogeneous imagery, maintaining the highly efficient interaction and global consistency of end-to-end features without degrading to tile processing is currently the greatest challenge at this level. Additionally, generalizing this holistic capability from specific tasks to universal geospatial understanding interfaces requires deeper exploration.

The future of image-level models lies in cost-effective adaptation to RS-specific tasks. First, exploring PEFT for holistic architectures is paramount. Applying techniques like low-rank adaptation to holistic models can enable them to adapt to different sensor parameters and geographical regions with minimal computational overhead. Second, constructing multimodal universal holistic interfaces is necessary. Future frameworks should uniformly process optical, synthetic aperture radar, and light detection and ranging data through universal prompt or instruction interfaces to achieve cross-modality end-to-end holistic parsing.

D. BREAKING THE HIERARCHY: THE FUTURE OF CROSS-REGIME TRANSFER

Reflecting on the four-level taxonomy proposed in this survey, most current research isolates optimization within a single granularity level, such as optimizing a tile model or a global model exclusively. However, in practical applications, this segregation creates an awkward situation where global models are too heavy to deploy, and local models are too light to capture the big picture. This hierarchical gap hinders the deployment of advanced algorithms in practical engineering tasks like real-time environmental monitoring.

Breaking the boundaries of the “pixel–patch–tile–image” hierarchy to achieve cross-granularity knowledge transfer represents the most promising frontier. First, utilizing cross-granularity knowledge distillation is a key direction. Massive holistic learning models can serve as teacher networks to extract global topological structures and geographical association information, guiding the training of lightweight tile-level student networks. Second, this paradigm will facilitate cross-hierarchical consistency modeling. This ensures that even when inference is restricted to local tiles, the model retains the macro perspective learned from the global model, thereby obtaining segmentation results with global consistency while maintaining high inference efficiency.

VII. CONCLUSION

This survey presents a structured review of RSIS in the DL era through a unified pixel–patch–tile–image taxonomy aligned with segmentation granularity and the training/inference pipeline. Under this operational hierarchy, the progression from pixel- and patch-level classification to tile- and image-level modeling is systematically examined. By analyzing supervision paradigms, feature extraction strategies, and multimodal fusion mechanisms, the co-evolution of data scale, architectural design, and representation learning is clarified.

The evolution of RSIS reflects a fundamental shift from local discriminative learning toward globally contextualized and data-driven representation learning. As receptive fields expand and model capacity increases, progress is increasingly governed by scalable training strategies, domain robustness, and multimodal integration rather than architectural complexity alone. The transition from convolutional models to transformer-based architectures and vision foundation models marks a movement toward more transferable and generalizable representations.

Despite substantial advances in common land-cover segmentation, challenges remain in annotation scarcity, modality heterogeneity, computational scalability, and cross-domain reliability. Future progress is expected to depend on harmonizing large-scale pretraining with task-specific adaptation, strengthening domain generalization, and developing unified multimodal frameworks for diverse sensing platforms.

REFERENCES

- [1] K. Kikaki, I. Kakogeorgiou, I. Hoteit, and K. Karantzas, “Detecting marine pollutants and sea surface features with deep learning in Sentinel-2 imagery,” *ISPRS J. Photogramm. Remote Sens.*, vol. 210, pp. 39–54, Apr. 2024.
- [2] Q. Feng, D. Zhu, J. Yang, and B. Li, “Multisource hyperspectral and LiDAR data fusion for urban land-use mapping based on a modified two-branch convolutional neural network,” *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 1, p. 28, Jan. 2019.
- [3] B. Chen, L. Wang, X. Fan, W. Bo, X. Yang, and T. Tjahjedi, “Semi-FCMNet: Semi-supervised learning for forest cover mapping from satellite imagery via ensemble self-training and perturbation,” *Remote Sens.*, vol. 15, no. 16, p. 4012, Aug. 2023.
- [4] F. Qu, Y. Sun, M. Zhou, L. Liu, H. Yang, J. Zhang, H. Huang, and D. Hong, “Vegetation land segmentation with multi-modal and multi-temporal remote sensing images: A temporal learning approach and a new dataset,” *Remote Sens.*, vol. 16, no. 1, p. 3, Dec. 2023.
- [5] N. Casagli, E. Intriari, V. Tofani, G. Gigli, and F. Raspini, “Landslide detection, monitoring and prediction with remote-sensing techniques,” *Nature Rev. Earth Environ.*, vol. 4, no. 1, pp. 51–64, Jan. 2023.
- [6] A. Yu, Y. Quan, R. Yu, W. Guo, X. Wang, D. Hong, H. Zhang, J. Chen, Q. Hu, and P. He, “Deep learning methods for semantic segmentation in remote sensing with small data: A survey,” *Remote Sens.*, vol. 15, no. 20, p. 4987, Oct. 2023.
- [7] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
- [8] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Springer, 2015, pp. 234–241.
- [9] D. Wang, J. Zhang, B. Du, M. Xu, L. Liu, D. Tao, and L. Zhang, “SAMRS: Scaling-up remote sensing segmentation dataset with segment anything model,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 8815–8827.
- [10] W. Chen, L. Bruzzone, B. Dang, Y. Gao, Y. Deng, J.-G. Yu, L. Yuan, and Y. Li, “REST: Holistic learning for end-to-end semantic segmentation of whole-scene remote sensing imagery,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 48, no. 1, pp. 693–710, Jan. 2026.
- [11] Z. Xue, X. Tan, X. Yu, B. Liu, A. Yu, and P. Zhang, “Deep hierarchical vision transformer for hyperspectral and LiDAR data classification,” *IEEE Trans. Image Process.*, vol. 31, pp. 3095–3110, 2022.
- [12] M. Imani and H. Ghassemian, “An overview on spectral and spatial information fusion for hyperspectral image classification: Current trends and challenges,” *Inf. Fusion*, vol. 59, pp. 59–83, Jul. 2020.
- [13] B. Rasti, D. Hong, R. Hang, P. Ghamisi, X. Kang, J. Chanussot, and J. A. Benediktsson, “Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox,” *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 60–88, Dec. 2020.
- [14] N. Zang, Y. Cao, Y. Wang, B. Huang, L. Zhang, and P. T. Mathiopoulos, “Land-use mapping for high-spatial resolution remote sensing image via deep learning: A review,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5372–5391, 2021.
- [15] X. Yuan, J. Shi, and L. Gu, “A review of deep learning methods for semantic segmentation of remote sensing imagery,” *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114417.
- [16] M. Ahmad, S. Shabbir, S. K. Roy, D. Hong, X. Wu, J. Yao, A. M. Khan, M. Mazzara, S. Distefano, and J. Chanussot, “Hyperspectral image classification—Traditional to deep models: A survey for future prospects,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 968–999, 2022.
- [17] L. Huang, B. Jiang, S. Lv, Y. Liu, and Y. Fu, “Deep-learning-based semantic segmentation of remote sensing images: A survey,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 8370–8396, 2023.
- [18] C. Liu, Y. Sun, Y. Xu, Z. Sun, X. Zhang, L. Lei, and G. Kuang, “A review of optical and SAR image deep feature fusion in semantic segmentation,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 12910–12930, 2024.
- [19] V. Kumar, R. S. Singh, M. Rambabu, and Y. Dua, “Deep learning for hyperspectral image classification: A survey,” *Comput. Sci. Rev.*, vol. 53, Aug. 2024, Art. no. 100658.
- [20] M. Akewar and M. Chandak, “An integration of natural language and hyperspectral imaging: A review,” *IEEE Geosci. Remote Sens. Mag.*, vol. 13, no. 1, pp. 32–54, Mar. 2025.
- [21] M. Z. U. Rehman, S. M. S. Islam, D. Blake, A. Ulhaq, and N. Janjua, “Deep learning for land use classification: A systematic review of HS-LiDAR imagery,” *Artif. Intell. Rev.*, vol. 58, no. 9, p. 272, Jun. 2025.
- [22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 60, 2017, pp. 84–90.
- [23] K. Topouzelis, V. Karathanassi, P. Pavlakis, and D. Rokos, “Detection and discrimination between oil spills and look-alike phenomena through neural networks,” *ISPRS J. Photogramm. Remote Sens.*, vol. 62, no. 4, pp. 264–270, Sep. 2007.
- [24] C. Deng, Y. Xue, X. Liu, C. Li, and D. Tao, “Active transfer learning network: A unified deep joint spectral–spatial feature learning model for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1741–1754, Mar. 2019.
- [25] Z. Sun, L. Di, and H. Fang, “Using long short-term memory recurrent neural network in land cover classification on Landsat and cropland data layer time series,” *Int. J. Remote Sens.*, vol. 40, no. 2, pp. 593–614, Jan. 2019.
- [26] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, “Deep learning-based classification of hyperspectral data,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.
- [27] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, “Deep convolutional neural networks for hyperspectral image classification,” *J. Sensors*, vol. 2015, no. 1, 2015, Art. no. 258619.
- [28] B. Chen, H. Zheng, L. Wang, O. Hellwich, C. Chen, L. Yang, T. Liu, G. Luo, A. Bao, and X. Chen, “A joint learning im-BiLSTM model for incomplete time-series sentinel-2A data imputation and crop classification,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 108, Apr. 2022, Art. no. 102762.
- [29] J. Li, Y. Cai, Q. Li, M. Kou, and T. Zhang, “A review of remote sensing image segmentation by deep learning methods,” *Int. J. Digit. Earth*, vol. 17, no. 1, Dec. 2024, Art. no. 2328827.

