The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-G-2025
ISPRS Geospatial Week 2025 “Photogrammetry & Remote Sensing for a Better Tomorrow…”, 6–11 April 2025, Dubai, UAE

Remote sensing semantic segmentation based on multimodal feature alignment and fusion

Boshen Chang[1], Timo Balz[1]

[1] The State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China,430079

Abstract

The accurate semantic segmentation of remote sensing data is of paramount importance to the success of geoscience research and applications. In comparison to traditional single-modal segmentation techniques, models based on multi-modal fusion have demonstrated superior performance and have been the subject of considerable attention in recent years. However, the majority of these models employ convolutional neural networks (CNNs) or visual transformers (ViTs) for fusion operations, which results in inadequate modelling and representation of local-global context. In this study, we propose a multi-layer multi-modal feature alignment and fusion scheme, designated as MFAFUNet, with the objective of providing a robust and effective multi-modal fusion backbone for semantic segmentation. The overarching algorithmic framework is analogous to that of the Unet model. First, the data in different modalities is aggregated and the image size is reduced through the use of multi-level downsampling modules based on the Haar wavelet transform. The high-frequency and low-frequency information of the features is extracted through a feature extraction module composed of a convolutional neural network (CNN) and a visual transformer (ViT). Second, through the semantic distribution alignment loss, the high-level features of different modal information are transformed into a common latent space, and their distributions are aligned to associate the complementary clues hidden in each modality. The effectiveness of the proposed method is demonstrated through experiments.

Keywords: Land Use, Haar Transform, Feature Aligning

1. Introduction

Land use classification based on remote sensing imagery plays a vital role in land resource management and planning, offering essential insights into how land is spatially utilized and how human activities drive its transformation over time(Yue et al., 2024). By identifying various land cover categories, the classification process provides a more intuitive understanding of land use patterns and purposes. In complex environments, land cover classification supports a wide range of applications, including urban development planning (Zhang et al., 2024), mapping of geospatial information (Luo et al., 2020), and initiatives in environmental protection (Fraser and Storie, 2016). Recent progress in Earth observation and sensor technologies has made it easier to access multi-source remote sensing data of the same geographical scene. Integrating the unique imaging traits of various data sources enables not only the handling of tasks beyond the scope of a single modality but also contributes to enhanced performance overall (Chen et al., 2017)(Ienco et al., 2019).

Optical and SAR imagery each offer distinct yet complementary advantages in identifying surface features. While both can discern different terrains with consistency, optical images highlight surface categories using spectral color cues, whereas SAR imagery emphasizes object structure and material properties. Hence, designing robust strategies for integrating the distinctive attributes of each data source is key to advancing multi-source land-use classification. In this study, we investigate multiple fusion schemes that aim to harness these complementary characteristics effectively.

Earlier land-use classification approaches in remote sensing primarily relied on hand-crafted feature extraction and manual labeling techniques (Fan and Lin, 2007). These traditional methods were labor-intensive, prone to subjectivity, and limited in their ability to extract deep semantic information. With the rise of deep learning, especially convolutional neural networks (CNNs) (LeCun et al., 1998), automated and more reliable classification techniques have gained prominence. Initial adaptations of deep learning for remote sensing tasks often stemmed from semantic segmentation models developed for natural imagery. For example, Zhao et al. (Zhao and Du, 2016) implemented a multi-level and multi-scale model to extract advanced spatial representations by capturing contextual cues within each pixel's receptive field. Similarly, the "From Contexts to Locality" framework (Li et al., n.d.) merges high-resolution spatial features with low-resolution semantic cues to enhance classification representation.

Relying exclusively on single-modality data limits improvements in dense prediction tasks. As sensor systems evolve, the field has entered a phase characterized by "multi-sourcing," where exploiting the synergy between different modalities is essential for accuracy improvements. A variety of strategies have been proposed to leverage this advantage. For instance, AFNet (Yang et al., 2021) incorporates spatial and channel attention to combine low- and high-level features from different sources, thereby improving classification near object boundaries. MCANet (Li et al., 2022) and CMX (Zhang et al., 2023) utilize a cross-attention (CA) mechanism that allows features extracted from different modalities at the same layer to interact and share complementary context, enabling better fusion performance and enhancing the capacity for global information exchange.

