This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3574396

Wavelet-based Deep Generative Framework for Super-Resolution of Low-Resolution Labelled Maps and Weak-Supervised Learning

Abhishek Singh and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—The unavailability of pixel-level detailed labels is a crucial challenge in the field of remote sensing image analysis. Deep learning models require a large number of labeled samples for an accurate estimation of a large number of trainable parameters. However, in remote sensing applications usually only a few reliable labeled data are available for the learning of a classifier, whereas often many weak/low-resolution unreliable labeled data can be collected from available land-cover maps. Accordingly, weak supervised learning may overcome the problems by using noisy and low-resolution labels in remote sensing. In this paper, we propose a deep adversarial model based on discrete wavelet transform to exploit weak/low-resolution label information for generating refined super-resolved Weak Reference Maps (WRM). Our contribution includes the development of a discrete wavelet transform based generator for enhancing the low-resolution labels to generate a refined high-resolution reference map. We also present an efficient framework for multi-source image fusion that incorporates the refined super-resolved WRM, synthetic aperture radar images and corresponding low-resolution labels. Our findings highlight the effectiveness of the refined super-resolved WRM. Additionally, we investigate the impact of the high-resolution reference maps on segmentation accuracy, which reveals their potential in improving the segmentation performance compared to other reference methods.**

***Index Terms*—Wavelet Convolution, Multi-Source Image Analysis, Semantic Segmentation, Generative Model, Deep Learning, Weak Labels and Remote Sensing.**

I. Introduction

MULTI-SOURCE remote sensing data fusion is the process of combining information obtained from multiple sources of remote sensing data to create a more comprehensive and accurate representation of the target area or object being observed. The sources of remote sensing data can include aerial imagery, satellite imagery, UAV data and other forms of data collection. The data are typically acquired using different sensors having varying characteristics, such as spatial resolution, spectral resolution and temporal resolution. The process of multi-source data fusion involves combining these different datasets into a single integrated dataset, either by physically merging the data or by integrating the information through mathematical models and algorithms.

Data fusion techniques have many applications in remote sensing. One of the key domains that can benefit from these techniques is land-cover classification and ground object identification, which can be improved by increasing the spatial resolution of the data. To address the challenges arising from the trade-off between spatial resolution and temporal frequency, various remote sensing data fusion methods have been proposed. These methods can be broadly categorized into image pair-based and spatial unmixing-based techniques. The image pair-based fusion techniques involve combining two or more images that were captured at different times or from different viewpoints. The techniques are often used in remote sensing applications, where multiple images of the same area are captured using different sensors or at different times. The goal of image pair-based fusion is to create a single image that contains the most relevant information from all of the input images. On the other hand, spatial unmixing-based fusion techniques decompose the input images into their constituent parts and then recombine them in a way that preserves the most important features of each image. The techniques are often used in applications where there is a high degree of spectral mixing, such as in hyperspectral imaging. Both image pair-based and spatial unmixing-based fusion techniques can be used to create specific images that contain reliable information.

Deep Learning (DL) algorithms, such as Convolutional Neural Networks (CNN) [1] and Recurrent Neural Networks (RNN) [2], have shown great potential in handling large and complex datasets, making them ideal for remote sensing data fusion. These algorithms can extract high-level features from different sources of data and combine them to provide more accurate and detailed information about the Earth's surface. Some of the applications of multi-source remote sensing data fusion using deep learning include land use and land cover classification, crop monitoring, urban planning, disaster response, and environmental monitoring.

A DL model requires a large amount of labeled data for the estimation of the network parameters. In many remote sensing applications, the unavailability of a sufficient number of labeled data results in a major limitation [3]. CNN plays a significant role in many applications. However, collecting sufficient pixel-level labeled data for training a CNN is an expensive task and requires expert knowledge [4], [5]. Thus it is important to develop techniques that require a limited number of reliable labels and can exploit weak-supervised learning. Recently, weak-supervised learning has raised large attention as one of the solutions to deal with either limited or noisy labels and to minimise the gap between the performance of supervised and semi-supervised classification.

Goodfellow et al. [6] proposed Generative Adversarial Networks (GANs), which have been used in several tasks (e.g.

A. Singh and L. Bruzzone are with Remote Sensing Laboratory (RSLab), Department of Information Engineering and Computer Science, University of Trento, Italy

TABLE I
LIST OF SYMBOLS USED FOR REPRESENTING THE VARIABLES AND PARAMETERS IN THIS PAPER

Symbols Used	Description
$I(.)$	*Input Image*
(p,q)	*Indicates spatial location of Image*
n	*Indicates spectral channel of Image*
C	*Convolutional Operation*
K	*Convolutional Kernel*
$u \times v$	*Size of Kernel*
φ	*Activation Function*
b	*Bias*
l	*Convolutional Layer*
$D(.)$	*Discriminator*
$G(.)$	*Generator*
x	*Real Label*
x'	*Refined Super-Resolved WRM*
y	*Generated Images*
z	*Random Noise*
$\mathcal{L}$	*Objective Function of Generative Adversarial Model*
$X \times Y$	*Size of Wavelet Decomposition of each sub-band*
$\psi_r(.)$	*Wavelet Decomposition*
$L_{i,j}$	*Low-Frequency Components of ith row and jth column*
$H_{i,j}$	*High-Frequency Components of ith row and jth column*
L	*Low-pass Matrices*
H	*High-pass Matrices*
FM	*Feature Matrix*
P	*Principal Components Matrix*
V	*Covariance Matrix*
F	*Fusion of Features*
k	*Number of Features*
f_k	*Input Feature for Fusion*
e_k	*Convolutional Filter*
N	*Number of all the input channels after Fusion*
θ	*Parameter vector of update rule*
α	*Learning Rate*
β	*Decay Rate*
ϵ	*Small Constant*
$J(.)$	*Cost Function*
t	*Current Time Step*
g_t	*Gradient of the Cost Function*
m_t	*Moving Average of the Gradient*
$\hat{m}_t$	*Bias Corrected Estimates of* m_t
$\mathcal{WCE}(.)$	*Weighted Cross-Entropy Loss*
c	*Number of Classes*
γ_h	*True Probability of* h^{th} *class*
$\hat{\gamma_h}$	*Predicted Probability of the* h^{th} *class*
$\mathcal{W}$	*Vector of class weights*

