Remote Sensing Forestry Similarity Convolution

Shikuan Wang[1] Yuangong Chen[2] Jianzhou Gong[1*] Lingyi Meng[3]
Mengquan Wu[3] Longxing Liu[4] Haiwei Yuan[1] Mingbin Guo[1]
[1]Guangzhou University [2]The Hong Kong Polytechnic University
[3]Ludong University [4]Nanjing University

Abstract

Recent advancements in convolutional neural networks (CNNs) have significantly propelled the field of remote sensing forestry mapping. However, traditional convolution operations exhibit inherent limitations in extracting complex forest features: their fixed receptive fields struggle to accommodate multi-scale forest attributes, and their insufficient focus on background information impairs the overall feature representation. To address these challenges, we propose Similar Convolution (SimConv), which introduces dynamic convolution kernel size selection by modeling feature relationships. SimConv adaptively adjusts the receptive field based on the semantic relevance of input features, enhancing the capture of forestry background information and improving the distinction between target features. Building upon this, we introduce SIMNet, a feature extraction network that integrates SimConv at its core. Experimental results across multiple remote sensing datasets demonstrate that SIMNet outperforms existing methods in terms of feature extraction accuracy. code in https://github.com/WangShiK/SimConv.

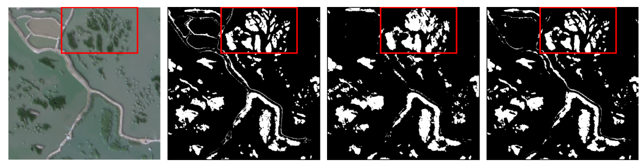

Figure 1. Mapping results of CNNs constructed using the proposed convolution and standard convolution. GT stands for Ground Truth. (a) shows CNNs constructed using standard convolution, and (b) shows CNNs constructed using our proposed SimConv. SimConv in detail mapping due to standard convolution.

1. Introduction

Forestlands are vital ecological assets, critical for global carbon cycle monitoring, biodiversity conservation, and high-precision mapping [1, 33, 34]. Amid escalating global climate change, accurate forest spatial distribution data is indispensable for effective forest resource management, ecological restoration, and sustainable development evaluation [12, 47]. Traditional forest mapping integrates multi-source remote sensing data to improve accuracy via complementary strengths [3], yet spatiotemporal resolution discrepancies, inconsistent collection protocols, and format differences often cause critical issues like registration errors and semantic inconsistencies in data fusion [56]. Worse, conventional methods (manual interpretation, rule-based classification) fail to capture forest structure complexity and heterogeneity—resulting in outcomes that cannot meet modern forestry's demands for precision, efficiency, and automation [19].

With the rapid advancement of high-resolution satellite remote sensing and widespread adoption of convolutional neural networks (CNNs), optical satellite imagery has become a foundational data source for forest resource surveys—offering sub-meter spatial resolution and rich textures [40, 41]. Its utility is well-recognized in key tasks like forest boundary delineation [3, 30, 47]. Despite CNNs' success in forestry mapping, standard convolution operations face inherent structural limitations in complex forest scenarios: they use fixed kernel sizes and cannot adaptively capture hierarchical, spatially varying feature relationships. This makes them struggle to handle the multi-scale characteristics and heterogeneous backgrounds of remote sensing imagery. Such limitations often lead models to miss critical fine-grained features, directly undermining mapping accuracy and robustness [11, 29, 38] (see Figure 1).

To address conventional convolutions' limitations in remote sensing forest mapping, we propose a domain-specific method—Similarity Convolution (SimConv). SimConv models intrinsic feature relationships spatially (x/y), then dynamically selects optimal convolution kernel sizes for adaptive, context-aware feature extraction. This enables it to capture subtle, complex features in remote sensing data, enhancing the detail and accuracy of forestry mapping out-

*Corresponding author: gongjzh@gzhu.edu.cn

puts.

The main contributions of this paper are as follows:

- We propose an efficient, task-specific feature extraction encoder based on SimConv, leveraging its dynamic receptive fields to capture more discriminative features from remote sensing imagery.
- To boost semantic integration, we introduce a novel Wavelet Feature Fusion Module (WFFM) in the decoder, which merges deep/shallow features to leverage multi-scale context for accurate mapping.
- By combining the SimConv encoder and WFFM-enhanced decoder, we build SIMNet—a model tailored for remote sensing forestry mapping that balances spatial adaptability with semantic richness for precise and efficient performance.
- Extensive evaluations on three benchmark datasets show SIMNet's consistent superiority over existing methods across diverse forestry mapping tasks, validating SimConv and SIMNet as a robust, practical framework for forest resource monitoring in complex remote sensing environments.

2. Related Works

In recent years, CNNs have driven significant breakthroughs in remote sensing image analysis, markedly advancing the precision and automation of optical imagery-based forestry mapping [15]. However, standard convolutions are limited by inherent fixed receptive fields, restricting effective capture of the complex, variable spatial structures in forested regions [26]. To address this, researchers have proposed structural enhancements to bridge the gap between static convolution mechanisms and the dynamic spatial characteristics of real-world scenes [8, 42, 55].

Forestry target recognition in remote sensing imagery presents additional challenges due to the wide range of target scales, from individual trees to expansive forest canopies. Traditional convolutional architectures with fixed receptive fields often struggle to accommodate this multi-scale variability, leading to suboptimal feature extraction performance [24]. Consequently, adaptive receptive fields and multi-scale feature perception have emerged as key research challenges in this domain.

Among existing approaches, Atrous Spatial Pyramid Pooling (ASPP, used in DeepLabv3+) [5] effectively models multi-scale context via parallel convolutions with different dilation rates. DenseASPP [48] extends this by densely stacking dilated convolutions—boosting semantic perception across continuous scales and adaptability to forest objects of different sizes. However, both rely heavily on manually designed receptive field structures and lack dynamic adaptability to targets' actual semantic scales. In complex backgrounds or overlapping scales, their generalization is limited [23].