- [30] V. Mnih and G. E. Hinton, "Learning to label aerial images from noisy data," in *Proc. 29th Int. Conf. Mach. Learn. (ICML-12)*, 2012, pp. 203–210.
- [31] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 680–688.
- [32] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," 2016, *arXiv:1606.02585*.
- [33] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 881–893, Feb. 2017.
- [34] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111322.
- [35] W. Zhao and S. Du, "Learning multiscale and deep representations for classifying remotely sensed imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 113, pp. 155–165, Mar. 2016.
- [36] H. Gao, Y. Yang, C. Li, L. Gao, and B. Zhang, "Multiscale residual network with mixed depthwise convolution for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3396–3408, Apr. 2021.
- [37] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.
- [38] Y. Xu, L. Zhang, B. Du, and F. Zhang, "Spectral–spatial unified networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5893–5909, Oct. 2018.
- [39] M. He, B. Li, and H. Chen, "Multi-scale 3D deep convolutional neural network for hyperspectral image classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3904–3908.
- [40] S. Pande and B. Banerjee, "HyperLoopNet: Hyperspectral image classification using multiscale self-looping convolutional networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 422–438, Jan. 2022.
- [41] X. Wang, J. Zhu, Y. Feng, and L. Wang, "MS2CANet: Multiscale spatial–spectral cross-modal attention network for hyperspectral image and LiDAR classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.
- [42] R. Guo, J. Liu, N. Li, S. Liu, F. Chen, B. Cheng, J. Duan, X. Li, and C. Ma, "Pixel-wise classification method for high resolution remote sensing imagery using deep neural networks," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 3, p. 110, Mar. 2018.
- [43] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.
- [44] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, "Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [45] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 192–194.
- [46] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 60–77, Nov. 2018.
- [47] R. Li, L. Wang, C. Zhang, C. Duan, and S. Zheng, "A2-FPN for semantic segmentation of fine-resolution remotely sensed images," *Int. J. Remote Sens.*, vol. 43, no. 3, pp. 1131–1155, Feb. 2022.
- [48] Y. Yi, Z. Zhang, W. Zhang, C. Zhang, W. Li, and T. Zhao, "Semantic segmentation of urban buildings from VHR remote sensing imagery using a deep convolutional neural network," *Remote Sens.*, vol. 11, no. 15, p. 1774, Jul. 2019.
- [49] K. Yang, H. Sun, C. Zou, and X. Lu, "Cross-attention spectral–spatial network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
- [50] Q. Hong, X. Zhong, W. Chen, Z. Zhang, B. Li, H. Sun, T. Yang, and C. Tan, "SATNet: A spatial attention based network for hyperspectral image classification," *Remote Sens.*, vol. 14, no. 22, p. 5902, Nov. 2022.
- [51] X. Mei, E. Pan, Y. Ma, X. Dai, J. Huang, F. Fan, Q. Du, H. Zheng, and J. Ma, "Spectral–spatial attention networks for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 8, p. 963, Apr. 2019.
- [52] W. Ma, Q. Yang, Y. Wu, W. Zhao, and X. Zhang, "Double-branch multi-attention mechanism network for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 11, p. 1307, Jun. 2019.
- [53] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Visual attention-driven hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8065–8080, Oct. 2019.
- [54] X. Tang, F. Meng, X. Zhang, Y.-M. Cheung, J. Ma, F. Liu, and L. Jiao, "Hyperspectral image classification based on 3-D octave convolution with spatial–spectral attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2430–2447, Mar. 2021.
- [55] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 449–462, Jan. 2021.
- [56] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3232–3245, May 2020.
- [57] Y. Dong, Q. Liu, B. Du, and L. Zhang, "Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 1559–1572, 2022.
- [58] Y. Li, Y. Luo, L. Zhang, Z. Wang, and B. Du, "MambaHSI: Spatial–spectral mamba for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5524216.
- [59] Q. Liu, J. Yue, Y. Fang, S. Xia, and L. Fang, "HyperMamba: A spectral–spatial adaptive mamba for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5536514.
- [60] X. Ma, X. Zhang, Z. Wang, and M.-O. Pun, "Unsupervised domain adaptation augmented by mutually boosted attention for semantic segmentation of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5400515.
- [61] L. Wang, R. Li, D. Wang, C. Duan, T. Wang, and X. Meng, "Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images," *Remote Sens.*, vol. 13, no. 16, p. 3065, Aug. 2021.
- [62] Y. Liu, Y. Zhang, Y. Wang, and S. Mei, "Rethinking transformers for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5617515.
- [63] J. Wu, J. Shi, Z. Zhao, Z. Liu, and R. Zhi, "FreeMix: Open-vocabulary domain generalization of remote-sensing images for semantic segmentation," *Remote Sens.*, vol. 17, no. 8, p. 1357, Apr. 2025.
- [64] X. Ma, X. Zhang, M.-O. Pun, and B. Huang, "A unified framework with multimodal fine-tuning for remote sensing semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5405015.
- [65] M. Liu, J. Dan, Z. Lu, Y. Yu, Y. Li, and X. Li, "CM-UNet: Hybrid CNN-mamba UNet for remote sensing image semantic segmentation," 2024, *arXiv:2405.10530*.
- [66] H. Zhang, Y. Li, Y. Zhang, and Q. Shen, "Spectral–spatial classification of hyperspectral imagery using a dual-channel convolutional neural network," *Remote Sens. Lett.*, vol. 8, no. 5, pp. 438–447, May 2017.
- [67] Q. Liu, Y. Dong, T. Huang, L. Zhang, and B. Du, "A universal knowledge embedded contrastive learning framework for hyperspectral image classification," 2024, *arXiv:2404.01673*.
- [68] X. Kang, B. Deng, P. Duan, X. Wei, and S. Li, "Self-supervised spectral–spatial transformer network for hyperspectral oil spill mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5507410.
- [69] T. Arshad and J. Zhang, "A light-weighted spectral–spatial transformer model for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 12008–12019, 2024.
- [70] F. Gao, T. Huang, J. Wang, J. Sun, A. Hussain, and E. Yang, "Dual-branch deep convolution neural network for polarimetric SAR image classification," *Appl. Sci.*, vol. 7, no. 5, p. 447, Apr. 2017.
- [71] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.
- [72] Z. Zhang, D. Liu, D. Gao, and G. Shi, "S³Net: Spectral–spatial–semantic network for hyperspectral image classification with the multiway attention mechanism," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.