image synthesis, image generation, image super-resolution, etc). They consist of a generator and a discriminator, which are trained simultaneously. Authors in [7] proposed a pixel-to-pixel mapping which utilizes the Conditional Generative Adversarial Networks (cGAN) as an approach to the image translation tasks to generate images by conditioning the input images. Image translation was successful in a range of tasks including colorization, image reconstruction, image synthesis, etc. Several researchers have used GANs for data augmentation tasks in remote sensing for obtaining better training data and thus classification results [8], [9], [10]. Generative Adversarial Networks (GAN)-based models are also widely used in change detection on multispectral remote sensing imagery for modeling the distribution between bi-temporal images. Other researchers have used adversarial training for annotated sample generation, super-resolution of remote sensing images, cloud removal, image classification, radar image sequence generation, synthetic aperture radar (SAR) image generation for target recognition and semi-supervised hyperspectral image classification.

Our work aims to exploit the generative characteristics of cGAN to progressively generate refined super-resolved WRM by incorporating the properties of wavelet transform. Wavelet Transform [11] can extract robust spatio-spectral features at different scales and orientations. The refined super-resolved WRM are then merged with features extracted from multi-spectral and SAR data in order to obtain weak/low-resolution labels for performing pixel-level classification.

The use of frequency-based methods, and in particular wavelet transforms, has been studied in computer vision research. Gal et al. [12] have used the style-based wavelet-driven generative model for content generation in the wavelet domain. Li et al. [13] proposed a framework for synthetic aperture radar images to optical images conversion using multiscale GAN based on wavelet feature learning. The main novelty of the proposed method consists in capturing weak label information of very low-resolution land-use-land-cover labels by using dedicated multiscale analysis using wavelet transform in order to generate refined super-resolved WRM. Within the scope of this paper, the term weak-supervised learning is limited to situations where the supervision provided is inexact and inaccurate, often referred to as label noise. However, there are several challenges associated with weakly supervised learning, including:

1) *Ambiguity in labels*: Weakly labeled data can be ambiguous or noisy, making it difficult for models to accurately learn from the data. For example, in image classification, an image may be labeled as belonging to a certain category, but it may also contain elements that are characteristic of other categories.
2) *Limited information*: Weakly labeled data often provides limited information about the underlying structure of the data. This can lead to models that are not as robust or accurate as those trained with reliable labeled data.
3) *Difficulty in evaluation*: Since weak-supervised learning often relies on imperfect labels, it can be difficult to evaluate the performance of models.

To address the aforementioned issues, we propose a pre-processing approach to refine weak land-use/land-cover labels for weak supervised learning. The main contributions of this paper are as follows:

1) We use a discrete-wavelet-transform-inspired generator to produce refined super-resolved WRM from low-resolution labels using the characteristics of principal components of the multispectral images.
2) We develop an efficient framework for multi-source image fusion which utilizes the characteristics of refined super-resolved WRM, SAR images and low-resolution labels.
3) We compare the effectiveness of the obtained feature map with those of other general generative model and study the impact of refined super-resolved WRM in terms of segmentation accuracy.

This paper is organized into six sections. Section II describes the background and state-of-the-art approaches. Section III presents the proposed approach to the fusion of multi-