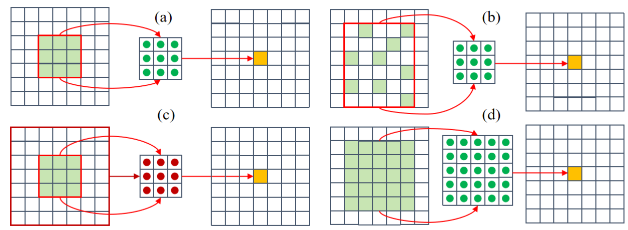

Figure 2. (a) Standard; (b) Deformable; (c) Dynamic: globally generates weights from input; (d) Proposed SimConv: adaptively selects kernel size from input.

HRNet [37] takes a different approach: it maintains parallel multi-resolution branches to preserve high-resolution spatial information—key for precise tasks like forest boundary delineation. Its simultaneous multi-scale feature learning significantly boosts accuracy and robustness. However, HRNet's high computational cost and structural complexity hinder practical deployment.

To address traditional convolution's rigidity, deformable convolution [7] uses learnable offsets to adapt kernel sampling positions to target geometry, boosting complex shape modeling. Extensions like LDConv [53] enhance spatial adaptability further, while dynamic methods (DyConv [6], ODConv [17]) generate input-dependent weights to improve contextual sensitivity. Yet these approaches still lack semantically guided receptive field adjustment—and as shown in Figure 2, kernel sizes remain effectively fixed, limiting their performance on scenes with complex semantic variability.

To address existing method shortcomings, we propose an innovative convolution operation—SimConv. It adaptively selects convolution kernel sizes based on input features for refined feature extraction and stronger semantic relevance. Targeting existing convolutions' fixed receptive fields and insufficient feature relationship modeling, SimConv solves the problem of intelligent receptive field adjustment in complex semantic scenarios. Experimental validation shows it delivers significant performance gains in remote sensing forestry mapping, offering new insights for future research.

3. Methodology

This section will provide a detailed introduction to the designs based on SimConv and SIMNet.

3.1. Overall Architecture of SIMNet

The proposed SIMNet architecture (Figure 3) introduces a hierarchical feature interaction mechanism as its core innovation. In the encoder, we design the SimBlock module based on SimConv, where the conventional pooling layers are replaced with the HWD [46] module. This enables progressive resolution reduction while preserving multi-

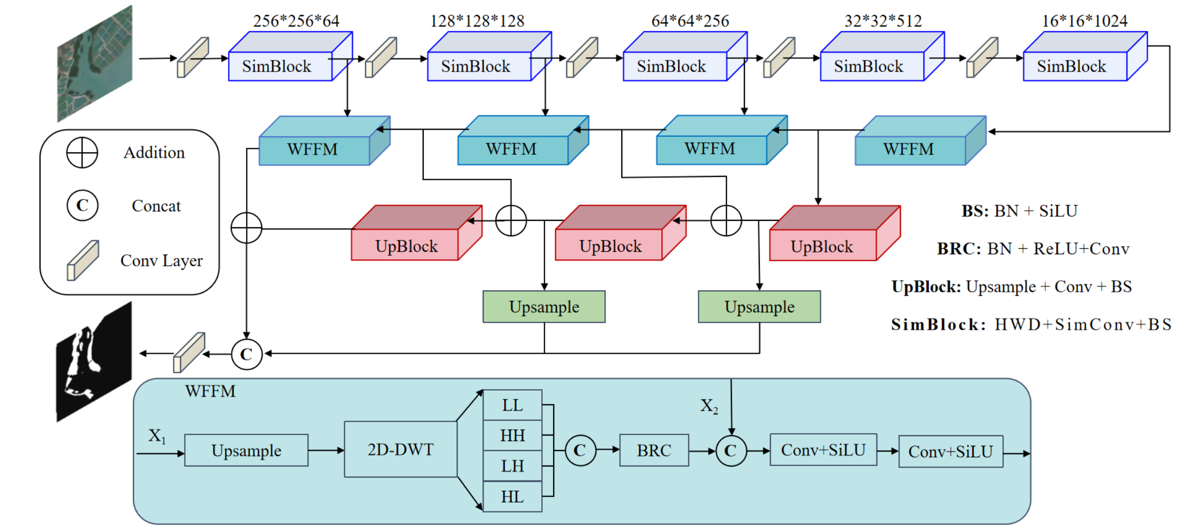

Figure 3. Overall Architecture of SIMNet.