- [73] S. Jia, J. Liao, M. Xu, Y. Li, J. Zhu, W. Sun, X. Jia, and Q. Li, "3-D Gabor convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.
- [74] S. K. Roy, A. Deria, C. Shah, J. M. Haut, Q. Du, and A. Plaza, "Spectral-spatial morphological attention transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5503615.
- [75] W. Fu, K. Ding, X. Kang, and D. Wang, "Local-global gated convolutional neural network for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.
- [76] T. Song, Z. Zeng, C. Gao, H. Chen, and J. Li, "Joint classification of hyperspectral and LiDAR data using height information guided hierarchical fusion-and-separation network," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5505315.
- [77] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 181, pp. 84–98, Nov. 2021.
- [78] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 196–214, Aug. 2022.
- [79] W. Huang, Y. Shi, Z. Xiong, and X. X. Zhu, "Decouple and weight semi-supervised semantic segmentation of remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 212, pp. 13–26, Jun. 2024.
- [80] Y. Li, T. Shi, Y. Zhang, and J. Ma, "SPGAN-DA: Semantic-preserved generative adversarial network for domain adaptive remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5406717.
- [81] X. Sun, M. Xia, and T. Dai, "Controllable fused semantic segmentation with adaptive edge loss for remote sensing parsing," *Remote Sens.*, vol. 14, no. 1, p. 207, Jan. 2022.
- [82] Y. Chen, P. Fang, X. Zhong, J. Yu, X. Zhang, and T. Li, "Hi-ResNet: Edge detail enhancement for high-resolution remote sensing segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 15024–15040, 2024.
- [83] F. Zhou, R. Hang, Q. Liu, and X. Yuan, "Hyperspectral image classification using spectral-spatial LSTMs," *Neurocomputing*, vol. 328, pp. 39–47, Feb. 2019.
- [84] W. Wang, S. Dou, Z. Jiang, and L. Sun, "A fast dense spectral-spatial convolution network framework for hyperspectral images classification," *Remote Sens.*, vol. 10, no. 7, p. 1068, Jul. 2018.
- [85] Q. Liu, Y. Dong, Y. Zhang, and H. Luo, "A fast dynamic graph convolutional network and CNN parallel network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5530215.
- [86] H. Zhang, Y. Li, Y. Jiang, P. Wang, Q. Shen, and C. Shen, "Hyperspectral classification based on lightweight 3-D-CNN with transfer learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5813–5828, Aug. 2019.
- [87] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and S. Nong, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 325–341.
- [88] Q. Chi, G. Lv, G. Zhao, and X. Dong, "A novel knowledge distillation method for self-supervised hyperspectral image classification," *Remote Sens.*, vol. 14, no. 18, p. 4523, Sep. 2022.
- [89] J. Yue, L. Fang, H. Rahmani, and P. Ghamisi, "Self-supervised learning with adaptive distillation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.
- [90] C. Shi, L. Fang, Z. Lv, and M. Zhao, "Explainable scale distillation for hyperspectral image classification," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108316.
- [91] W. Zhao, R. Peng, Q. Wang, C. Cheng, W. J. Emery, and L. Zhang, "Life-long learning with continual spectral-spatial feature distillation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5544214.
- [92] Z. Li, S. Xia, J. Yue, and L. Fang, "HyperKD: Lifelong hyperspectral image classification with cross-spectral-spatial knowledge distillation," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5509017.
- [93] Z. Dong, G. Gao, T. Liu, Y. Gu, and X. Zhang, "Distilling segmenters from CNNs and transformers for remote sensing images' semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5613814.
- [94] W. Zhou, Y. Li, J. Huan, Y. Liu, and Q. Jiang, "MSTNet-KD: Multilevel transfer networks using knowledge distillation for the dense prediction of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4504612.
- [95] W. Zhou, P. Yang, W. Qiu, and F. Qiang, "STONet-S*: A knowledge-distilled approach for semantic segmentation in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4414413.
- [96] Y. Sun, D. Liang, S.-Y. Li, S. Chen, and S.-J. Huang, "Handling noisy annotation for remote sensing semantic segmentation via boundary-aware knowledge distillation," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 4408720.
- [97] W. Zhou, X. Fan, W. Yan, S. Shan, Q. Jiang, and J.-N. Hwang, "Graph attention guidance network with knowledge distillation for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4506015.
- [98] K. Zheng, Y. Chen, J. Wang, Z. Liu, S. Bao, J. Zhan, and N. Shen, "Enhancing remote sensing semantic segmentation accuracy and efficiency through transformer and knowledge distillation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 4074–4092, 2025.
- [99] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.
- [100] Y. Kong, X. Wang, and Y. Cheng, "Spectral-spatial feature extraction for HSI classification based on supervised hypergraph and sample expanded CNN," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4128–4140, Nov. 2018.
- [101] J. Feng, L. Wang, H. Yu, L. Jiao, and X. Zhang, "Divide-and-conquer dual-architecture convolutional neural network for classification of hyperspectral images," *Remote Sens.*, vol. 11, no. 5, p. 484, Feb. 2019.
- [102] W. Yu, S. Wan, G. Li, J. Yang, and C. Gong, "Hyperspectral image classification with contrastive graph convolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5503015.
- [103] B. Liu, A. Yu, X. Yu, R. Wang, K. Gao, and W. Guo, "Deep multiview learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7758–7772, Sep. 2021.
- [104] Y. Chang, Q. Liu, Y. Zhang, and Y. Dong, "Unsupervised multiview graph contrastive feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5524314.
- [105] H. Li, Y. Li, G. Zhang, R. Liu, H. Huang, Q. Zhu, and C. Tao, "Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618014.
- [106] D. Muhtar, X. Zhang, and P. Xiao, "Index your position: A novel self-supervised learning method for remote sensing images semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4411511.
- [107] W. Miao, Z. Xu, J. Geng, and W. Jiang, "ECAE: Edge-aware class activation enhancement for semisupervised remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5625014.
- [108] L. Wang, P. Xiao, X. Zhang, and X. Chen, "A fine-grained unsupervised domain adaptation framework for semantic segmentation of remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4109–4121, 2023.
- [109] B. Benjdira, Y. Bazi, A. Koubaa, and K. Ouni, "Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images," *Remote Sens.*, vol. 11, no. 11, pp. 1369–1392, Jun. 2019.
- [110] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "ColorMapGAN: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7178–7193, Oct. 2020.
- [111] S. Ji, D. Wang, and M. Luo, "Generative adversarial network-based full-space domain adaptation for land cover classification from multiple-source remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 3816–3828, May 2021.
- [112] Y. Cai, Y. Yang, Y. Shang, Z. Chen, Z. Shen, and J. Yin, "IterDANet: Iterative intra-domain adaptation for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5629517.

- [113] J. Yang, Y.-Q. Zhao, and J. C.-W. Chan, "Learning and transferring deep joint spectral-spatial features for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4729–4742, Aug. 2017.
- [114] B. Huang, B. Zhao, and Y. Song, "Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery," *Remote Sens. Environ.*, vol. 214, pp. 73–86, Sep. 2018.
- [115] X. He, Y. Chen, and P. Ghamisi, "Heterogeneous transfer learning for hyperspectral image classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3246–3263, May 2020.
- [116] C. Zhong, J. Zhang, S. Wu, and Y. Zhang, "Cross-scene deep transfer learning with spectral feature adaptation for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2861–2873, 2020.
- [117] E. Zhang, J. Liu, A. Cao, Z. Sun, H. Zhang, H. Wang, L. Sun, and M. Song, "RS-SAM: Integrating multi-scale information for enhanced remote sensing image segmentation," in *Proc. Asian Conf. Comput. Vis.*, 2024, pp. 994–1010.
- [118] X. Lu and Q. Weng, "Multi-LoRA fine-tuned segment anything model for urban man-made object extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5637519.
- [119] W. Wu, M. S. Wong, X. Yu, G. Shi, C. Y. T. Kwok, and K. Zou, "Compositional oil spill detection based on object detector and adapted segment anything model from SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.
- [120] S. Chen, Y. Yu, Y. Li, Z. Wang, X. Li, and J. Han, "Multiscale adapter based on SAM for remote sensing semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 6806–6819, 2025.
- [121] H. Yang, Z. Jiang, Y. Zhang, Y. Wu, H. Luo, P. Zhang, and B. Wang, "A high-resolution remote sensing land use/land cover classification method based on multi-level features adaptation of segment anything model," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 141, Jul. 2025, Art. no. 104659.
- [122] B. Liu, X. Yu, A. Yu, P. Zhang, G. Wan, and R. Wang, "Deep few-shot learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2290–2304, Apr. 2019.
- [123] K. Gao, B. Liu, X. Yu, and A. Yu, "Unsupervised meta learning with multiview constraints for hyperspectral image small sample set classification," *IEEE Trans. Image Process.*, vol. 31, pp. 3449–3462, 2022.
- [124] Z. Li, M. Liu, Y. Chen, Y. Xu, W. Li, and Q. Du, "Deep cross-domain few-shot learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5501618.
- [125] S. Zhang, Z. Chen, D. Wang, and Z. J. Wang, "Cross-domain few-shot contrastive learning for hyperspectral images classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [126] Y. Xu, Y. Zhang, T. Yue, C. Yu, and H. Li, "Graph-based domain adaptation few-shot learning for hyperspectral image classification," *Remote Sens.*, vol. 15, no. 4, p. 1125, Feb. 2023.
- [127] J. Zeng, Z. Xue, L. Zhang, Q. Lan, and M. Zhang, "Multistage relation network with dual-metric for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5510017.
- [128] J. Li, Z. Zhang, R. Song, Y. Li, and Q. Du, "SCFormer: Spectral coordinate transformer for cross-domain few-shot hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 33, pp. 840–855, 2024.
- [129] S. Feng, H. Zhang, B. Xi, C. Zhao, Y. Li, and J. Chanussot, "Cross-domain few-shot learning based on decoupled knowledge distillation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5534414.
- [130] X. Jiang, N. Zhou, and X. Li, "Few-shot segmentation of remote sensing images using deep metric learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [131] B. Wang, Z. Wang, X. Sun, H. Wang, and K. Fu, "DMML-Net: Deep metametric learning for few-shot geographic object segmentation in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5611118.
- [132] C. Lang, G. Cheng, B. Tu, and J. Han, "Global rectification and decoupled registration for few-shot segmentation in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5617211.
- [133] Q. Cao, Y. Chen, C. Ma, and X. Yang, "Few-shot rotation-invariant aerial image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5600513.
- [134] G. Puthumanaim and U. Verma, "Texture based prototypical network for few-shot semantic segmentation of forest cover: Generalizing for different geographical regions," *Neurocomputing*, vol. 538, Jun. 2023, Art. no. 126201.
- [135] S. Li, F. Liu, L. Jiao, X. Liu, P. Chen, and L. Li, "Mask-guided correlation learning for few-shot segmentation in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5636714.
- [136] J. Chen, X. Wang, L. Hong, and M. Liu, "Cross-domain few-shot segmentation for remote sensing image based on task augmentation and feature disentanglement," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 9360–9375, 2024.
- [137] Z. Li, F. Lu, J. Zou, L. Hu, and H. Zhang, "Generalized few-shot meets remote sensing: Discovering novel classes in land cover mapping via hybrid semantic segmentation framework," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 2744–2754.
- [138] Y. Zhang, W. Li, M. Zhang, Y. Qu, R. Tao, and H. Qi, "Topological structure and semantic information transfer network for cross-scene hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 6, pp. 2817–2830, Jun. 2023.
- [139] Y. Zhang, W. Li, M. Zhang, S. Wang, R. Tao, and Q. Du, "Graph information aggregation cross-domain few-shot learning for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 1912–1925, Feb. 2024.
- [140] X. Zhao, M. Zhang, R. Tao, W. Li, W. Liao, and W. Philips, "Cross-domain classification of multisource remote sensing data using fractional fusion and spatial-spectral domain adaptation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5721–5733, 2022.
- [141] Y. Huang, J. Peng, W. Sun, N. Chen, Q. Du, Y. Ning, and H. Su, "Two-branch attention adversarial domain adaptation network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5540813.
- [142] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.
- [143] Z. Liu, L. Ma, and Q. Du, "Class-wise distribution adaptation for unsupervised classification of hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 508–521, Jan. 2021.
- [144] J. Gao, X. Ji, G. Chen, Y. Huang, and F. Ye, "Pseudo-class distribution guided multi-view unsupervised domain adaptation for hyperspectral image classification," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 136, Feb. 2025, Art. no. 104356.
- [145] B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty, "DeepJDOT: Deep joint distribution optimal transport for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 447–463.
- [146] Y. He, K. P. Seng, L. M. Ang, B. Peng, and X. Zhao, "HyperCycleGAN: A new adversarial neural network architecture for cross-domain hyperspectral data generation," *Appl. Sci.*, vol. 15, no. 8, p. 4188, Apr. 2025.
- [147] M. Ye, Z. Meng, and Y. Qian, "Building cross-domain mapping chains from multi-CycleGAN for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5524117.
- [148] Z. Xin, Z. Li, M. Xu, L. Wang, G. Ren, J. Wang, and Y. Hu, "Feature disentanglement based domain adaptation network for cross-scene coastal wetland hyperspectral image classification," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 129, May 2024, Art. no. 103850.
- [149] J. Feng, T. Zhang, J. Zhang, R. Shang, W. Dong, G. Shi, and L. Jiao, "S4DL: Shift-sensitive spatial-spectral disentangling learning for hyperspectral image unsupervised domain adaptation," 2024, *arXiv:2408.15263*.
- [150] O. Tasar, Y. Tarabalka, A. Giros, P. Alliez, and S. Clerc, "StandardGAN: Multi-source domain adaptation for semantic segmentation of very high resolution satellite images by data standardization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 747–756.
- [151] D. Wittich and F. Rottensteiner, "Appearance based deep domain adaptation for the classification of aerial images," *ISPRS J. Photogramm. Remote Sens.*, vol. 180, pp. 82–102, Oct. 2021.
- [152] D. Peng, H. Guan, Y. Zang, and L. Bruzzone, "Full-level domain adaptation for building extraction in very-high-resolution optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.