Nevertheless, many of these approaches focus solely on either spatial or channel-wise fusion and project the fused features into a one-dimensional semantic embedding space, which may be insufficient due to the complexity and redundancy of multi-modal features. Additionally, traditional CNN and Transformer extractors may struggle to establish fine-grained semantic con-

nections between pixels. To overcome these limitations, and drawing inspiration from the Convolutional Block Attention Module (Woo et al., 2018) and the work of Chen et al. (Chen et al., 2022), we introduce a Feature Decomposition Module. This module leverages CNNs for extracting high-frequency, localized patterns and utilizes transformers to capture broader, low-frequency global context—thereby enhancing feature representation completeness.

Our framework adopts a U-Net-like structure in which multimodal features are combined and spatial resolution is reduced through successive downsampling layers built on the Haar wavelet transform. Following this, a semantic distribution alignment loss function is employed to project high-level features from each modality into a shared latent space. This alignment facilitates the integration of complementary information across modalities and strengthens the network's semantic consistency.

2. Method

In this paper, we present a multilevel feature decomposition and reconstruction image fusion framework designed to effectively integrate complementary information from different sensors, specifically SAR and optical images. The framework consists of several key components: a feature decomposition module, a feature reconstruction module, and a cross-feature fusion block, forming a deep fusion network that follows a U-Net-like structure. An overview of the framework is shown in Fig. 1.

The input SAR and optical images are first transformed using the Haar wavelet, with their resolution reduced to half of the original size before being passed through the feature decomposition module via channel splitting. The feature decomposition module is divided into high-frequency and low-frequency feature extraction branches, with the high-frequency branch capturing texture details and the low-frequency branch focusing on global background structure information. To further optimize the fusion process, the framework adopts a multi-level processing approach. At each level, the ratio of high- and low-frequency feature extraction modules is adjusted by different channel splitting coefficients. This multi-level decomposition allows the model to gradually extract rich semantic information while preserving local details.

At each level, the cross-modal feature fusion module enables fine-grained interaction between features extracted from both source images. The extracted features from both modalities, along with the fused features and features sampled for reconstruction at the next level, are input into the feature reconstruction module. After multi-level feature decomposition and fusion, the structural information from the SAR image and the texture and color information from the optical image are gradually combined, ensuring the full utilization of image details. The final fusion result retains high-resolution details while capturing global semantic information.

We will now describe the various modules of the framework in detail, including the downsampling module, the feature decomposition module, the cross-fertilization module, and the feature reconstruction module. We will also outline the training process and loss function design for the self-supervised learning setup.

2.1 Feature Decomposition Module(FDM)

Once the input data is fed into the feature decomposition module, it undergoes downsampling using the Haar wavelet transform, reducing its resolution to half. The downsampled data is then directed into the high-frequency and low-frequency feature extraction branches through a channel splitting mechanism, ensuring the accurate extraction of features at different frequencies. The module's structure is illustrated in Fig. 2 and Fig. 3.

2.1.1 Downsampling Module According to the Nyquist-Shannon sampling theorem, frequencies higher than the Nyquist frequency (half the sampling rate) are lost during downsampling. For example, frequencies above 1/4 are aliased in a 2-fold downsampling operation (e.g., using a 1x1 convolutional layer with a stride of 2, resulting in a sampling rate of 1/2). To mitigate information loss, we utilize a first-order Haar wavelet transform layer for downsampling. This reduces the spatial resolution of the feature maps to half of the original, while preserving essential frequency information. Additionally, to capture long-range dependencies, a Transformer with spatial self-attention is used. To balance performance and computational efficiency, we adopt the LT block, which extracts low-frequency basis features (LFE) with reduced computational complexity. Instead of using a fully connected neural network, we use a 1x1 convolution to reduce computational effort. The LFE is responsible for extracting low-frequency features from the input, as shown in the following formula(1):

$$\begin{aligned}\Phi_i^O &= \left(\mathrm{ReLU}\left(\mathrm{Batch}\left(\left(\mathrm{Conv}_{1\times1}\right)\left(\mathrm{HWT}\left(\left(\mathrm{sar}\right)_{i-1}\right)\right)\right)\right)\right)\\ \Phi_i^S &= \left(\mathrm{ReLU}\left(\mathrm{Batch}\left(\left(\mathrm{Conv}_{1\times1}\right)\left(\mathrm{HWT}\left(\left(\mathrm{sar}\right)_{i-1}\right)\right)\right)\right)\right)\end{aligned} \tag{1}$$