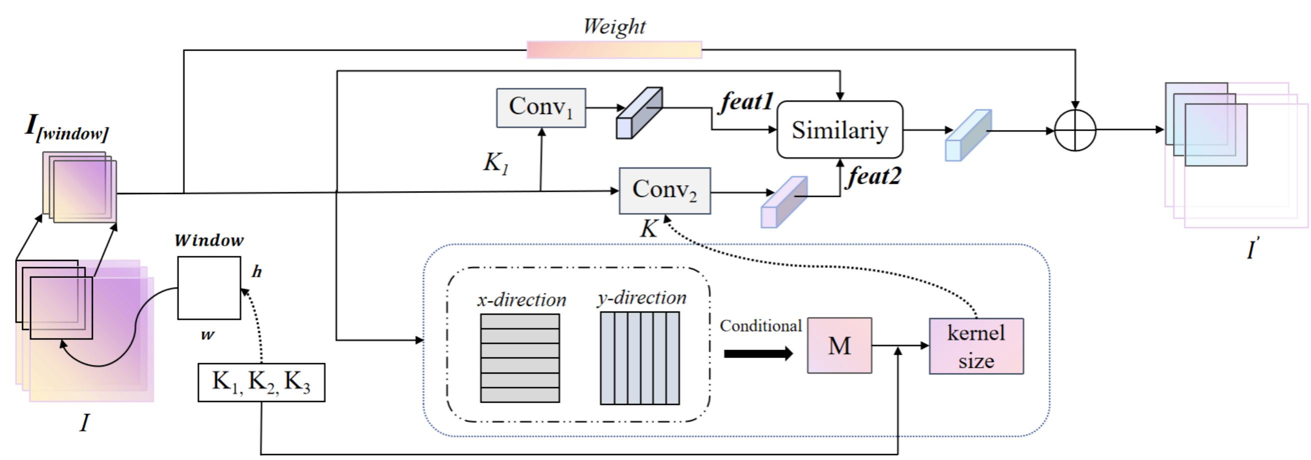

Figure 4. Similar convolution architecture

scale spatial context. In the decoder, deep features X_1 are upsampled using WFFM and decomposed into sub-band representations (LH, HH, LH, HL) via a 2D discrete wavelet transform. These are concatenated with shallow features X_2, dimensionally aligned through the BRC module, thereby facilitating efficient interaction between high-level semantics and fine-grained spatial details.

3.2. Similarity Convolution

Standard convolution, defined in Eq. 1, applies a fixed-size kernel $k \in \mathbb{R}^{m\times n}$ over the input feature map $I \in \mathbb{R}^{B\times C\times H\times W}$ via a sliding window. Despite its simplicity, this design faces two key limitations: (1) the receptive field is predetermined and cannot adapt to scale variations in the input, and (2) the local sampling mechanism processes features independently, neglecting their intrinsic correlations. Consequently, conventional convolutions often fail to capture discriminative multi-scale context in complex scenes, thereby constraining the representational capacity of learned features.

$$O(i,j) = \sum_m \sum_n I(i-m, j-n) \cdot k(m,n) + b \tag{1}$$

To overcome these limitations, we propose SimConv, an adaptive convolution operator that dynamically selects kernel sizes. As illustrated in Figure 4, a set of candidate kernels (K_1, K_2, K_3) is predefined according to the target scale distribution. SimConv then determines the local window size $h \times w$, defined as the least common multiple (LCM) of the candidate kernel side lengths. This design is motivated by two key considerations: (1) ensuring that the receptive field of each candidate kernel is fully contained within the window, thereby preserving complete multi-scale feature information; and (2) selecting the minimal window size compatible with all kernels, which avoids redundant computation and balances accuracy with efficiency. The resulting window is subsequently used to extract the local feature representation:

$$h, w = LCM(K_1, K_2, K_3) \tag{2}$$

$$I_{[window]} = I_{[h,w]} \tag{3}$$

$I_[window] \in R^{B\times C\times h\times c}$ denotes the feature information within the local window of feature I. To account for the relationship between local and global context, the local-global feature weight $Window$ must be obtained:

$$Weight = Conv_{1\times 1}(Reshape(I_{[window]})) \tag{4}$$

Next, SimConv adaptively selects the convolution kernel size based on the background information within each local window. Specifically, we compute the correlation R between feature responses along the two spatial dimensions (x and y) of the window to guide kernel selection. The correlation in the x-direction is first calculated as:

$$R_{I_{[winodw]_x}} = I_{[window]}(x+1, y) - I_{[window]}(x, y) \tag{5}$$

For connections in the y-direction:

$$R_{I_{[winodw]_y}} = I_{[window]}(x, y+1) - I_{[window]}(x, y) \tag{6}$$

To evaluate feature correlations in the x and y directions, we introduce threshold R_θ (0.1): values below R_θ indicate significant correlation (assigned 1), otherwise 0, generating a Boolean matrix M reflecting local feature correlations. Based on the proportion of 1s in M and comparison against the threshold M_θ (0.4), the convolutional kernel is adaptively selected: when the proportion exceeds M_θ, a large kernel captures broader spatial dependencies; otherwise, a small kernel is used. This strategy enables SimConv to dynamically adjust receptive fields based on local context, θ denotes the threshold. Each threshold has theoretical support: R_θ references statistical correlation-based feature selection methods (e.g., Pearson's coefficient) [27], while M_θ is grounded in information theory (incorporating concepts like entropy and mutual information for feature association evaluation) [9].

$$K = \begin{cases} K_2, \frac{M[1]}{M[1]+M[0]} > M_\theta \\ K_3, \frac{M[1]}{M[1]+M[0]} \leq M_\theta, \end{cases} \tag{7}$$

Given the local window context, the adaptive kernel K is selected to extract feature $feat2$. To evaluate its quality and completeness—particularly in regions with missing features or weak signals—we introduce an additional correlation mechanism with the original input. A reference feature $feat1$ is first obtained using a baseline kernel K_1. The similarity between $feat1$ and $feat2$ is then measured via cosine similarity, which captures directional consistency by computing the angular distance between feature vectors. This similarity serves as a criterion for feature selection, allowing the model to prioritize more representative and semantically consistent features.

$$Similarity = \frac{\sum_i x_i \cdot y_i}{\sqrt{\sum_i x_i^2} \times \sqrt{\sum_i y_i^2}} \tag{8}$$

Then, by substituting the features to be evaluated, we obtain:

$$E = \frac{\sum_{i=1}^n feat_i \cdot I[window]_i}{\sqrt{\sum_{i=1}^n feat_i^2 \times \sqrt{\sum_{i=1}^n I[window]_i}}} \tag{9}$$

Subsequently, an evaluation threshold is applied to assess the consistency between the features produced by different convolution kernels. Specifically, we compute two evaluation values: E_1, derived from the standard convolution kernel K_1 (with output feature denoted as $feat1$), and E_2, derived from the adaptively selected kernel K (with output feature denoted as $feat2$). The absolute difference $D = |E_1 - E_2|$ quantifies the consistency between $feat1$ and $feat2$: a smaller D indicates higher consistency, while a larger D implies lower consistency.