- [153] X. Lu, Y. Zhong, Z. Zheng, and J. Wang, "Cross-domain road detection based on global-local adversarial learning framework from very high resolution satellite imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 180, pp. 296–312, Oct. 2021.
- [154] L. Wu, M. Lu, and L. Fang, "Deep covariance alignment for domain adaptive remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5620811.
- [155] Z. Zhang, K. Doi, A. Iwasaki, and G. Xu, "Unsupervised domain adaptation of high-resolution aerial images via correlation alignment and self training," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 4, pp. 746–750, Apr. 2021.
- [156] A. Zheng, M. Wang, C. Li, J. Tang, and B. Luo, "Entropy guided adversarial domain adaptation for aerial image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
- [157] X. Chen, S. Pan, and Y. Chong, "Unsupervised domain adaptation for remote sensing image semantic segmentation using region and category adaptive domain discriminator," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4412913.
- [158] W. Li, H. Gao, Y. Su, and B. M. Momanyi, "Unsupervised domain adaptation for remote sensing semantic segmentation with transformer," *Remote Sens.*, vol. 14, no. 19, p. 4942, Oct. 2022.
- [159] J. Wang, A. Ma, Y. Zhong, Z. Zheng, and L. Zhang, "Cross-sensor domain adaptation for high spatial resolution urban land-cover mapping: From airborne to spaceborne imagery," *Remote Sens. Environ.*, vol. 277, Aug. 2022, Art. no. 113058.
- [160] L. Hoyer, D. Dai, and L. Van Gool, "DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9914–9925.
- [161] J. Zhu, Y. Guo, G. Sun, L. Yang, M. Deng, and J. Chen, "Unsupervised domain adaptation semantic segmentation of high-resolution remote sensing imagery with invariant domain-level prototype memory," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5603518.
- [162] A. Ma, C. Zheng, J. Wang, and Y. Zhong, "Domain adaptive land-cover classification via local consistency and global diversity," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5606317.
- [163] B. Zhang, T. Chen, and B. Wang, "Curriculum-style local-to-global adaptation for cross-domain remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.
- [164] Y. Zhang, W. Li, W. Sun, R. Tao, and Q. Du, "Single-source domain expansion network for cross-scene hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 32, pp. 1498–1512, 2023.
- [165] H. Zhao, J. Zhang, L. Lin, J. Wang, S. Gao, and Z. Zhang, "Locally linear unbiased randomization network for cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5526512.
- [166] L. Dong, J. Geng, and W. Jiang, "Spectral-spatial enhancement and causal constraint for hyperspectral image cross-scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5507013.
- [167] H. Cai, H. Yu, Y. Liu, X. An, C. Xiao, and J. Hu, "Domain generalization based on dynamic style enhancement and dual consistency constraint network for hyperspectral image cross-scene classification," *Int. J. Remote Sens.*, vol. 46, no. 9, pp. 3642–3665, May 2025.
- [168] Y. Zhang, M. Zhang, W. Li, S. Wang, and R. Tao, "Language-aware domain generalization network for cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501312.
- [169] X. Li, C. Jin, Y. Tang, K. Xing, and X. Yu, "Semantic-aware co-parallel network for cross-scene hyperspectral image classification," *Sensors*, vol. 25, no. 21, p. 6688, Nov. 2025.
- [170] B. Qin, S. Feng, C. Zhao, B. Xi, W. Li, and R. Tao, "FDGNet: Frequency disentanglement and data geometry for domain generalization in cross-scene hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 6, pp. 10297–10310, Jun. 2025.
- [171] J. Gao, X. Ji, F. Ye, and G. Chen, "Invariant semantic domain generalization shuffle network for cross-scene hyperspectral image classification," *Expert Syst. Appl.*, vol. 273, May 2025, Art. no. 126818.
- [172] Z. Han, C. Zhang, L. Gao, Z. Zeng, M. K. Ng, B. Zhang, and J. Chanussot, "Multisource collaborative domain generalization for cross-scene remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5535815.
- [173] D. Wang et al., "HyperSIGMA: Hyperspectral intelligence comprehension foundation model," 2024, *arXiv:2406.11519*.
- [174] Y. Zhang, W. Li, M. Zhang, J. Han, R. Tao, and S. Liang, "SpectralX: Parameter-efficient domain generalization for spectral remote sensing foundation models," 2025, *arXiv:2508.01731*.
- [175] R. Iizuka, J. Xia, and N. Yokoya, "Frequency-based optimal style mix for domain generalization in semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4501114.
- [176] C. Liang, W. Li, Y. Dong, and W. Fu, "Single domain generalization method for remote sensing image segmentation via category consistency on domain randomization," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5616116.
- [177] Y. Wang, X. Hu, Y. Hu, Z. Sun, S. Zhang, and L. Chen, "Boosting domain generalization in remote sensing image segmentation via style mapping and general prototypical contrast," *Int. J. Comput. Vis.*, vol. 133, no. 12, pp. 8526–8545, Dec. 2025.
- [178] M. Luo, Y. Zan, K. Khoshelham, and S. Ji, "Domain generalization for semantic segmentation of remote sensing images via vision foundation model fine-tuning," *ISPRS J. Photogramm. Remote Sens.*, vol. 230, pp. 126–146, Dec. 2025.
- [179] A. Yaghmour, M. M. Crawford, and S. Prasad, "A sensor agnostic domain generalization framework for leveraging geospatial foundation models: Enhancing semantic segmentation via synergistic pseudo-labeling and generative learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2025, pp. 3047–3056.
- [180] C. Zhao, S. Chen, J. Zhang, X. Fan, and M. Liu, "CRLMDG-LM: Causal representation learning-guided multi-target domain generalization network for fine-grained landslide mapping from high-resolution remote sensing images," *Knowl.-Based Syst.*, vol. 330, Nov. 2025, Art. no. 114796.
- [181] Z. Gong, Z. Wei, D. Wang, X. Hu, X. Ma, H. Chen, Y. Jia, Y. Deng, Z. Ji, X. Zhu, X. Yang, N. Yokoya, J. Zhang, B. Du, J. Yan, and L. Zhang, "CrossEarth: Geospatial vision foundation model for domain generalizable remote sensing semantic segmentation," 2024, *arXiv:2410.22629*.
- [182] L. Mou, P. Ghamisi, and X. X. Zhu, "Unsupervised spectral-spatial feature learning via deep residual conv-deconv network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 391–406, Jan. 2018.
- [183] S. Mei, J. Ji, Y. Geng, Z. Zhang, X. Li, and Q. Du, "Unsupervised spatial-spectral feature learning by 3D convolutional autoencoder for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6808–6820, Sep. 2019.
- [184] D. Ibanez, R. Fernandez-Beltran, F. Pla, and N. Yokoya, "Masked auto-encoding spectral-spatial transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5542614.
- [185] L. Scheibenreif, M. Mommert, and D. Borth, "Masked vision transformers for hyperspectral image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 2166–2176.
- [186] Z. Cao, X. Li, and L. Zhao, "Unsupervised feature learning by autoencoder and prototypical contrastive learning for hyperspectral classification," 2020, *arXiv:2009.00953*.
- [187] X. Hu, T. Li, T. Zhou, Y. Liu, and Y. Peng, "Contrastive learning based on transformer for hyperspectral image classification," *Appl. Sci.*, vol. 11, no. 18, p. 8670, Sep. 2021.
- [188] S. Hou, H. Shi, X. Cao, X. Zhang, and L. Jiao, "Hyperspectral imagery classification based on contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.
- [189] P. Guan and E. Y. Lam, "Cross-domain contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528913.
- [190] W. Li, H. Chen, and Z. Shi, "Semantic segmentation of remote sensing images with self-supervised multitask representation learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6438–6450, 2021.
- [191] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang, Q. He, G. Yang, R. Wang, J. Lu, and K. Fu, "RingMo: A remote sensing foundation model with masked image modeling," *IEEE Trans. Geosci. Remote Sens.*, 2022, Art. no. 5612822.
- [192] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, and S. Ermon, "SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 197–211.