Where $HWT()$ denotes the Haar wavelet transform, $(Conv)_{1\times1}$ denotes 1×1 convolution, $Batch$ represents batch normalization, $Relu$ represents the ReLU activation function, Φ_i^O and Φ_i^S represent the processed outputs of optical and SAR images, respectively.

2.1.2 low-frequency Feature Extraction Module In contrast to BTE, DCE extracts high-frequency detail information from the assigned channel features. Since high-frequency features play a critical role in capturing edge and texture information, which is essential for successful image fusion, the CNN architecture in DCE is designed to preserve as much detailed information as possible. The formula for this operation is presented as follows.

The INN module facilitates the preservation of input information by establishing a feedback loop between the input and output features. It serves as a lossless feature extraction module, making it ideal for this application. Therefore, we employ INN blocks with affine coupling layers, where the processing for each modality is identical. For an optical image, the transformation can be expressed as follows.

$$\Phi_i^O = \mathcal{L}\left(\Phi_i^O\right), \Phi_i^S = \mathcal{L}\left(\Phi_i^S\right) \tag{2}$$

The formula $\mathcal{L}$ denotes the low-frequency feature extractor.

2.1.3 high-frequency Feature Extraction Module In contrast to BTE, DCE extracts high-frequency detail information

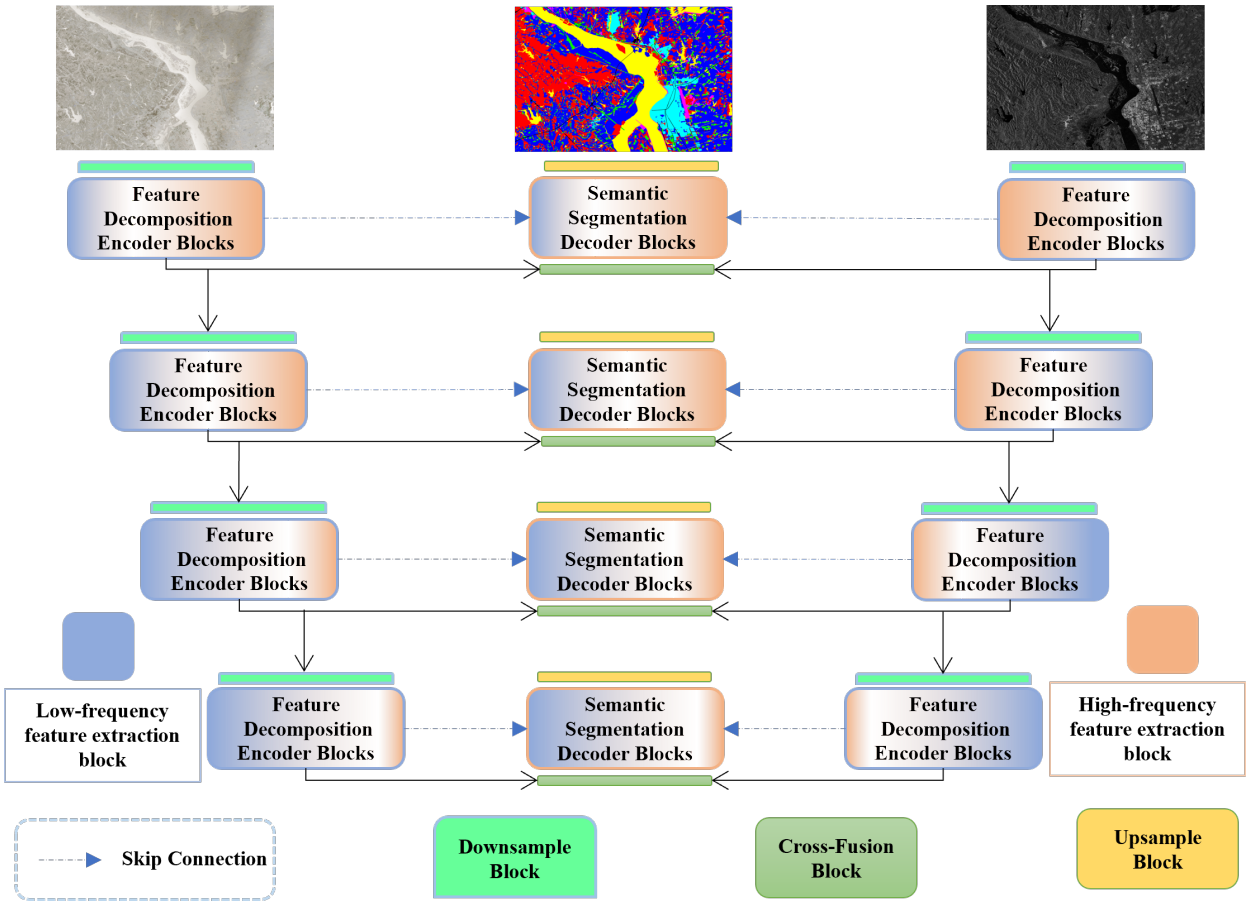

Figure 1. The overall framework of a network for SAR and optical image fusion based on multi-level feature decomposition.

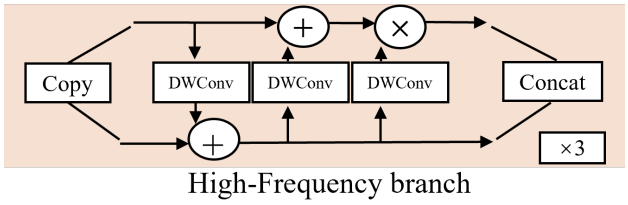