This difference D is then compared against a predefined threshold E_θ (set to 0.15 [25, 32]). To incorporate this consistency information into feature selection, we define the following rule for determining the output feature within a $I'_{[window]}$:

$$I'[window] = Weight + \begin{cases} feat1, if D > E_\theta \\ feat2, if D < E_\theta \\ \frac{feat1+feat2}{2}, other \end{cases} \tag{10}$$

If the difference is less than the threshold, the corresponding convolutional feature is considered valid and retained. Only the features that satisfy this criterion are selected for the final output, ensuring reliability and robustness in the extracted representation. The choice of thresholds is further examined in detail through ablation experiments.

3.3. Wavelet feature fusion module

To more effectively exploit both deep and shallow feature information, we propose a novel feature fusion module named Wavelet feature fusion module (WFFM), as illustrated in Figure 3. Let X_1 denote the deep-layer features and X_2 the shallow-layer features. First, X_1 is upsampled by a factor of 4 to match the spatial resolution of

X_2. The upsampled deep features are then processed via a wavelet transform to decompose them into high-frequency and low-frequency components. These components are subsequently integrated to form a refined representation, denoted as:

$$LL, HH, LH, HL = \text{2D-DWT}(Upsample_{\times 4}(X_1)) \tag{11}$$

$$X_1^{'} = Concat(LL, HH, LH, HL) \tag{12}$$

Traditional upsampling methods such as bilinear or bicubic interpolation often introduce jagged artifacts, which can degrade the visual quality and impair the effectiveness of feature fusion. To mitigate this issue, we employ a wavelet transform, which leverages its inherent multi-resolution and directional sensitivity to produce smoother pixel transitions and suppress visual distortions. This results in more natural and accurate feature representations. Following the wavelet decomposition, a standard convolution operation is applied to restore the channel dimensionality, enabling effective integration with the shallow-layer features X_2. The proposed wavelet transform feature fusion module thus facilitates the efficient utilization of both deep and shallow feature information, enhances semantic feature representation, and ultimately improves the model's learning performance.

4. Experiments and Results

4.1. Datasets

To evaluate the proposed method in forestry mapping, we constructed a benchmark using three datasets with varying spatial resolutions. The first is the GaoFen-2 **mangrove dataset**[50], a 2 class set with 1 m resolution (512×512 pixels), characterized by patchy mangrove distributions against complex coastal wetlands. The second is **LoveDA** [39], containing 5,987 high-resolution images at 0.3 m GSD (originally 1024×1024, resized to 512×512), spanning 7 land-cover categories. The third is **WHDLD** [35, 36], a multi-label dataset of 4,940 images with 2 m resolution (256×256 pixels) annotated with 6 land-cover types.

4.2. Experimental Setup

All models were implemented in PyTorch and trained on a single NVIDIA RTX 4090 GPU. To accelerate convergence and mitigate class imbalance, we used the AdamW optimizer with a hybrid loss combining Focal Loss and Dice Loss. Training was conducted from scratch without pre-trained weights. The initial learning rate was set to 5e-5 and scheduled via cosine annealing. Training configurations were dataset-specific: 50 epochs with batch size 4 for the Mangrove dataset, and 100 epochs for LoveDA (batch size 4) and WHDLD (batch size 8). To enhance generalization, standard augmentations (random scaling, horizontal flip, and translation) were applied during training, while validation used unaugmented data for fairness. Performance was evaluated using Overall Accuracy (OA), Mean Intersection over Union (mIoU), and F1-score, computed from cumulative confusion matrices for consistency and statistical reliability.

4.3. Result

This section benchmarks the proposed model against mainstream baselines. For the Mangrove and LoveDA datasets, we compare three representative architectures: (1) CNN-based, (2) Vision Transformer (ViT), and (3) state-space (Mamba). For the densely annotated WHDLD dataset, we focus on classical CNN segmentation networks, further incorporating advanced convolutional variants (DyConv, ODConv, LDConv) and feature fusion strategies (attention-gated skip connections, multi-scale aggregation) for cross-validation. Across all three datasets with distinct characteristics, the proposed model consistently achieves superior performance.

Table 1. Quantitative analysis of the mangrove dataset.

Method	BG	Mangrove	F1	OA	mIoU
ABCNet [22]	95.82	72.94	91.16	96.24	84.38
CCNet [14]	96.42	77.42	92.73	96.81	86.92
CGNet [44]	96.84	79.90	93.61	97.19	88.37
Deeplabv3+ [5]	96.05	74.98	91.85	96.47	85.51
DenseASPP [48]	96.86	80.13	93.69	97.21	88.50
DSAT-Net [51]	96.28	75.81	92.21	96.67	86.05
ENet [28]	96.22	76.43	92.36	96.63	86.32
MANet [21]	96.48	77.52	92.78	96.86	87.00
PSPNet [54]	96.16	75.75	92.12	96.57	85.95
SDSC-UNet [52]	96.55	78.11	92.98	96.93	87.33
SegNet [2]	95.45	72.17	90.76	95.93	83.81
CMTFNet [43]	96.68	78.81	93.23	97.04	87.74
DBBANet [20]	96.18	76.00	92.21	96.59	86.09
CARENet [31]	96.78	79.58	93.50	97.14	88.18
SFFNet [49]	96.29	76.69	92.46	96.70	86.49
SIMNet	**97.24**	**82.36**	**94.47**	**97.56**	**89.80**

Mangrove Dataset: As shown in Table 1, the proposed method exhibits notable performance advantages in the task of fine-grained mangrove mapping. Specifically, it achieves an Intersection over Union (IoU) of 82.36% for the mangrove class, marking an absolute improvement of 2.23% over the next best-performing model. Furthermore, in terms of the mIoU metric, the proposed approach yields a 1.3% increase compared to the baseline, underscoring its effectiveness in handling complex coastal wetland scenes with discrete vegetation patterns.