- [193] M. Cai, H. Chen, T. Zhang, Y. Zhuang, and L. Chen, "Consistency regularization based on masked image modeling for semisupervised remote sensing semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 17442–17460, 2024.
- [194] X. Cao, H. Lin, S. Guo, T. Xiong, and L. Jiao, "Transformer-based masked autoencoder with contrastive loss for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5524312.
- [195] V. Marsocci, S. Scardapane, and N. Komodakis, "MARE: Self-supervised multi-attention REsU-Net for semantic segmentation in remote sensing," *Remote Sens.*, vol. 13, no. 16, p. 3275, Aug. 2021.
- [196] Z. Dong, T. Liu, and Y. Gu, "Spatial and semantic consistency contrastive learning for self-supervised semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5621112.
- [197] Q. Liu, L. Xiao, J. Yang, and Z. Wei, "CNN-enhanced graph convolutional network with pixel- and superpixel-level feature fusion for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8657–8671, Oct. 2021.
- [198] H. Wu and S. Prasad, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1259–1270, Mar. 2018.
- [199] M. Seydgar, S. Rahnamayan, P. Ghamisi, and A. A. Bidgoli, "Semisupervised hyperspectral image classification using a probabilistic pseudo-label generation framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5535218.
- [200] B. Liu, X. Yu, P. Zhang, X. Tan, A. Yu, and Z. Xue, "A semi-supervised convolutional neural network for hyperspectral image classification," *Remote Sens. Lett.*, vol. 8, no. 9, pp. 839–848, Sep. 2017.
- [201] R. Hang, F. Zhou, Q. Liu, and P. Ghamisi, "Classification of hyperspectral images via multitask generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1424–1436, Feb. 2021.
- [202] L. Huang, Y. Chen, and X. He, "Spectral-spatial masked transformer with supervised and contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5508718.
- [203] W. Huang, Y. Shi, Z. Xiong, and X. X. Zhu, "AdaptMatch: Adaptive matching for semisupervised binary segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5625416.
- [204] J.-X. Wang, S.-B. Chen, C. H. Q. Ding, J. Tang, and B. Luo, "Semi-supervised semantic segmentation of remote sensing images with iterative contrastive network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [205] X. Lu, L. Jiao, F. Liu, S. Yang, X. Liu, Z. Feng, L. Li, and P. Chen, "Simple and efficient: A semisupervised learning framework for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5543516.
- [206] J. Li, B. Sun, S. Li, and X. Kang, "Semisupervised semantic segmentation of remote sensing images with consistency self-training," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.
- [207] J. Jin, W. Lu, H. Yu, X. Rong, X. Sun, and Y. Wu, "Dynamic and adaptive self-training for semi-supervised remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5639814.
- [208] J. Wang, C. H. Q. Ding, S. Chen, C. He, and B. Luo, "Semi-supervised remote sensing image semantic segmentation via consistency regularization and average update of pseudo-label," *Remote Sens.*, vol. 12, no. 21, p. 3603, Nov. 2020.
- [209] J.-X. Wang, S.-B. Chen, C. H. Q. Ding, J. Tang, and B. Luo, "RanPaste: Paste consistency and pseudo label for semisupervised remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.
- [210] L. Lv and L. Zhang, "Advancing data-efficient exploitation for semi-supervised remote sensing images semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5619213.
- [211] J. Chen, G. Chen, L. Zhang, M. Huang, J. Luo, M. Ding, and Y. Ge, "Category-sensitive semi-supervised semantic segmentation framework for land-use/land-cover mapping with optical remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 134, Nov. 2024, Art. no. 104160.
- [212] J. Chen, B. Sun, L. Wang, B. Fang, Y. Chang, Y. Li, J. Zhang, X. Lyu, and G. Chen, "Semi-supervised semantic segmentation framework with pseudo supervisions for land-use/land-cover mapping in coastal areas," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102881.
- [213] Y. Guo, F. Wang, Y. Xiang, and H. You, "Semisupervised semantic segmentation with certainty-aware consistency training for remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2900–2914, 2023.
- [214] J. Kang, Z. Wang, R. Zhu, X. Sun, R. Fernandez-Beltran, and A. Plaza, "PiCoCo: Pixelwise contrast and consistency learning for semisupervised building footprint segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10548–10559, 2021.
- [215] Y. He, J. Wang, C. Liao, B. Shan, and X. Zhou, "ClassHyPer: ClassMix-based hybrid perturbations for deep semi-supervised semantic segmentation of remote sensing imagery," *Remote Sens.*, vol. 14, no. 4, p. 879, Feb. 2022.
- [216] J. Yang, B. Du, D. Wang, and L. Zhang, "ITER: Image-to-pixel representation for weakly supervised HSI classification," *IEEE Trans. Image Process.*, vol. 33, pp. 257–272, 2024.
- [217] K. Fu, W. Lu, W. Diao, M. Yan, H. Sun, Y. Zhang, and X. Sun, "WSF-NET: Weakly supervised feature-fusion network for binary segmentation in remote sensing image," *Remote Sens.*, vol. 10, no. 12, p. 1970, Dec. 2018.
- [218] Y. Cao and X. Huang, "A coarse-to-fine weakly supervised learning method for green plastic cover segmentation using high-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 188, pp. 157–176, Jun. 2022.
- [219] F. Fang, D. Zheng, S. Li, Y. Liu, L. Zeng, J. Zhang, and B. Wan, "Improved pseudomasks generation for weakly supervised building extraction from high-resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1629–1642, 2022.
- [220] Z. Li, X. Zhang, P. Xiao, and Z. Zheng, "On the effectiveness of weakly supervised semantic segmentation for building extraction from high-resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3266–3281, 2021.
- [221] J. Chen, F. He, Y. Zhang, G. Sun, and M. Deng, "SPMF-Net: Weakly supervised building segmentation by combining superpixel pooling and multi-scale feature fusion," *Remote Sens.*, vol. 12, no. 6, p. 1049, Mar. 2020.
- [222] J. Iqbal and M. Ali, "Weakly-supervised domain adaptation for built-up region segmentation in aerial and satellite imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 263–275, Sep. 2020.
- [223] M. Lu, L. Fang, M. Li, B. Zhang, Y. Zhang, and P. Ghamisi, "NFANet: A novel method for weakly supervised water extraction from high-resolution remote-sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5617114.
- [224] Y. Wei and S. Ji, "Scribble-based weakly supervised deep learning for road surface extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602312.
- [225] Z. Li, H. Zhang, F. Lu, R. Xue, G. Yang, and L. Zhang, "Breaking the resolution barrier: A low-to-high network for large-scale high-resolution land-cover mapping using low-resolution labels," *ISPRS J. Photogramm. Remote Sens.*, vol. 192, pp. 244–267, Oct. 2022.
- [226] V. Slavkovikj, S. Verstockt, W. De Neve, S. Van Hoecke, and R. Van de Walle, "Hyperspectral image classification with convolutional neural networks," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1159–1162.
- [227] Y. Duan, F. Liu, L. Jiao, P. Zhao, and L. Zhang, "SAR image segmentation based on convolutional-wavelet neural network and Markov random field," *Pattern Recognit.*, vol. 64, pp. 255–267, Apr. 2017.
- [228] A. Wang, M. Wang, H. Wu, K. Jiang, and Y. Iwahori, "A novel LiDAR data classification algorithm combined CapsNet with ResNet," *Sensors*, vol. 20, no. 4, p. 1151, Feb. 2020.
- [229] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.
- [230] Y. Mao, N. Wang, W. Zhou, and H. Li, "Joint inductive and transductive learning for video object segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9659.
- [231] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
- [232] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., Red Hook, NY, USA: Curran Associates, 2015, pp. 1–9.