Figure 2. The architecture of the High-Frequency branch

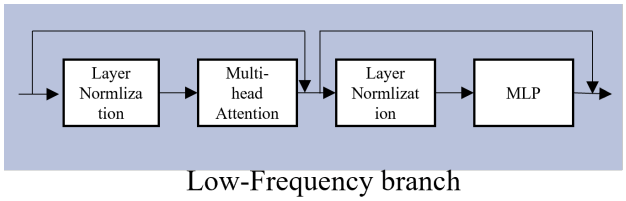

Figure 3. The architecture of the Low-Frequency branch

from the assigned channel features. Since high-frequency features play a critical role in capturing edge and texture information, which is essential for successful image fusion, the CNN architecture in DCE is designed to preserve as much detailed information as possible. The formula for this operation is presented as follows.

$$\Phi_i^O = \mathcal{H}\left(\Phi_i^O\right), \Phi_i^S = \mathcal{H}\left(\Phi_i^O\right) \tag{3}$$

Where $\mathcal{H}$ denotes the high-frequency feature extractor.

The INN module facilitates the preservation of input information by establishing a feedback loop between the input and output features. It serves as a lossless feature extraction module, making it ideal for this application. Therefore, we employ INN blocks with affine coupling layers, where the processing for each modality is identical. For an optical image, the transformation can be expressed as follows.

$$\begin{aligned}
\Phi_{i,c2}^O &= \Phi_{i,c2}^O + \mathrm{DWConv}\left(\Phi_{i,c1}^O\right) \\
\Phi_{i,c1}^O &= \Phi_{i,c1}^O \odot \exp\left(\mathrm{DWConv}\left(\Phi_{i,c2}^O\right)\right) + \mathrm{DWConv}\left(\Phi_{i,c2}^O\right) \\
\Phi_i^O &= \mathrm{Concat}\left(\Phi_{i,c1}^O, \Phi_{i,c2}^O\right)
\end{aligned} \tag{4}$$

In Eq. (4), the DWConv denotes the separable convolution, $\Phi_{i,c1}^O$ and $\Phi_{i,c2}^O$ denotes the C1 and C2 channel branches that

have been divided equally at the i^{th} level, and the $\odot$ denotes the Hadamard product.