LoveDA Dataset: As shown in Table 2, SIMNet demonstrates a significant performance advantage in forest extraction tasks. Notably, it is the only model whose forest extraction accuracy exceeds the 75% threshold, underscoring its

Table 2. Quantitative analysis of the LoveDA dataset.

Method	BG	Building	Road	Water	Barren	Forest	Agriculture	F1	OA	mIoU
ABCNet [22]	**59.49**	56.20	54.98	70.67	39.00	74.07	62.88	74.33	77.92	59.61
CCNet [14]	57.35	62.30	61.55	73.31	46.50	74.35	63.66	76.79	78.21	62.72
CARENet [31]	56.15	62.85	63.39	72.68	44.69	74.13	64.36	76.66	77.98	62.61
DFANet [18]	55.34	57.08	54.16	69.50	30.21	72.58	58.99	71.88	75.63	56.84
DMNet [10]	59.20	62.10	63.53	73.35	42.90	74.74	64.50	76.89	**78.75**	62.90
DSAT-Net [51]	56.15	59.83	58.07	69.44	37.94	73.21	58.70	73.77	76.19	59.05
ENet [28]	57.37	61.71	62.01	72.41	44.86	73.66	62.88	76.27	77.76	62.13
MANet [21]	57.40	61.39	61.74	72.60	44.29	73.87	63.63	76.29	77.97	62.13
SDSCU-Net [52]	57.75	62.52	59.71	71.23	39.62	73.76	61.68	75.22	77.47	60.94
SegNet [2]	55.70	60.97	59.57	72.59	42.28	73.45	62.60	75.39	77.17	61.02
DBBANet [20]	55.07	61.92	61.32	**74.47**	46.08	73.79	64.27	76.63	77.77	62.42
SFFNet [49]	55.60	61.51	**63.66**	72.16	41.13	74.36	62.54	75.86	77.47	61.57
SIMNet	58.00	**63.72**	63.33	73.16	**46.57**	**75.04**	**65.03**	**77.50**	78.71	**63.55**

Table 3. Quantitative analysis of the WHDLD dataset.

Method	Building	Road	Pavement	Vegetation	Bare soil	Water	F1	OA	mIoU
ABCNet [22]	54.34	52.84	37.16	78.58	35.41	92.24	72.17	82.14	58.43
CCNet [14]	51.66	57.25	38.99	77.71	40.35	91.58	72.99	81.71	59.59
CGNet [44]	51.66	59.09	40.56	78.80	41.84	92.58	74.00	82.41	60.76
Deeplabv3+ [5]	51.49	56.97	40.96	78.35	38.91	92.50	73.37	82.03	59.86
DenseASPP [48]	53.30	60.26	41.27	79.11	40.94	92.80	74.36	82.76	61.30
DFANet [18]	46.19	46.84	33.41	74.53	32.26	88.22	67.82	78.71	53.58
ENet [10]	47.46	56.19	39.31	77.97	40.37	92.32	72.60	81.36	58.94
GCNet [4]	51.97	58.88	40.89	78.45	40.86	92.16	73.85	82.26	60.54
ISANet [13]	52.70	59.14	40.99	78.62	40.73	92.36	73.98	82.36	60.76
MANet [21]	53.84	58.57	41.57	79.19	40.66	92.80	74.23	82.75	61.11
SegNet [2]	45.00	47.35	37.22	74.92	34.91	90.73	69.08	79.08	55.02
UPerNet [45]	54.10	57.52	40.69	78.93	40.38	93.05	73.90	82.72	60.78
SIMNet	**56.57**	**60.47**	**43.46**	**80.71**	**41.96**	**94.02**	**75.72**	**83.93**	**62.87**

effectiveness in capturing forest-related features. Furthermore, SIMNet achieves the highest mIoU among all compared methods, highlighting its strong overall segmentation capability in complex land cover scenarios.

WHDLD Dataset: As shown in Table 3, SIMNet continued to exhibit its technical superiority on the WHDLD dataset. Among all CNN-based methods, SIMNet achieved the highest classification accuracy for forest land categories at 80.71%, representing an absolute improvement of 1.52% over the next best approach. Additionally, it maintained a leading margin of 1.57% in the mIoU metric. This quantitative advantage directly confirms the effectiveness of the proposed method in interpreting complex land features and underscores its strong potential for practical applications in forest resource mapping.

Comparison of Different Convolution Methods: Table 4 provides a detailed comparison between the proposed SimConv operator and several state-of-the-art convolutional techniques, including LDConv, ODConv, and DyConv. In the context of the challenging forest mapping task, SimConv demonstrates consistent superiority across all evaluation metrics and semantic categories. Notably, this performance gain is achieved with only a marginal increase in computational cost—specifically, an additional 0.0016 GFLOPs and 0.0004 million parameters. In particularly difficult-to-segment classes such as vegetation and water bodies, SimConv attains IoU scores of 80.71% and 94.02%, respectively, representing a substantial improvement over existing methods. These results highlight SimConv's enhanced capability to adaptively capture contextual and structural information through dynamic kernel size selection. This adaptive behavior is particularly beneficial in remote sensing scenarios like forestry mapping, where high spatial variability and complex terrain features necessitate

Table 4. Quantitative analysis of the convolution method. FLOPs and Params are a single convolution operation.