- [233] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.
- [234] S. Luo, Y. Li, P. Gao, Y. Wang, and S. Serikawa, "Meta-seg: A survey of meta-learning for image segmentation," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108586.
- [235] D. Wang, L. Ren, X. Sun, L. Gao, and J. Chanussot, "Nonlocal and local feature-coupled self-supervised network for hyperspectral anomaly detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 6981–6993, 2025.
- [236] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15979–15988.
- [237] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 1096–1103.
- [238] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1691–1703.
- [239] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 22243–22255.
- [240] X. Qi, Y. Mao, Y. Zhang, Y. Deng, H. Wei, and L. Wang, "PICS: Paradigms integration and contrastive selection for semisupervised remote sensing images semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5602119.
- [241] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, Jan. 2018.
- [242] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
- [243] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. Van-Den Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 36–43.
- [244] X. Du, X. Zheng, X. Lu, and A. A. Douudkin, "Multisource remote sensing data classification with graph fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10062–10072, Dec. 2021.
- [245] Y. Yang, D. Zhu, T. Qu, Q. Wang, F. Ren, and C. Cheng, "Single-stream CNN with learnable architecture for multisource remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5409218.
- [246] Y. Yu, T. Jiang, J. Gao, H. Guan, D. Li, S. Gao, E. Tang, W. Wang, P. Tang, and J. Li, "CapViT: Cross-context capsule vision transformers for land cover classification with airborne multispectral LiDAR data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 111, Jul. 2022, Art. no. 102837.
- [247] Y. Dong, Z. Yang, Q. Liu, R. Zuo, and Z. Wang, "Fusion of GaoFen-5 and Sentinel-2B data for lithological mapping using vision transformer dynamic graph convolutional network," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 129, May 2024, Art. no. 103780.
- [248] X. Zhao, R. Tao, W. Li, H.-C. Li, Q. Du, W. Liao, and W. Philips, "Joint classification of hyperspectral and LiDAR data using hierarchical random walk and deep CNN architecture," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7355–7370, Oct. 2020.
- [249] Y. Chen, C. Li, P. Ghamisi, X. Jia, and Y. Gu, "Deep fusion of remote sensing data for accurate classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1253–1257, Aug. 2017.
- [250] H. Li, P. Ghamisi, U. Soergel, and X. Zhu, "Hyperspectral and LiDAR fusion using deep three-stream convolutional neural networks," *Remote Sens.*, vol. 10, no. 10, p. 1649, Oct. 2018.
- [251] H.-C. Li, W.-Y. Wang, L. Pan, W. Li, Q. Du, and R. Tao, "Robust capsule network based on maximum correntropy criterion for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 738–751, 2020.
- [252] Z. Xue, B. Liu, A. Yu, X. Yu, P. Zhang, and X. Tan, "Self-supervised feature representation and few-shot land cover classification of multimodal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5541618.
- [253] Y. Cai, Z. Zhang, P. Ghamisi, B. Rasti, X. Liu, and Z. Cai, "Transformer-based contrastive prototypical clustering for multimodal remote sensing data," *Inf. Sci.*, vol. 649, Nov. 2023, Art. no. 119655.
- [254] X. Li, L. Lei, Y. Sun, M. Li, and G. Kuang, "Multimodal bilinear fusion network with second-order attention-based channel selection for land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1011–1026, 2020.
- [255] X. Zhao, M. Zhang, R. Tao, W. Li, W. Liao, L. Tian, and W. Philips, "Fractional Fourier image transformer for multimodal remote sensing data classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 2314–2326, Feb. 2024.
- [256] Q. Liu, T. Huang, Y. Dong, and W. Xiang, "Enhancing oil spill detection with controlled random sampling: A multimodal fusion approach using SAR and HSI imagery," *Remote Sens. Appl., Soc. Environ.*, vol. 38, pp. 1–23, Apr. 2025.
- [257] F. Guo, Z. Li, Q. Meng, L. Wang, and J. Zhang, "Dual graph convolution joint dense networks for hyperspectral and LiDAR data classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2022, pp. 1141–1144.
- [258] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018.
- [259] S. K. Roy, A. Deria, D. Hong, M. Ahmad, A. Plaza, and J. Chanussot, "Hyperspectral and LiDAR data classification using joint CNNs and morphological feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5530416.
- [260] K. Yue, L. Yang, R. Li, W. Hu, F. Zhang, and W. Li, "TreeUNet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation," *ISPRS J. Photogramm. Remote Sens.*, vol. 156, pp. 1–13, Oct. 2019.
- [261] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet—A: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.
- [262] Z. Xie, U. K. Haritashya, V. K. Asari, M. P. Bishop, J. S. Kargel, and T. H. Aspiras, "GlacierNet2: A hybrid multi-model learning architecture for Alpine glacier mapping," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102921.
- [263] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, "Dense semantic labeling of very-high-resolution aerial imagery and LiDAR with fully-convolutional neural networks and higher-order CRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1561–1570.
- [264] N. Audebert, B. Le Saux, and S. Lefèvre, "Semantic segmentation of Earth observation data using multimodal and multi-scale deep networks," in *Proc. Asian Conf. Comput. Vis. Cham, Switzerland: Springer*, 2016, pp. 180–196.
- [265] X. Liu, H. Zou, S. Wang, Y. Lin, and X. Zuo, "Joint network combining dual-attention fusion modality and two specific modalities for land cover classification using optical and SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3236–3250, 2024.
- [266] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, and Q. Liu, "Classification of hyperspectral and LiDAR data using coupled CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4939–4950, Jul. 2020.
- [267] W. Song, Z. Gao, and Y. Zhang, "Discriminative feature extraction and fusion for classification of hyperspectral and LiDAR data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2022, pp. 2271–2274.
- [268] M. Zhang, X. Zhao, W. Li, Y. Zhang, R. Tao, and Q. Du, "Cross-scene joint classification of multisource data with multilevel domain adaption network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 8, pp. 11514–11526, Aug. 2024.
- [269] X. Zhao, R. Tao, W. Li, W. Philips, and W. Liao, "Fractional Gabor convolutional network for multisource remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.
- [270] Y. Zhang, S. Xu, D. Hong, H. Gao, C. Zhang, M. Bi, and C. Li, "Multimodal transformer network for hyperspectral and LiDAR classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514317.
- [271] J. Li, Y. Liu, R. Song, W. Liu, Y. Li, and Q. Du, "HyperMLP: Superpixel prior and feature aggregated perceptron networks for hyperspectral and LiDAR hybrid classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5505614.
- [272] H. Gao, H. Feng, Y. Zhang, S. Xu, and B. Zhang, "AMSSE-Net: Adaptive multiscale spatial-spectral enhancement network for classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5531317.