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3574396

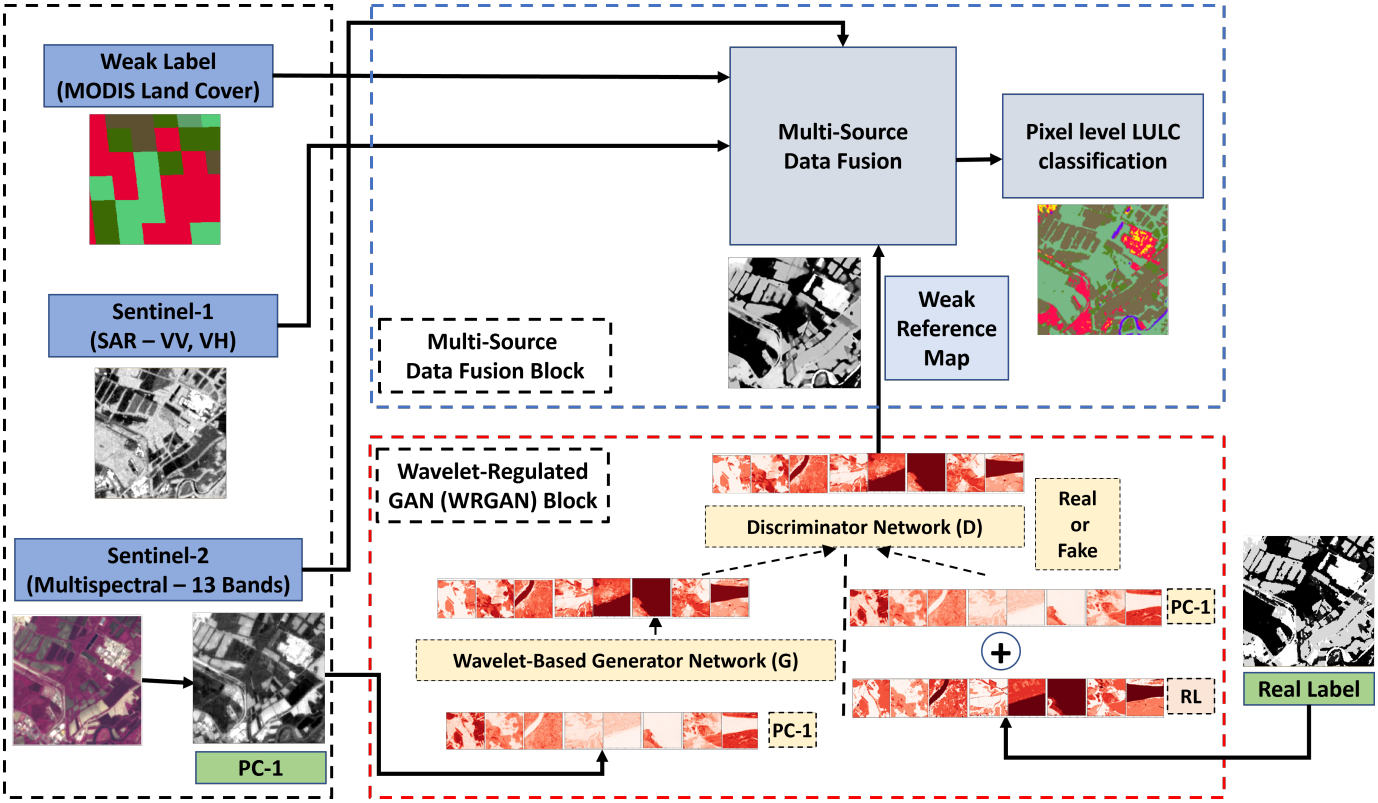

Fig. 1. Proposed framework for the weak-supervised based multi-source image classification

source satellite images and thus the classification method. Section IV describes the experimental data set and the design of the experimental analysis carried out for assessing the effectiveness of the proposed approach. Section V draws the conclusions of this paper.

II. Background

Supervised learning is a commonly used approach in DL and requires a massive amount of training data. Due to changes in natural landscapes and man-made objects in the physical world, annotated data get outdated over time. Another big challenge of DL in remote sensing is label noise. It is very difficult to arrange perfect large-scale data labeling due to the complexity of remote sensing data. But the availability of noisy labeled data can bias the model, whereas a small amount of reliable labels leads to overfitting. In order to deal with such challenges several researchers have used transfer learning [14] by executing pre-training on established benchmark datasets for various tasks that do not have sufficient labels. However, the pre-trained model performs well on similar datasets but it is not reliable on different datasets having different characteristics.

To improve classification accuracy, many researchers have worked on either enhancing the resolution of remote sensing data or incorporating auxiliary information. Using the classical DL models to solve the mapping on low and medium-resolution data has become challenging for many researchers in the remote sensing domain [15]. DL methods like CNN [1], RNN [16], GAN [6] and many more have been frequently employed in the remote sensing domain. At present, it is also common to incorporate the attention mechanism within the DL architectures [17], [18]. Ronneberger et al. [19] proposed a UNet CNN which was successful in utilising the labeled samples more effectively due to the data enhancement capability of the network. The modified network based on UNet has been used in various applications for the purpose of classification and segmentation [15], [20].

The increasing gap between the availability of large amount of remote sensing data and the lack of reliable labeled data opens an opportunity for the researchers to exploit the noisy labels, few labels, weak labels to learn valuable information. Semi-supervised and weak supervised learning offers a paradigm to approach the above challenges.

A. Weak-Supervised Learning (WSL)

With the development of DL model, Weak-Supervised Learning (WSL) has become a research topic in the field of computer vision. In an actual scenario, collecting supervision information requires a high cost and thus it is desirable to opt for weak supervision. In computer vision, the purpose of WSL is to utilise a few labels in order to achieve high accuracy in classification/segmentation. WSL uses high-level and often noisy labels to generate larger training sets. WSL is an umbrella term that addresses the variety of studies to construct predictive models by focusing on learning with incomplete, inexact and inaccurate supervision [21].

Incomplete supervision deals with a small amount of labeled data and abundant unlabeled data [22]. Semi-supervised learn-

ing and active learning come under this category [21]. In [23], [24], [25], examples of incomplete supervision are presented in terms of semi-supervised based scene classification. Inexact supervision deals with the situation in which some information is available, but not as exact as required [26]. In this category low-resolution or weak labels are usually considered for supervision. Multi-instance learning is an example of inexact supervision. Inaccurate supervision deals with a situation in which the supervision information is not always correct. In this category labeled samples may be affected by noise and errors [27]. The purpose of inaccurate supervision is to minimise the impact of noisy labels. Noisy labels refer to the labels that are different from their real land cover classes. Several researchers have used crowd-sourcing to collect labeled samples for remote sensing image classification [28], crop mapping [29], etc.

In pixel-level classification, weakly supervised learning has had a significant impact by enabling models to learn from partially labeled or imperfectly labeled data. This has the potential to significantly reduce the cost and time required for manual annotation of large datasets.

1) Relation with Semi-Supervised Learning: Both semi-supervised learning and weak-supervised learning approaches deal with situations where labeled data are limited or partially available. While there are similarities between these two methods, they differ in terms of the level of supervision they require and the techniques used to leverage the available data. Both semi-supervised learning and weakly supervised learning aim to overcome the limitations of fully supervised learning, where a large amount of accurately labeled data is required. Instead, they utilize partially labeled or weakly labeled data to train models.