Method	Building	Road	Pavement	Vegetation	Bare soil	Water	F1	OA	mIoU	FLOPs(G)	Params(M)
SimConv	**56.57**	**60.47**	43.46	**80.71**	**41.96**	**94.02**	**75.72**	**83.93**	**62.87**	0.0456	0.0011
LDConv [53]	44.60	32.01	31.45	71.93	27.60	82.91	62.80	75.01	48.42	0.0185	0.0001
ODConv [17]	54.40	59.97	**43.76**	80.01	41.08	93.75	75.31	83.17	62.18	0.0440	0.0007
DyConv [6]	52.48	57.93	41.48	78.37	37.75	90.48	73.35	82.04	59.75	0.0440	0.0004

fine-grained feature extraction.

Comparison of Mainstream feature fusion methods: Table 5 presents a comparison of mainstream feature fusion methods for semantic segmentation of remote sensing images. Both category-specific segmentation accuracy and overall performance metrics highlight the advantages of WFFM. In particular, vegetation category segmentation accuracy improves by at least 1.38% compared with other fusion methods, while the mIoU, reflecting overall performance, surpasses competing approaches by at least 2.04%. These results demonstrate that WFFM provides a significant improvement over existing strategies by more effectively integrating deep semantic features with shallow detail features, thereby enhancing the model's ability to capture both semantic information and spatial details in remote sensing images.

5. Ablation Studies

This section will systematically analyze the mechanism of the SIMNet core module and deeply analyze the impact of different similarity threshold settings in the SimConv convolution kernel on model performance.

5.1. SIMNet Feature Extraction and Feature Fusion

To evaluate the individual contributions of SIMNet's core components, we conducted ablation experiments on three benchmark datasets, with results summarized in Table 6.

Feature Extraction Performance Validation: The removal of the proposed feature extraction module results in a significant decline in performance across all datasets. This confirms the module's effectiveness in capturing salient object features and fine-grained details, while also leveraging contextual background information. Moreover, scale-specific evaluations demonstrate the module's robust multi-scale feature extraction capability, ensuring consistent performance across varying resolutions and scene complexities.

Feature Fusion Performance Validation: Similarly, removing the proposed feature fusion mechanism also leads to a notable performance drop. This validates the efficacy of our fusion strategy in integrating deep semantic and shallow spatial information, enhancing the model's ability to exploit multi-scale features. The improved utilization of hierarchical feature representations significantly boosts the accuracy and robustness of object recognition and classification in complex remote sensing scenes.

5.2. Choosing the threshold

To validate the rationale behind the threshold settings of the SimConv module, we conducted control experiments on the WHDLD dataset, focusing on the key thresholds introduced in Section 3: (1) the feature correlation threshold R_θ, which evaluates feature correlations in the x- and y-directions, and (2) the convolution kernel selection threshold M_θ, which determines kernel size. These thresholds were systematically varied, and their impact on kernel size selection was assessed. The results, presented in Table 7, show that when image resolution is reduced, the combination of R_θ=0.1 and M_θ=0.4 yields the slowest decline in mIoU while maintaining superior performance. This demonstrates that the chosen parameter settings enable precise, adaptive adjustment of kernel size, thereby allowing the model to achieve optimal performance.

To further assess the influence of thresholds on feature quality evaluation, we performed ablation experiments as reported in Table 8. The results indicate that a threshold of 0.15 provides the most accurate reflection of feature validity during similarity-based evaluation, enabling SimConv to make more precise and context-aware feature selection decisions. Together, these experiments validate the empirical design choices for threshold selection in SimConv and confirm their critical role in enhancing the model's feature extraction and representation capabilities.

5.3. Convolution Kernel List Combination

To validate the rationality of the threshold settings in the SimConv module, we conducted control experiments on the WHDLD dataset, focusing on the third layer. In addition, to verify the effectiveness of the convolution kernel list proposed in Section 3.2, we performed ablation experiments using multiple kernel combinations on the WHDLD dataset. The results are presented in Table 9. As shown, the hybrid convolution kernel strategy—constructing the encoder with kernels of sizes 1, 3, 5, 7, and 9—achieved optimal performance across core metrics such as classification accuracy and overall consistency. At the same time, it maintained a lightweight design, with a parameter count of only 73.78,

Table 5. Quantitative analysis of mainstream feature fusion methods.

Method	Building	Road	Pavement	Vegetation	Bare soil	Water	F1	OA	mIoU
Multi-scale feature addition [16]	53.34	56.19	41.17	79.33	**42.06**	92.88	74.08	82.75	60.83
Attention-gated skip connections [27]	44.47	47.86	37.46	77.22	34.31	91.19	69.61	79.80	55.42
WFFM	**56.57**	**60.47**	**43.46**	**80.71**	41.96	**94.02**	**75.72**	**83.93**	**62.87**

Table 6. Ablation study of each component in the proposed SIM-Net model.

Dataset	**Size**	**SimConv**	**WFFM**	**mIoU**	**F1**
Mangrove	512	✓	✓	89.80	94.47
	512	✓	×	89.45	94.25
	512	×	✓	88.55	93.72
	256	✓	✓	88.39	93.62
	256	✓	×	87.98	93.38
	256	×	✓	86.05	92.18
LoveDA	512	✓	✓	63.55	77.50
	512	✓	×	61.79	75.94
	512	×	✓	62.27	77.80
	256	✓	✓	63.86	77.68
	256	✓	×	62.15	76.34
	256	×	✓	62.86	76.99
WHDLD	256	✓	✓	62.87	75.72
	256	✓	×	61.37	74.57
	256	×	✓	62.10	75.07
	128	✓	✓	61.41	74.64
	128	✓	×	60.17	73.64
	128	×	✓	59.07	72.50

Table 7. Ablation study of each component in the proposed SIM-Net model.