- [273] H. Zhang, J. Yao, L. Ni, L. Gao, and M. Huang, "Multimodal attention-aware convolutional neural networks for classification of hyperspectral and LiDAR data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3635–3644, 2023.
- [274] X. Du, X. Zheng, X. Lu, and X. Wang, "Hyperspectral and LiDAR representation with spectral-spatial graph network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 9231–9245, 2023.
- [275] K. Ni, D. Wang, Z. Zheng, and P. Wang, "MHST: Multiscale head selection transformer for hyperspectral and LiDAR classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 5470–5483, 2024.
- [276] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.
- [277] C. Peng, Y. Li, L. Jiao, Y. Chen, and R. Shang, "Densely based multi-scale and multi-modal fully convolutional networks for high-resolution remote-sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2612–2626, Aug. 2019.
- [278] Q. Wang, W. Chen, Z. Huang, H. Tang, and L. Yang, "MultiSenseSeg: A cost-effective unified multimodal semantic segmentation model for remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4703724.
- [279] X. He, M. Wang, T. Liu, L. Zhao, and Y. Yue, "SFAF-MA: Spatial feature aggregation and fusion with modality adaptation for RGB-thermal semantic segmentation," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–10, 2023.
- [280] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, "FusAtNet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and LiDAR classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 416–425.
- [281] W. Dong, T. Zhang, J. Qu, S. Xiao, T. Zhang, and Y. Li, "Multibranch feature fusion network with self- and cross-guided attention for hyperspectral and LiDAR classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5530612.
- [282] D. Xiu, Z. Pan, Y. Wu, and Y. Hu, "MAGE: Multisource attention network with discriminative graph and informative entities for classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5539714.
- [283] J. Wang, J. Li, Y. Shi, J. Lai, and X. Tan, "AM³Net: Adaptive mutual-learning-based multimodal data fusion network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5411–5426, Aug. 2022.
- [284] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5515620.
- [285] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514415.
- [286] G. Zhao, Q. Ye, L. Sun, Z. Wu, C. Pan, and B. Jeon, "Joint classification of hyperspectral and LiDAR data using a hierarchical CNN and transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5500716.
- [287] M. Zhang, W. Li, Y. Zhang, R. Tao, and Q. Du, "Hyperspectral and LiDAR data classification based on structural optimization transmission," *IEEE Trans. Cybern.*, vol. 53, no. 5, pp. 3153–3164, May 2023.
- [288] J. Wang and X. Tan, "Mutually beneficial transformer for multimodal data fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7466–7479, Dec. 2023.
- [289] S. K. Roy, A. Sukul, A. Jamali, J. M. Haut, and P. Ghamisi, "Cross hyperspectral and LiDAR attention transformer: An extended self-attention for land use and land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5512815.
- [290] J. Li, Y. Ma, R. Song, B. Xi, D. Hong, and Q. Du, "A triplet semisupervised deep network for fusion classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5540513.
- [291] C. Li, R. Hang, and B. Rasti, "EMFNet: Enhanced multisource fusion network for land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4381–4389, 2021.
- [292] S. Fang, K. Li, and Z. Li, "S²ENet: Spatial-spectral cross-modal enhancement network for classification of hyperspectral and LiDAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [293] T. Zhang, S. Xiao, W. Dong, J. Qu, and Y. Yang, "A mutual guidance attention-based multi-level fusion network for hyperspectral and LiDAR classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [294] X. Wang, Y. Feng, R. Song, Z. Mu, and C. Song, "Multi-attentive hierarchical dense fusion net for fusion classification of hyperspectral and LiDAR data," *Inf. Fusion*, vol. 82, pp. 1–18, Jun. 2022.
- [295] Y. Feng, L. Song, L. Wang, and X. Wang, "DSHFNet: Dynamic scale hierarchical fusion network based on multiattention for hyperspectral image and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5522514.
- [296] Y. Zhang, Y. Peng, B. Tu, and Y. Liu, "Local information interaction transformer for hyperspectral and LiDAR data classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1130–1143, 2023.
- [297] K. Li, D. Wang, X. Wang, G. Liu, Z. Wu, and Q. Wang, "Mixing self-attention and convolution: A unified framework for multisource remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5523216.
- [298] Y. Zhang, H. Gao, J. Zhou, C. Zhang, P. Ghamisi, S. Xu, C. Li, and B. Zhang, "A cross-modal feature aggregation and enhancement network for hyperspectral and LiDAR joint classification," *Expert Syst. Appl.*, vol. 258, Dec. 2024, Art. no. 125145.
- [299] F. Gao, X. Jin, X. Zhou, J. Dong, and Q. Du, "MSFMamba: Multiscale feature fusion state space model for multisource remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5504116.
- [300] B. Ren, S. Ma, B. Hou, D. Hong, J. Chanussot, J. Wang, and L. Jiao, "A dual-stream high resolution network: Deep fusion of GF-2 and GF-3 data for land cover classification," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102896.
- [301] P. Zhang, P. Du, C. Lin, X. Wang, E. Li, Z. Xue, and X. Bai, "A hybrid attention-aware fusion network (HAFNet) for building extraction from high-resolution imagery and LiDAR data," *Remote Sens.*, vol. 12, no. 22, p. 3764, Nov. 2020.
- [302] X. Yang, S. Li, Z. Chen, J. Chanussot, X. Jia, B. Zhang, B. Li, and P. Chen, "An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 177, pp. 238–262, Jul. 2021.
- [303] L. Ferrari, F. Dell'Acqua, P. Zhang, and P. Du, "Integrating EfficientNet into an HAFNet structure for building mapping in high-resolution optical Earth observation data," *Remote Sens.*, vol. 13, no. 21, p. 4361, Oct. 2021.
- [304] X. Ma, X. Zhang, and M.-O. Pun, "A crossmodal multiscale fusion network for semantic segmentation of remote sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3463–3474, 2022.
- [305] Y. Sun, Z. Fu, C. Sun, Y. Hu, and S. Zhang, "Deep multimodal fusion network for semantic segmentation using remote sensing image and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.
- [306] X. Li, G. Zhang, H. Cui, S. Hou, S. Wang, X. Li, Y. Chen, Z. Li, and L. Zhang, "MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 106, Feb. 2022, Art. no. 102638.
- [307] X. Ma, X. Zhang, M.-O. Pun, and M. Liu, "A multilevel multimodal fusion transformer for remote sensing semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5403215.
- [308] X. Li, L. Lei, Y. Sun, M. Li, and G. Kuang, "Collaborative attention-based heterogeneous gated fusion network for land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 3829–3845, May 2021.
- [309] A. Zhang, L. Jia, J. Wang, and C. Wang, "SAR image classification using gated channel attention based convolutional neural network," *Remote Sens.*, vol. 15, no. 2, p. 362, Jan. 2023.
- [310] Z. Li, A. Zhang, G. Sun, Z. Han, and X. Jia, "Automatic impervious surface mapping in subtropical China via a terrain-guided gated fusion network," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 127, Mar. 2024, Art. no. 103608.
- [311] J. Huang, X. Zhang, Q. Xin, Y. Sun, and P. Zhang, "Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network," *ISPRS J. Photogramm. Remote Sens.*, vol. 151, pp. 91–105, May 2019.
- [312] C. Yue, Y. Zhang, J. Yan, Z. Luo, Y. Liu, and P. Guo, "BCLNet: Boundary contrastive learning with gated attention feature fusion and multi-branch spatial-channel reconstruction for land use classification," *Knowl.-Based Syst.*, vol. 302, Oct. 2024, Art. no. 112387.
- [313] W. Zhou, J. Jin, J. Lei, and J.-N. Hwang, "CEGFNet: Common extraction and gate fusion network for scene parsing of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2022.

- [314] H. Hosseinpour, F. Samadzadegan, and F. D. Javan, "CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 96–115, Feb. 2022.
- [315] X. Geng, L. Jiao, L. Li, F. Liu, X. Liu, S. Yang, and X. Zhang, "Multisource joint representation learning fusion classification for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4406414.
- [316] K. Wei, J. Dai, D. Hong, and Y. Ye, "MGFNet: An MLP-dominated gated fusion network for semantic segmentation of high-resolution multi-modal remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 135, Dec. 2024, Art. no. 104241.
- [317] W. Kang, Y. Xiang, F. Wang, and H. You, "CFNet: A cross fusion network for joint land cover classification using optical and SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1562–1574, 2022.
- [318] K. Ding, T. Lu, W. Fu, S. Li, and F. Ma, "Global-local transformer network for HSI and LiDAR data joint classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5541213.
- [319] D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, "Deep encoder-decoder networks for classification of hyperspectral and LiDAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [320] M. Zhang, W. Li, R. Tao, H. Li, and Q. Du, "Information fusion for classification of hyperspectral and LiDAR data using IP-CNN," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.
- [321] T. Lu, K. Ding, W. Fu, S. Li, and A. Guo, "Coupled adversarial learning for fusion classification of hyperspectral and LiDAR data," *Inf. Fusion*, vol. 93, pp. 118–131, May 2023.
- [322] J. Zhang, J. Lei, W. Xie, G. Yang, D. Li, and Y. Li, "Multimodal informative ViT: Information aggregation and distribution for hyperspectral and LiDAR classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7643–7656, Aug. 2024.
- [323] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2022.
- [324] M. Zhang, W. Li, Q. Du, L. Gao, and B. Zhang, "Feature extraction for classification of hyperspectral and LiDAR data using patch-to-patch CNN," *IEEE Trans. Cybern.*, vol. 50, no. 1, pp. 100–111, Jan. 2020.
- [325] S. Pande and B. Banerjee, "Self-supervision assisted multimodal remote sensing image classification with coupled self-looping convolution networks," *Neural Netw.*, vol. 164, pp. 1–20, Jul. 2023.
- [326] D. Hong, J. Yao, D. Meng, Z. Xu, and J. Chanussot, "Multimodal GANs: Toward crossmodal hyperspectral–multispectral image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5103–5113, Jun. 2021.
- [327] Y. Chen, M. Zhao, and L. Bruzzone, "Incomplete multimodal learning for remote sensing data fusion," 2023, *arXiv:2304.11381*.
- [328] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.
- [329] R. Hang, X. Qian, and Q. Liu, "Cross-modality contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5532812.
- [330] X. Li, G. Zhang, H. Cui, S. Hou, Y. Chen, Z. Li, H. Li, and H. Wang, "Progressive fusion learning: A multimodal joint segmentation framework for building extraction from optical and SAR images," *ISPRS J. Photogramm. Remote Sens.*, vol. 195, pp. 178–191, Jan. 2023.
- [331] W. Li, K. Sun, W. Li, J. Wei, S. Miao, S. Gao, and Q. Zhou, "Aligning semantic distribution in fusing optical and SAR images for land use classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 199, pp. 272–288, May 2023.
- [332] D. Gupta, A. Golder, R. Zhu, K. Cui, W. Tang, F. Yang, O. Csillik, S. Alqahtani, and V. Paul Pauca, "MoSAiC: Multimodal multi-label supervision-aware contrastive learning for remote sensing," 2025, *arXiv:2507.08683*.
- [333] T. Chen, J. Pan, J. Xu, J. Hu, L. Cao, and J. Li, "MDCA-Net: A multi-directional alignment and dynamic context aggregation network for optical and SAR image fusion," *Geo-spatial Inf. Sci.*, vol. 29, no. 2, pp. 1–19, Mar. 2026.
- [334] J. Li, Z. Wang, N. Xu, and Z. You, "Semantic segmentation with scale alignment and contextual information fusion for multimodal remote sensing images," *Inf. Fusion*, vol. 126, Feb. 2026, Art. no. 103671.
- [335] Y. Hu, H. He, and L. Weng, "Hyperspectral and LiDAR data land-use classification using parallel transformers," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 703–706.
- [336] P. Duan, X. Kang, P. Ghamisi, and S. Li, "Hyperspectral remote sensing benchmark database for oil spill detection with an isolation forest-guided unsupervised detector," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5509711.
- [337] Y. Zhong, X. Wang, Y. Xu, S. Wang, T. Jia, X. Hu, J. Zhao, L. Wei, and L. Zhang, "Mini-UAV-borne hyperspectral remote sensing: From observation and processing to applications," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 4, pp. 46–62, Dec. 2018.
- [338] C. Yi, L. Zhang, X. Zhang, W. Yueming, Q. Wenchao, T. Senlin, and P. Zhang, "Aerial hyperspectral remote sensing classification dataset of Xiongan new area (matiwang village)," *Nat. Remote Sens. Bull.*, vol. 24, no. 11, pp. 1299–1306, 2020.
- [339] A. Rangnekar, N. Mokashi, E. J. Ientilucci, C. Kanan, and M. J. Hoffman, "AeroRIT: A new scene for hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 8116–8124, Nov. 2020.
- [340] J. Li, X. Huang, and L. Tu, "WHU-OHS: A benchmark dataset for large-scale hersepctral image classification," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 113, Sep. 2022, Art. no. 103022.
- [341] H. Alemohammad and K. Booth, "LandCoverNet: A global benchmark land cover classification training dataset," 2020, *arXiv:2012.03111*.
- [342] M. Krestenitis, G. Orfanidis, K. Ioannidis, K. Avgerinakis, S. Vrochidis, and I. Kompatsiaris, "Oil spill identification from satellite images using deep neural networks," *Remote Sens.*, vol. 11, no. 15, p. 1762, Jul. 2019.
- [343] Q. Zhu, Y. Zhang, Z. Li, X. Yan, Q. Guan, Y. Zhong, L. Zhang, and D. Li, "Oil spill contextual and boundary-supervised detection network based on marine SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2022.
- [344] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3226–3229.
- [345] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "DeepGlobe 2018: A challenge to parse the Earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.
- [346] S. M. Azimi, C. Henry, L. Sommer, A. Schumann, and E. Vig, "SkyScapes–fine-grained semantic understanding of aerial scenes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7392–7402.
- [347] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "UAVid: A semantic segmentation dataset for UAV imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 165, pp. 108–119, Jul. 2020.
- [348] A. Boguszewski, D. Batorski, N. Ziemia-Jankowska, T. Dziedzic, and A. Zambrycka, "LandCoverAI: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1102–1110.
- [349] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," 2022, *arXiv:2110.08733*.
- [350] M. Rahnemounfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy, "FloodNet: A high resolution aerial imagery dataset for post flood scene understanding," *IEEE Access*, vol. 9, pp. 89644–89654, 2021.
- [351] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, "Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model," *ISPRS J. Photogramm. Remote Sens.*, vol. 178, pp. 68–80, Aug. 2021.
- [352] D. Hong, B. Zhang, H. Li, Y. Li, J. Yao, C. Li, M. Werner, J. Chanussot, A. Zipf, and X. X. Zhu, "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, Dec. 2023, Art. no. 113856.
- [353] J. Hu, R. Liu, D. Hong, A. Camero, J. Yao, M. Schneider, F. Kurz, K. Segl, and X. X. Zhu, "MDAS: A new multimodal benchmark dataset for remote sensing," *Earth Syst. Sci. Data*, vol. 15, no. 1, pp. 113–131, Jan. 2023.
- [354] M. P. Barbato, F. Piccoli, and P. Napoletano, "Ticino: A multi-modal remote sensing dataset for semantic segmentation," *Expert Syst. Appl.*, vol. 249, Sep. 2024, Art. no. 123600.