2) Distinctions with Semi-Supervised Learning: Semi-supervised learning assumes that a small portion of the data is labeled, while the majority remains unlabeled. The labeled data are used to guide the learning process, while the unlabeled data helps in capturing the underlying structure of the data distribution. In contrast, weakly supervised learning deals with situations where only weak or noisy labels are available. This means that the labels provided for training may be incomplete, inexact, or inaccurate.

In remote sensing, semi-supervised learning has been studied in multispectral image processing [30], [31], hyperspectral image processing [32] and SAR-optical classification [33]. Semi-supervised methods have been used with classical machine learning (ML) methods in various remote sensing applications [34], [35]. Recently, SSL methods have been addressed with DL methods for scene classification [23], semantic segmentation [36], [37], cross-modal learning [38], etc. Semi-supervised methods have been also used to address the limitation of dimensionality reduction in hyperspectral data [39], [40]. In [39], authors used an iterative multitask regression framework for low-dimensional subspace by considering labeled and unlabeled data. In [40], semi-supervised local fisher discriminant analysis with pseudo labels is used to perform non-linear dimensionality reduction. Zhao et al. [41] exploited image inpainting as a pretext task for remote sensing scene classification with limited labeled samples. In [42], the authors utilized two pretext tasks in training through rotation and contrastive prediction for few-shot scene classification.

B. Generative Model in Remote Sensing

In recent years, semi-supervised deep learning methods have become the primary tool for remote sensing image analysis. Among them, the GAN [6] is used by researchers because it helps in solving the problem related to the lack of training data. Generative models are widely used in remote sensing applications. Autoencoder and GAN have been widely exploited to learn representation from the dataset. Autoencoder is a deep neural network consisting of a hidden layer in between an input and an output layer of the same size that is trained with backpropagation in an unsupervised way [43]. Autoencoders have the capability to reconstruct an input from the most significant features of the data. Zhang et al. [44] exploited a stacked autoencoder to learn multispectral image features for the change detection task. In remote sensing image analysis, autoencoders are used for image classification [45], [46], [47], hyperspectral image denoising [48], hyperspectral image unmixing [49], generative feature extraction [50], [51], etc.

Goodfellow et al., [6], introduced the basis for estimating generative tasks through GAN. Radford et al., [52] proposed deep convolutional GAN that can learn an unsupervised representation of images for generative modeling. Several studies have analyzed the state-of-the-art generative models for image-to-image translation for paired and unpaired images, and successfully synthesized images from label maps and edge maps [7]. Isola et al., [7] presented pixel-to-pixel mapping, which utilizes the conditional generative adversarial networks as a tool for the image-to-image translation by conditioning the input images to generate synthetic output images. They conducted several experiments in which image-to-image translation was successful in a range of tasks, including colorization, image reconstruction, and image synthesis.

Adversarial training has proven its efficacy in various tasks such as road detection from remote sensing images [53] and road segmentation from aerial remote sensing images [54]. Moreover, GANs have shown effectiveness in change detection on multispectral imagery by modeling the distribution between bi-temporal images [55]. This approach has also been applied in annotated sample generation [56], super-resolution of remote sensing images [57], cloud removal [58], [59], remote sensing image classification [8], digital surface model simulation [60], radar image sequence generation [61] and semi-supervised hyperspectral image classification [62]. Additionally, Wasserstein GAN loss has been proposed for hyperspectral unmixing by Ozkan et al. [63], while Cheng et al. [64] introduced perturbation-seeking GAN for remote sensing scene classification. Ren et al. [65] proposed sample weighting and class adversarial training strategies for synthetic aperture radar (SAR) image classification. In another study [66], spatiotemporal fusion was achieved for data augmentation using CycleGAN, leveraging its characteristics to generate synthetic images for image fusion purposes.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3574396