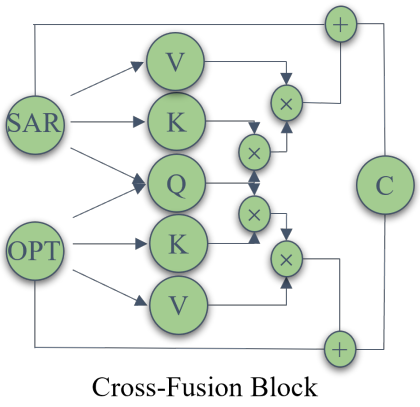

Figure 4. The architecture of the Cross-Fusion Block

2.2 Cross feature fusion Block (CFFB)

In this module, images from the two modalities are initially transformed linearly to generate their respective key and value vectors. To capture joint features from both modalities, query vectors are created by concatenating the features from both the SAR and optical images. The features are then weighted and integrated by calculating attention weights between the key and query vectors. The SAR and optical image features are reconstructed by combining their respective weighted value vectors, and the reconstructed features from both modalities are fused to produce the final output. This approach enables effective cross-modal information fusion by adaptively assigning feature weights to each modality. The block structure is shown in Fig. 4, with the corresponding formula provided below.

$$\Phi_i = Concat\left(C\left(\Phi_i^O, \Phi_i^s\right)\right) \tag{5}$$

Where $\mathcal{C}$ denotes the interactive feature fusion module, and Φ_i denotes the fusion feature information.

2.3 Loss Design

Images from different modalities often have varying visual characteristics, and these differences can obscure the underlying semantic information, making it challenging for the model to capture shared semantics between modalities. Consequently, the network might focus more on superficial appearance differences than on the common meaning shared by both images, leading to semantic inconsistency. Thus, ensuring semantic alignment across feature distributions and extracting complementary cues from each modality becomes essential for effective fusion (Li et al., 2023).

In recent work, maximum mean discrepancy (MMD) has emerged as a common metric for estimating the divergence between two distributions due to its non-parametric nature. The fundamental idea behind MMD is that two distributions can be considered equivalent if their statistical properties are the same. This is achieved by projecting features into a reproducing kernel Hilbert space (RKHS), where the distance between distributions is computed. In our framework, we leverage high-dimensional feature outputs to construct a semantic distribution alignment loss, guiding the model to emphasize semantic consistency. To simplify the implementation, we use an upper-bound approximation of the original MMD formulation, making the method more interpretable and computationally manageable.

$$Loss_{MMD} = \left\| E_p\left[\varphi\left(F^{optical}\right)\right] - E_q\left[\varphi\left(F^{SAR}\right)\right]\right\|_{\mathcal{H}} \tag{6}$$

The transformed distributions of $F^{optical}$ and F^{SAR} through $\varphi(x)$ lie in the reproducing kernel Hilbert space (RKHS) and are denoted by p and q, respectively, where E_p and E_q are their expected values. When the maximum mean discrepancy satisfies $MMD(p, q) = 0$, it indicates that the distributions p and q are statistically equivalent.

3. Results

In this section, we conduct a comparative analysis of our method and other methods such as UNet from multiple evaluation metrics. Through two sets of experiments, we respectively evaluate the per - class segmentation results of image - level fusion on the WHU - OPT - SAR dataset (see Table 1) and visualize the detailed results of this dataset (see Table 2). These experimental results show that different methods have their own advantages and disadvantages in different indicators and categories, but our method demonstrates significant advantages.

In Table 1 (Some visualization of detailed results), our method also performs well in the aAcc and mIoU indicators. The aAcc reaches 84.72, slightly higher than 84.69 of UNet(Optical + SAR); the mIoU is 36.79, also leading. Among various categories, the accuracy of our method for the ”houses” category reaches 84.9, which is significantly higher than other methods, indicating that our method has obvious advantages in identifying the house category.

In Table 2 (Per - class segmentation results of image - level fusion), in terms of the average accuracy (aAcc), our method reaches 84.61, which is higher than 74.99 of UNet(SAR), 82.5 of UNet(Optical), and 82.39 of UNet(Optical + SAR). Regarding the mean intersection over union (mIoU), our method is 36.83, also outperforming several other UNet - related methods. In terms of the segmentation accuracy of each category, for the ”bareground” category, the accuracy of our method reaches 93.9, far exceeding other methods; for the ”vegetation” category, the accuracy of our method is 76.3, which is also higher than other UNet methods. This indicates that our method has higher accuracy in identifying land categories such as bare ground and vegetation.