Input size	R_θ	M_θ	**F1**	**OA**	**mIoU**
256	0.05	0.30	75.25	83.68	62.42
	0.10	0.40	75.72	83.93	62.87
	0.20	0.50	75.41	83.69	62.49
128	0.05	0.30	74.25	82.42	60.75
	0.10	0.40	74.64	82.87	61.41
	0.20	0.50	74.20	82.59	60.97

significantly lower than all other kernel combinations. By contrast, fixed-size kernel schemes (e.g., 1, 5, 9 or 1, 7, 9) occasionally achieved performance close to that of the hybrid approach, but required substantially more parameters (e.g., the combination of 1, 5, 9 reached 165.5), far exceeding the hybrid configuration. These findings clearly demonstrate that hybrid convolution kernels more effectively capture multi-scale features from remote sensing images while maintaining parameter efficiency, making them particularly well suited for feature mapping tasks.

Table 8. Quantitative analysis of threshold ablation experiments

E_θ	F1	OA	mIoU
0.10	75.64	83.83	62.73
0.15	75.72	83.93	62.87
0.20	75.56	83.73	62.62

Table 9. Quantitative Analysis of Convolution Kernel Combinations.

K_1, K_2, K_3	F1	OA	mIoU	Params(M)
1,3,9	75.46	83.65	62.45	165.56
1,7,9	75.37	83.63	62.44	199.08
3,3,5	75.14	83.36	62.01	81.74
3,5,7	74.71	83.32	61.40	137.61
Hybrid	75.72	83.93	62.87	73.78

6. Conclusion

In summary, we propose the novel Similarity Convolution (SimConv) method, which adaptively selects convolution sizes by focusing on intrinsic feature relationships. This dynamic approach enables more efficient and accurate feature extraction, overcoming the fixed receptive field limitations of traditional convolutions. Built upon SimConv, the SIM-Net model demonstrates outstanding performance across multiple forestry mapping datasets, achieving top accuracy and evaluation scores. These results validate SimConv's effectiveness and present SIMNet as a powerful, competitive solution for forestry survey applications.

7. Acknowledgement

This work was supported by the Projects of the National Natural Science Foundation of China (NSFC) (Grant No. 42571358, 42293270, 4243000531).

References

[1] Syed Ashraful Alam, Sonja Kivinen, Heini Kujala, Topi Tanhuanpää, and Martin Forsius. Integrating carbon sequestration and biodiversity impacts in forested ecosystems: Concepts, cases, and policies. *Ambio*, 52(11):1687–1696, 2023. 1

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 5, 6

[3] Mattia Balestra, Suzanne Marselis, Temuulen Tsagaan Sankey, Carlos Cabo, Xinlian Liang, Martin Mokroš, Xi Peng, Arunima Singh, Krzysztof Stereńczak, Cedric Vega, et al. Lidar data fusion to improve forest attribute estimates: A review. *Current Forestry Reports*, 10(4):281–297, 2024. 1

[4] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 6

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2, 5, 6

[6] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11030–11039, 2020. 2, 7

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 2

[8] Mingzhe Feng, Xin Sun, Junyu Dong, and Haoran Zhao. Gaussian dynamic convolution for semantic segmentation in remote sensing images. *Remote Sensing*, 14(22):5736, 2022. 2

[9] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003. 4

[10] Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multiscale filters for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3562–3572, 2019. 6

[11] Yan He, Bing Tu, Bo Liu, Yunyun Chen, Jun Li, and Antonio Plaza. Hybrid multi-scale spatial-spectral transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1

[12] Benyamin Hosseiny, Mahdieh Zaboli, and Saeid Homayouni. Forest change mapping using multi-source satellite sar, optical, and lidar remote sensing data. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10:163–168, 2024. 1

[13] Lang Huang, Yuhui Yuan, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Interlaced sparse self-attention for semantic segmentation. *arXiv preprint arXiv:1907.12273*, 2019. 6

[14] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019. 5, 6

[15] Teja Kattenborn, Jens Leitloff, Felix Schiefer, and Stefan Hinz. Review on convolutional neural networks (cnn) in vegetation remote sensing. *ISPRS journal of photogrammetry and remote sensing*, 173:24–49, 2021. 2

[16] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. 8

[17] Chao Li, Aojun Zhou, and Anbang Yao. Omni-dimensional dynamic convolution. *arXiv preprint arXiv:2209.07947*, 2022. 2, 7

[18] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun. Dfanet: Deep feature aggregation for real-time semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9522–9531, 2019. 6

[19] Jiaxin Li, Danfeng Hong, Lianru Gao, Jing Yao, Ke Zheng, Bing Zhang, and Jocelyn Chanussot. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, 112:102926, 2022. 1

[20] Jiepan Li, Yipan Wei, Tiangao Wei, and Wei He. A comprehensive deep-learning framework for fine-grained farmland mapping from high-resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025. 5, 6

[21] Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Jianlin Su, Libo Wang, and Peter M Atkinson. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021. 5, 6

[22] Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Libo Wang, and Peter M Atkinson. Abcnet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. *ISPRS journal of photogrammetry and remote sensing*, 181:84–98, 2021. 5, 6

[23] Denghui Liu, Lin Zhong, Haiyang Wu, Songyang Li, and Yida Li. Remote sensing image super-resolution reconstruction by fusing multi-scale receptive fields and hybrid transformer. *Scientific Reports*, 15(1):2140, 2025. 2

[24] Jiahang Liu, Donghao Yang, and Fei Hu. Multiscale object detection in remote sensing images combined with multi-receptive-field features and relation-connected attention. *Remote Sensing*, 14(2):427, 2022. 2