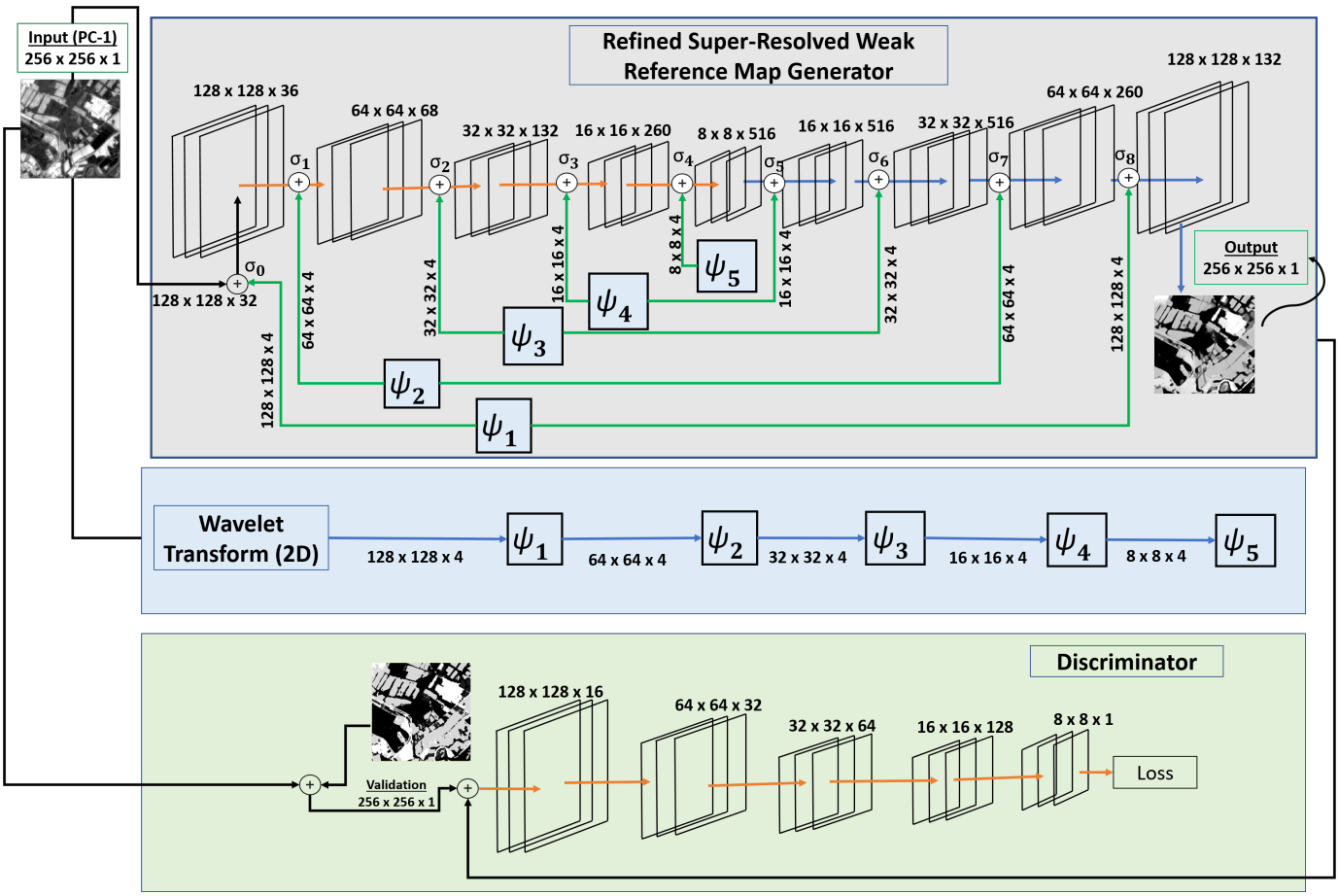

Fig. 2. Proposed methodology of the proposed Wavelet Regulated GAN (WRGAN) for refined super-resolved weak reference map generation.

C. Mathematics of Convolutional Neural Network (CNN) and Generative Adversarial Network (GAN)

CNN is a state-of-the-art tool to process 2D and 3D data [1]. It consists of three key components: convolutional layers, activation function, and pooling. Let us consider an image $I(p,q,n)$, where (p,q) indicates the spatial location, and n indicates the spectral band in input to a CNN with Kernel K, activation function φ and bias b of layer l, the convolution and pooling operations can be defined as:

$$C^{l}_{(p,q;n)} = \varphi^{l+1}(\sum_{u}\sum_{v} I(p+u-1, q+v-1; n)^{l} K(u,v;n) l + b) \quad (1)$$

$$pool(I^{l}-1)_{p,q,n} = \rho^{l}(I_{p+u-1,q+v-1,n}) \quad (2)$$

where the kernel has fixed u rows and v columns.

The GANs exploit the min-max operator in which the discriminator aims to maximize the objective function whereas the generator aims to minimize it. The min-max operator minimizes the divergence between the data and the distribution of deep features [67]. The objective function of the GAN can be written as:

$$\mathcal{L} = \min_{G}\max_{D} \mathbb{E}_{x}[log(D(x))] + \mathbb{E}_{z}[log(1 - D(G(z)))] \quad (3)$$

where $D(x)$ and $G(z)$ are the discriminative and generative image samples, respectively, x represents the input image and z is the noise used for generating synthetic training samples.

A conditional version of generative adversarial networks can control the method used for data generation, incorporating additional information such as class labels to monitor the artificial data generation process [68]. In conditional generative adversarial networks, conditional information can be exploited in the generator and the discriminator. The objective function of cGAN can be written as:

$$\mathcal{L} = \min_{G}\max_{D} \mathbb{E}_{x,y}[log(D(x,y))] + \mathbb{E}_{x,z}[log(1 - D(x, G(x,z)))] \quad (4)$$

where $D(x,y)$ and $G(x,z)$ are the discriminative and generative image samples, respectively, x represents the input image, y represents generated image and z is the noise used for generating synthetic training samples.

III. Proposed Method

Fig. 1 shows the overall framework for the proposed weak-supervised multi-source image classification approach.

A. Discrete Wavelet Transform Based Deep Adversarial Model

Fig. 2 illustrates the proposed Wavelet-Regulated Generative Adversarial Networks (WRGAN) for the refined super-resolved WRM generation using a wavelet-based generative framework. Our method is based on a generative approach and centered on discrete wavelet transform (WT), which decomposes an image into a sequence of channels that represent

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3574396

TABLE II
FEATURE MATRIX AFTER EACH BLOCK OF OPERATIONS IN THE PROPOSED GENERATOR; FM_{ψ}: MATRICES GENERATED AFTER WAVELET CONVOLUTION, $FM_{enc,dec}$: MATRICES GENERATED AFTER EACH LEVEL OF ENCODING AND DECODING