Overall, based on the results of the two tables, our method outperforms other comparative methods in multiple indicators. This advantage benefits from the spatial - aware circular module we adopted. It enhances the transferability of features between pixels and establishes a cross - modal receptive field, effectively correlating features from different modalities. At the same time, the supervision mechanism we designed ensures that the two modalities share a common semantic representation, explores complementary cues between modalities, creates favorable conditions for feature fusion, and prevents the network from only focusing on appearance differences, thus improving the overall segmentation performance.

Our method achieves the highest overall accuracy (OA) of 84.2 and mean intersection over union (mIoU) of 58.5, outperforming several other methods. Compared to the latest MCANet,

Table 1. Some visualization of detailed results on the YYX dataset.

Methods	aAcc	mIoU	Class					
			bareground	vegetation	trees	houses	water	roads
UNet(SAR)	84.19	34.2	85.8	65.3	23.7	65.5	29.7	43.1
UNet(Optical)	79.15	29.88	84.1	68.7	26.9	61	12	28.5
UNet(Optical + SAR)	84.69	36.73	88.8	68.9	23.5	67	48.4	42.9
Ours	84.72	36.79	89.8	81.9	22.6	84.9	10.1	39.3

Table 2. Per - class segmentation results of image - level fusion on the WHU-OPT-SAR dataset.

Methods	aAcc	mIoU	Class					
			farmland	city	village	water	forest	roads
UNet(SAR)	74.99	22.1	85.8	23.9	27.2	50.6	–	39.6
UNet(Optical)	82.5	29.96	86.8	50.6	32.4	76.1	7.8	47.9
UNet(Optical + SAR)	82.39	30.59	80.3	74.1	36	60.8	3.1	22.1
Ours	84.61	36.83	93.9	76.3	21	73.4	3.7	40.9

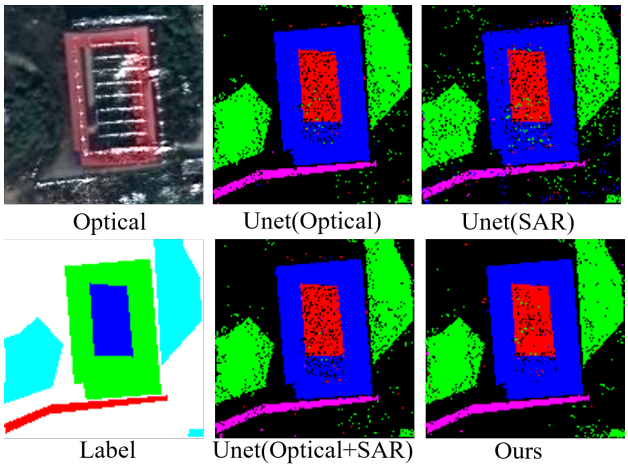

Figure 5. Per - class segmentation results of image - level fusion on the YYX dataset.

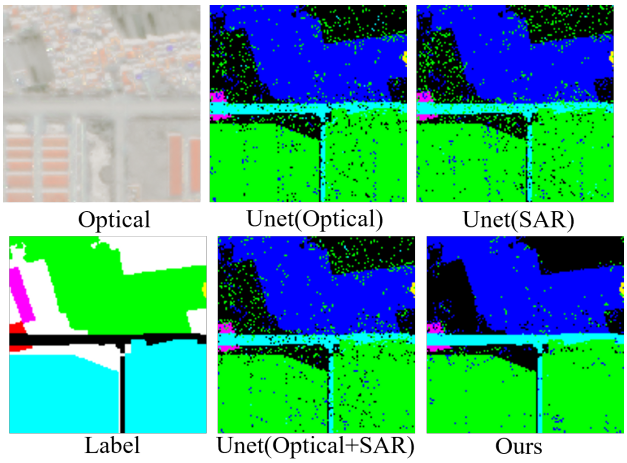

Figure 6. Some visualization of detailed results on the WHU-OPT-SAR dataset.

our approach improves OA by 2.1 and mIoU by 3.0. This improvement is due to our spatial-aware circular module, which globally enhances the transferability of features across pixels and establishes a cross-modal receptive field that correlates features from different modalities. Additionally, by ensuring that both modalities share a common semantic representation, we design a supervision mechanism to explore complementary cues between them. This supervision creates an ideal condition for feature fusion and prevents the network from focusing solely on appearance disparities.