- [355] N. Li, S. Jiang, J. Xue, S. Ye, and S. Jia, "Texture-aware self-attention model for hyperspectral tree species classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5502215.
- [356] Z. Yuan, L. Mou, Y. Hua, and X. X. Zhu, "RRSIS: Referring remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5613312.
- [357] S. Liu, Y. Ma, X. Zhang, H. Wang, J. Ji, X. Sun, and R. Ji, "Rotated multi-scale interaction network for referring remote sensing image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 26658–26668.
- [358] Z. Dong, Y. Sun, T. Liu, W. Zuo, and Y. Gu, "Cross-modal bidirectional interaction model for referring remote sensing image segmentation," 2024, *arXiv:2410.08613*.
- [359] K. Li, Z. Xin, L. Pang, C. Pang, Y. Deng, J. Yao, G. Xia, D. Meng, Z. Wang, and X. Cao, "SegEarth-r1: Geospatial pixel reasoning via large language model," 2025, *arXiv:2504.09644*.
- [360] D. Wang, L. Gao, Y. Qu, X. Sun, and W. Liao, "Frequency-to-spectrum mapping GAN for semisupervised hyperspectral anomaly detection," *CAAI Trans. Intell. Technol.*, vol. 8, no. 4, pp. 1258–1273, Dec. 2023.
- [361] D. Wang, L. Zhuang, L. Gao, X. Sun, X. Zhao, and A. Plaza, "Sliding dual-window-inspired reconstruction network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5504115.



Engineering and the Centre for AI and Data Science Innovation, James Cook University, Cairns, QLD, Australia.

His research interests include remote image segmentation, hyperspectral image processing, multimodal remote sensing fusion, and deep learning.

QUANWEI LIU (Student Member, IEEE) received the B.S. degree in exploration technology and engineering from Henan University of Engineering, Zhengzhou, China, in 2020, and the M.S. degree in resources and environment from China University of Geosciences, Wuhan, China, in 2023. He is currently pursuing the Ph.D. degree in multimodal image fusion, hyperspectral image classification, and remote-sensing image segmentation with the College of Science and



Research Fellow. He is currently a Senior Lecturer with James Cook University (JCU), Cairns, QLD, Australia, where he is the Co-Deputy Director of the Center for AI and Data Science Innovation (CADSI), JCU. He has received several awards, including the Australian Postgraduate Award, the Engineering Research Award from the University of New South Wales, the Best Paper Award at the IEEE Wireless Communications and Networking Conference, in 2011, the IEEE Outstanding Leadership Award, in 2022, the IEEE Access Outstanding Associate Editor Recognition, in 2023, 2024, and 2025, two Springer Nature Editor of Distinction Awards 2026 (Editorial Contribution Award and Author + Service Award), and the JCU Citation for Outstanding Contribution to Student Learning, in 2022. He serves as the Vice Chair of the IEEE Northern Australia Section and the Chair of the local MTT-S/ComSoc Chapter. He has also held leadership roles at international conferences, including the General Co-Chair, the Publicity Co-Chair, the Workshop Co-Chair, the Publication Co-Chair, the Technical Program Committee Chair, the Program Vice Chair, and the Symposium Chair. He is an Associate Editor of *Scientific Reports*, *IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY*, *IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING*, *IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY*, *IEEE ACCESS*, *Frontiers in Artificial Intelligence*, and *IET Communications*.

TAO HUANG (Senior Member, IEEE) received the B.Eng. degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2003, the M.Eng. degree in sensor system signal processing from The University of Adelaide, Adelaide, SA, Australia, in 2007, and the Ph.D. degree in electrical engineering from the University of New South Wales, Sydney, NSW, Australia, in 2016. He was an Endeavor Australia Cheung Kong



He is currently a Postdoctoral Research Associate with the University of Wisconsin–Madison, Madison, WI, USA. Her research interests include deep learning, satellite image processing, hyperspectral image interpretation, and multimodal data analysis.

JIAQI YANG (Member, IEEE) received the B.S. degree in geographic information science from Dalian Maritime University, Dalian, China, in 2019, the M.E. degree in geomatics engineering from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2021, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, in 2024.



consecutive years, from 2020 to 2025, he has consistently ranked among Stanford University's world's top 2% Scientists for both his single-year and career-long impacts. Due to his instrumental leadership in establishing Australia's first accredited Internet of Things Engineering degree program, he was inducted into the Percy Foundation's Hall of Fame, in October 2018. He has published more than 450 peer-reviewed articles, including three books, and nearly 300 journal articles. His research interests include the Internet of Things, wireless communications, machine learning for IoT data analytics, and computer vision. He is a TEDx speaker and an elected fellow of IET in U.K. and Engineers Australia. He has been awarded several prestigious fellowship titles. He was named a Queensland International Fellow by the Queensland Government of Australia, from 2010 to 2011, an Endeavour Research Fellow by the Commonwealth Government of Australia, from 2012 to 2013, a Smart Futures Fellow by the Queensland Government of Australia, from 2012 to 2015, and a JSPS Invitational Fellow jointly by the Australian Academy of Science and Japanese Society for Promotion of Science, from 2014 to 2015. He received the TNQ Innovation Award, in 2016, the Pearcey Entrepreneurship Award, in 2017, and the Engineers Australia Cairns Engineer of the Year, in 2017. He was a co-recipient of the four Best Paper Awards at WiSATS'2019, WCSP'2015, IEEE WCNC'2011, and ICWMC'2009. He has served in a large number of international conferences in the capacity of General Co-Chair, the TPC Co-Chair, and the Symposium Chair. He was the Vice Chair of the IEEE Northern Australia Section, from 2016 to 2020. He is an Associate Editor of *IEEE COMMUNICATIONS SURVEYS AND TUTORIALS*, *IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY*, *IEEE INTERNET OF THINGS JOURNAL*, *IEEE ACCESS*, and *Scientific Reports* (Nature).

WEI XIANG (Senior Member, IEEE) is currently a La Trobe Distinguished Professor and the Cisco Research Chair of AI and IoT of La Trobe University. He founded two world-first research centers, namely the CiscoLa Trobe Centre for AI and IoT, in 2020, and the Australian Centre for AI and Medical Innovation (ACAMI), in 2024. Previously, he was the Foundation Chair and the Head of Discipline of IoT Engineering of James Cook University, Cairns, Australia. For the past six

...