Feature Matrices		
Input: $256 \times 256 \times 1$		
FM_{ψ}	**Spectral concatenation**	$FM_{enc,dec}$
$128 \times 128 \times 4$	$128 \times 128 \times 36$	$128 \times 128 \times 32$
$64 \times 64 \times 4$	$64 \times 64 \times 68$	$64 \times 64 \times 64$
$32 \times 32 \times 4$	$32 \times 32 \times 132$	$32 \times 32 \times 128$
$16 \times 16 \times 4$	$16 \times 16 \times 260$	$16 \times 16 \times 256$
$8 \times 8 \times 4$	$8 \times 8 \times 516$	$8 \times 8 \times 512$
$16 \times 16 \times 4$	$16 \times 16 \times 516$	$16 \times 16 \times 512$
$32 \times 32 \times 4$	$32 \times 32 \times 516$	$32 \times 32 \times 512$
$64 \times 64 \times 4$	$64 \times 64 \times 260$	$64 \times 64 \times 256$
$128 \times 128 \times 4$	$128 \times 128 \times 132$	$128 \times 128 \times 128$
		$128 \times 128 \times 64$
		$256 \times 256 \times 32$
Output: $256 \times 256 \times 1$		

different ranges of frequencies. We use Haar WT with the convolutional operation that processes the data by computing the sum and difference of neighboring elements. Our generative model works with five-level decomposition, where each image is transformed into four sub-images, i.e., LL, LH, HL and HH components. The first sub-image, LL, corresponds to a low-frequency component, and the remaining sub-images, LH, HL, and HH, correspond to high-frequency components in the horizontal, vertical and diagonal directions, respectively. Iterating the use of these components on a given input image I, it is possible to obtain a multiscale decomposition:

$$\psi_r(I) = [\psi_{r-1}(L_{i,j}(I)), H_{i,j}(I)] \tag{5}$$

where, $\psi_0(I) = I$. $L_{i,j}, H_{i,j}$ are the low- and high-frequency components, respectively, of the i^{th} row and j^{th} column and $L = [L_{i,j}]$, $H = [H_{i,j}]$ are the low-pass and high-pass matrices, respectively.

B. Network Architecture and Training Procedure of Refined Super-Resolved Weak Reference Map Generator

Our approach integrates wavelet-aware techniques into the framework of the general conditional generative adversarial network. In every level of wavelet decomposition, each $2X \times 2Y$ image is represented by four sub-bands of size $X \times Y$. Within the generator, encoder blocks are composed of repeated convolutional layers, batch normalization and Leaky ReLU activation functions. The generator objective is to exploit the characteristics of the wavelet transform to produce refined super-resolved WRM, utilizing principal component features from corresponding high-resolution multispectral imagery. Meanwhile, the discriminator aims to distinguish between the refined label and the original label pairs. The principal component is calculated on multispectral images represented as a 3D array I of size $(p \times q \times n)$, where each pixel in the image has n spectral bands. Let V be the covariance matrix of I, then its eigenvectors and eigenvalues are given by:

$$V = \frac{1}{s-1} \sum_{i=1}^{s} (I_i - \overline{P})(I_i - \overline{I})^T \tag{6}$$

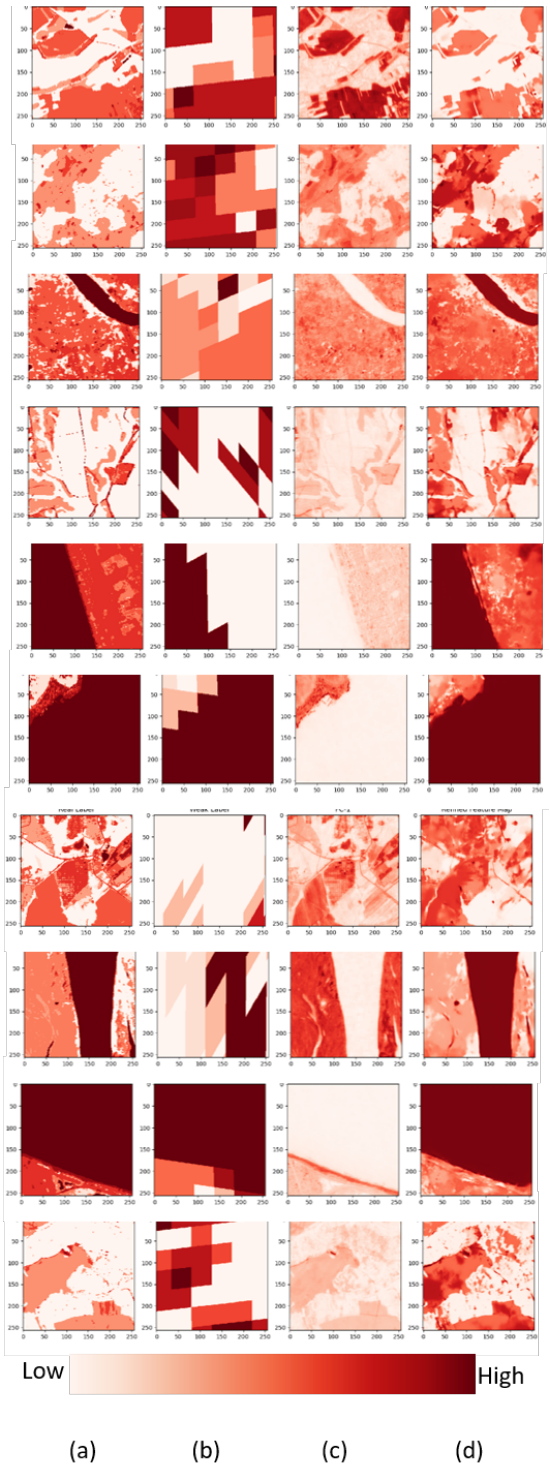

Fig. 3. (a) High-Resolution Map (Real Label), (b) Modis Land Cover (Weak Label), (c) Principal Component (PC-1) and (d) Weak Reference Map generated through the proposed Generator.

where $s = p \times q$ is the number of pixels in the image, and $\overline{I}$ is the mean vector of I.