Figure 6 displays several visualized results. Compared with U-Net trained solely on optical images, the version using combined optical and SAR inputs (U-Net-RGBNS) performs worse in distinguishing features such as water areas and farmland. This performance drop is attributed to the naïve fusion of multi-modal inputs—stacking them directly can lead the model to overly attend to visual differences between modalities, causing interference in feature learning. In contrast, more advanced fusion frameworks that explicitly handle modality differences are able to reduce this interference and improve classification accuracy.

4. Conclusion

This paper provides a thorough investigation into land-use classification. Key contributions include: (1) the proposal of a novel network for fusing optical and SAR images for land-use classification, featuring the spatial-aware circular module that explores cross-modal correlations with a global receptive field, (2) a reevaluation of land-use classification challenges, particularly recognizing that different modalities should share a common semantic representation, which led to the design of a semantic distribution alignment loss function for matching semantic distributions and revealing complementary cues between modalities, and (3) experimental results demonstrating the significant advantages of our method. However, our approach does not consider the association of different land objects. For instance, combinations like city and road appear more frequently than city and farmland. This association information could be leveraged to further optimize the model

References

Chen, B., Huang, B., Xu, B., 2017. Multi-source remotely sensed data fusion for improving land cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 124, 27–39.

Chen, K., Li, W., Chen, J., Zou, Z., Shi, Z., 2022. Resolution-agnostic remote sensing scene classification with implicit neural representations. *IEEE Geoscience and Remote Sensing Letters*, 20, 1–5.

Fan, S.-K. S., Lin, Y., 2007. A multi-level thresholding approach using a hybrid optimal estimation algorithm. *Pattern recognition letters*, 28(5), 662–669.

Fraser, S., Storie, J. L., 2016. Using Geoindicators to Prioritize Regional Wetland Locations for Flood Attenuation in Manitoba's Red River Basin. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(8), 3578–3587.

Ienco, D., Interdonato, R., Gaetano, R., Minh, D. H. T., 2019. Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for land cover mapping via a multi-source deep learning architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 158, 11–22.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

Li, Q., Yang, W., Liu, W., Yu, Y., He, S., n.d. From Contexts to Locality: Ultra-high Resolution Image Segmentation via Locality-aware Contextual Correlation (Supplementary Material).

Li, W., Sun, K., Li, W., Wei, J., Miao, S., Gao, S., Zhou, Q., 2023. Aligning semantic distribution in fusing optical and SAR images for land use classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 199, 272–288.

Li, X., Zhang, G., Cui, H., Hou, S., Wang, S., Li, X., Chen, Y., Li, Z., Zhang, L., 2022. MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 106, 102638.

Luo, X., Tong, X., Pan, H., 2020. Integrating multiresolution and multitemporal Sentinel-2 imagery for land-cover mapping in the Xiongan New Area, China. *IEEE Transactions on Geoscience and Remote Sensing*, 59(2), 1029–1040.

Woo, S., Park, J., Lee, J.-Y., Kweon, I. S., 2018. Cbam: Convolutional block attention module. *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Yang, X., Li, S., Chen, Z., Chanussot, J., Jia, X., Zhang, B., Li, B., Chen, P., 2021. An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 177, 238–262.

Yue, C., Zhang, Y., Yan, J., Luo, Z., Liu, Y., Guo, P., 2024. BCLNet: Boundary contrastive learning with gated attention feature fusion and multi-branch spatial-channel reconstruction for land use classification. *Knowledge-Based Systems*, 302, 112387.

Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R., Stiefelhagen, R., 2023. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*.

Zhang, Y., Ye, M., Zhu, G., Liu, Y., Guo, P., Yan, J., 2024. FFCA-YOLO for small object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*.

Zhao, W., Du, S., 2016. Learning multiscale and deep representations for classifying remotely sensed imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 113, 155–165.