[25] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999. 4

[26] Keiller Nogueira, Mauro Dalla Mura, Jocelyn Chanussot, William Robson Schwartz, and Jefersson Alex Dos Santos. Dynamic multicontext segmentation of remote sensing im-

ages based on convolutional networks. *IEEE Transactions on Geoscience and Remote Sensing*, 57(10):7503–7520, 2019. 2

[27] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 4, 8

[28] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016. 5, 6

[29] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 367–376, 2021. 1

[30] Ruiliang Pu. Mapping tree species using advanced remote sensing technologies: a state-of-the-art review and perspective. *Journal of remote sensing*, 2021. 1

[31] Zhilin Qu, Mingzhe Li, and Zehua Chen. Carenet: Satellite imagery road extraction via context aware and road enhancement. *IEEE Geoscience and Remote Sensing Letters*, 2025. 5, 6

[32] Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer, 2010. 4

[33] Francesco Maria Sabatini, Rafael Barreto de Andrade, Yoan Paillet, Péter Ódor, Christophe Bouget, Thomas Campagnaro, Frédéric Gosselin, Philippe Janssen, Walter Mattioli, Juri Nascimbene, et al. Trade-offs between carbon stocks and biodiversity in european temperate forests. *Global Change Biology*, 25(2):536–548, 2019. 1

[34] Serajis Salekin, Yvette L Dickinson, Mark Bloomberg, and Dean F Meason. Carbon sequestration potential of plantation forests in new zealand-no single tree species is universally best. *Carbon Balance and Management*, 19(1):11, 2024. 1

[35] Zhenfeng Shao, Ke Yang, and Weixun Zhou. Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset. *Remote Sensing*, 10(6):964, 2018. 5

[36] Zhenfeng Shao, Weixun Zhou, Xueqing Deng, Maoding Zhang, and Qimin Cheng. Multilabel remote sensing image retrieval based on fully convolutional network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:318–328, 2020. 5

[37] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019. 2

[38] Luoma Wan, Hongsheng Zhang, Guanghui Lin, and Hui Lin. A small-patched convolutional neural network for mangrove mapping at species level using high-resolution remote-sensing image. *Annals of GIS*, 25(1):45–55, 2019. 1

[39] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 5

[40] Shikuan Wang, Xingwen Cao, Mengquan Wu, Changbo Yi, Zheng Zhang, Hang Fei, Hongwei Zheng, Haoran Jiang, Yanchun Jiang, Xianfeng Zhao, et al. Detection of pine wilt disease using drone remote sensing imagery and improved yolov8 algorithm: A case study in weihai, china. *Forests (19994907)*, 14(10), 2023. 1

[41] Shikuan Wang, Mengquan Wu, Xinghua Wei, Xiaodong Song, Qingtong Wang, Yanchun Jiang, Jinkun Gao, Lingyi Meng, Zhipeng Chen, Qiyue Zhang, et al. An advanced multi-source data fusion method utilizing deep learning techniques for fire detection. *Engineering Applications of Artificial Intelligence*, 142:109902, 2025. 1

[42] Xuan Wang, Yue Zhang, Tao Lei, Yingbo Wang, Yujie Zhai, and Asoke K Nandi. Dynamic convolution self-attention network for land-cover classification in vhr remote-sensing images. *Remote Sensing*, 14(19):4941, 2022. 2

[43] Honglin Wu, Peng Huang, Min Zhang, Wenlong Tang, and Xinyu Yu. Cmtfnet: Cnn and multiscale transformer fusion network for remote-sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–12, 2023. 5

[44] Tianyi Wu, Sheng Tang, Rui Zhang, Juan Cao, and Yongdong Zhang. Cgnet: A light-weight context guided network for semantic segmentation. *IEEE Transactions on Image Processing*, 30:1169–1179, 2020. 5, 6

[45] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 6

[46] Guoping Xu, Wentao Liao, Xuan Zhang, Chang Li, Xinwei He, and Xinglong Wu. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern recognition*, 143:109819, 2023. 2

[47] Feng Yang and Zhenzhong Zeng. Refined fine-scale mapping of tree cover using time series of planet-nicfi and sentinel-1 imagery for southeast asia (2016–2021). *Earth System Science Data*, 15(9):4011–4021, 2023. 1

[48] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3684–3692, 2018. 2, 5, 6

[49] Yunsong Yang, Genji Yuan, and Jinjiang Li. Sffnet: A wavelet-based spatial and frequency domain fusion network for remote sensing segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024. 5, 6

[50] Liu Yequ, Zhang Li, Guo Kangli, Dang Er-sha, and Tang Shilin. A dataset of mangrove vector in the Guangdong province during 2015–2020, 2021. 5

[51] Renhe Zhang, Zhechun Wan, Qian Zhang, and Guixu Zhang. Dsat-net: Dual spatial attention transformer for building extraction from aerial images. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. 5, 6

[52] Renhe Zhang, Qian Zhang, and Guixu Zhang. Sdsc-unet: Dual skip connection vit-based u-shaped model for building extraction. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. 5, 6

[53] Xin Zhang, Yingze Song, Tingting Song, Degang Yang, Yichen Ye, Jie Zhou, and Liming Zhang. Ldconv: Linear

deformable convolution for improving convolutional neural networks. *Image and Vision Computing*, 149:105190, 2024. 2, 7

[54] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 5

[55] Yunji Zhao, Zhihao Zhang, Wenming Bao, Xiaozhuo Xu, and Zhifang Gao. Hyperspectral image classification based on channel perception mechanism and hybrid deformable convolution network. *Earth Science Informatics*, 17(3): 1889–1906, 2024. 2

[56] Bai Zhu, Liang Zhou, Simiao Pu, Jianwei Fan, and Yuanxin Ye. Advances and challenges in multimodal remote sensing image registration. *IEEE Journal on Miniaturization for Air and Space Systems*, 4(2):165–174, 2023. 1