The first r eigenvectors with the largest eigenvalues are used to construct the reduced dimensional representation of the data, which is given by:

$$P_{reduced} = P\mathbf{V}_r \tag{7}$$

TABLE III
COMPARISON BETWEEN THE PROPOSED WAVELET REGULATED GAN (WRGAN) AND THE CGAN ARCHITECTURE IN TERMS OF STRUCTURAL SIMILARITY INDEX (SSIM) AND MEAN SQUARED ERROR (MSE) ON THE TEST SET OF THE DFC2020 DATA THE AND VALIDATION SET OF CHESAPEAKE NEWYORK DATA.

Data	Metrics	cGAN	Proposed WRGAN
DFC2020 Test Dataset	**Mean Squared Error (MSE)**	33.60	21.18
	Structural Similarity Index (SSIM)	0.286	0.328
Chesapeake NewYork Validation Dataset	**Mean Squared Error (MSE)**	20.75	12.37
	Structural Similarity Index (SSIM)	0.029	0.244

TABLE IV
ANALYSIS OF THE EFFECTIVENESS OF THE PROPOSED WRGAN ACHIEVED ON THE VERY HIGH-RESOLUTION VALIDATION SET OF CHESAPEAKE NEWYORK DATA IN TERMS OF CLASSIFICATION METRICS OBTAINED BY USING PARAFORMER (L2HNET-V2) [69].

Metrics	L2HNet-v2	L2HNet-v2 + cGAN	L2HNet-v2 + Proposed WRGAN	Test Pixels
mIoU	43.50	45.08	**45.83**	143870568
OA	86.36	89.30	**89.43**	143870568
κ	76.76	81.47	**81.52**	143870568
Classes		Class-wise Scores		
Water	95.39	95.01	**95.81**	10010560
Forest	90.54	**91.56**	91.34	73174685
Low-Vegetation/Crop-Fields	85.25	88.81	**88.87**	59465016
Impervious	22.69	27.07	**31.94**	1220307

where $\mathbf{V}_r$ is a matrix containing the first r eigenvectors as columns.

We applied 2D-wavelet convolution at every stage of encoding and decoding operations. For instance, ψ_1, ψ_2, ψ_3, ψ_4, and ψ_5 represent the wavelet decompositions at the first, second, third, fourth and fifth levels, respectively derived from the input image I. Within the generator's decoder block, we employed iterative transposed convolution operations, followed by a stride operation and rectified linear unit (ReLU). This process reduces the number of spectral features by half while doubling the spatial feature size. The feature matrices resulting from wavelet convolution at each level are then concatenated spectrally with the feature maps obtained from each encoding and decoding stage. Table II illustrates the feature matrix after each operation block within the proposed generator.

The generator loss is calculated between the generated high-resolution weak reference images and real label images. The objective function of the proposed generative method is defined as follows:

$$\mathcal{L} = \min_{G_{\text{Wavelet}}} \max_{D} \mathbb{E}_{x,x'}[\log(D(x,x'))] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x,z)))] \quad (8)$$

where Generator (G) tries to minimise the loss against an adversarial Discriminator (D) that tries to maximise it. The resulting method is used for refining weak labels in which x represents the valid label, and z is the random noise used for generating refined super-resolved WRM. A prior random noise (z) and principal component information are given as input to the generator to produce the output (x'). The valid label (x), refined super-resolved WRM (x') and principal component information are fed to the discriminator model to calculate the adversarial loss.

C. Fusion of Features

The deep learning model is trained using a set of low-resolution labeled data that includes multispectral and SAR images, as well as the MODIS land cover data. The objective of the training is to optimize the weights and biases of the network to minimize the error between the predicted and actual land cover labels. The fusion of features from multispectral images, SAR images, refined super-resolved WRM and weak labels can be achieved using a deep learning model such as a convolutional neural network (CNN). Here is the general equation for fusing these features:

$$F = \varphi\left(\sum_{k=1}^{N} e_k * f_k + b\right) \quad (9)$$

where f_k is the k^{th} input feature, which are a multispectral band, a SAR channel, a refined super-resolved WRM and a weak label, e_k is the convolutional filter for the k^{th} feature, N is the number of input channels, F is the output feature map of the convolutional layer, b is the bias term and φ is the activation function. The fused features are used for land cover classification. The effectiveness of the fusion approach depends on the quality and relevance of the input features, the suitability of the network architecture and training strategy, and the complexity and variability of the target task and environment.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Experimental Data and Model Setup for Refined Super-Resolved Weak Reference Maps Generation

We have considered open source DFC2020 [70] and Chesapeake Land Cover [71] datasets for the refined super-resolved weak reference map generation.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3574396

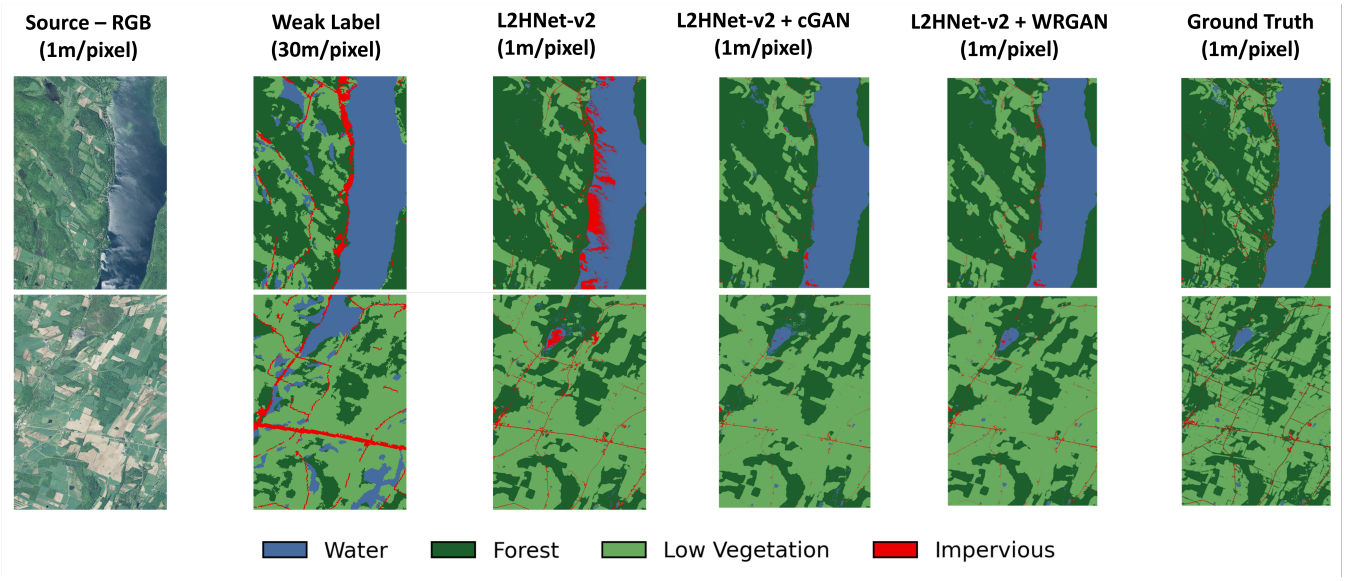

Fig. 4. Comparison of LULC maps generated after the inclusion of WRM in the very high-spatial resolution Chespeake NewYork dataset. Maps have been obtained by using using Paraformer (L2HNet-v2) [69].

1) DFC2020 Data: The experiments were conducted on a part of the open-source DFC2020 dataset in which we have considered low-resolution 500m-MODIS-derived land-cover maps and semi-manually derived high-resolution land-cover maps. This dataset originally consisted of a quadruple of Sentinel-1 SAR data, Sentinel-2 multispectral imagery, low-resolution land-cover and corresponding high-resolution land-cover maps. Each patch size is 256×256 pixels. We have used 2986 low-resolution/high-resolution pair samples and 3128 pair samples from the DFC2020 dataset to train and test the proposed WRGAN respectively.

2) Chesapeake Land Cover: To further validate the effectiveness of the proposed approach, we conducted experiments on the Chesapeake dataset, prepared by Microsoft Research. This dataset is made up of 1m very high-resolution imagery paired with a 30m low-resolution land-cover product for training, along with 1m very high-resolution ground truth data for evaluation. For our study, we focused on a subset of the dataset covering the New York region, which includes 125 large image tiles, each approximately 6000×7500 pixels in size. Out of these, we selected 20 tiles to train the proposed WRGAN, 100 tiles to train the classifier, and the remaining 5 tiles for validating both the generator and the classifier. Each tile was further divided into 224×224 patches for model training and evaluation.

For the training of the model, the Adam optimizer was used with a learning rate of 0.0002, the dropout rate was fixed to 0.5, stride and kernel size were set to 2×2 and 3×3, respectively. We have used mean absolute error and mean square error for calculating the generator loss and binary cross-entropy to analyze discriminator accuracy after each epoch. During the experimental training, the batch size is set to 25 for DFC2020 and 10 for Chesapeake NewYork data, and all models are trained for 500 epochs. Adam (Adaptive Moment Estimation) is an optimizer that combines the benefits of Root Mean Square Propagation (RMSprop) and momentum. RMSprop is a stochastic gradient descent optimizer that modifies the step size of the gradient descent algorithm in order to take steps proportional to the average of recent magnitudes of the gradients. It adjusts the step size of the gradient descent algorithm based on the moving average of both the gradient and its squared value, as well as a momentum term that accelerates learning in a consistent direction. The Adam update rule for a parameter vector θ with learning rate α, decay rates β_1 and β_2, and small constant ϵ can be expressed as:

$$g_t = \nabla_\theta J(\theta) \tag{10}$$

$$\hat{m}_t = \frac{\beta_1 m_{t-1} + (1-\beta_1) g_t}{1-\beta_1^t} \tag{11}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \tag{12}$$

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \tag{13}$$

where g_t is the gradient of the cost function $J(\theta)$ with respect to the parameter vector θ at time step t, m_t is the moving average of the gradient and its squared value, $\hat{m}_t$ is the bias-corrected estimates of m_t and t is the current time step.

B. Experimental Data and Model Setup for Multi-Source Image Fusion

For the multi-source feature fusion, the experiments were conducted on the multi-source SEN12MS dataset [72], which consists of triplet of Sentinel-1 SAR data, Sentinel-2 multispectral imagery and MODIS low-resolution land-cover maps. We have used $162,555$ triplet samples from the SEN12MS to train the model and $18,106$ triplet samples for validation. Each sample consists of 256×256 pixels. For the training of the classifier, the RMSprop optimizer was used with a learning rate of 0.01, the stride and kernel size were set to 2×2 and 3×3, respectively. RMSprop helps to smooth out the convergence process and prevent oscillations in the

optimization process. We have used cross-entropy to analyze loss after each epoch. During the experimental training, the batch size is set to 32, and all models are trained. The RMSprop update rule for a parameter vector θ with learning rate α and decay rate β can be written as:

$$E[g^2]_t = \beta E[g^2]_{t-1} + (1-\beta) g_t^2 \tag{14}$$

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{E[g^2]_t + \epsilon}} g_t \tag{15}$$

where g_t is the gradient of the cost function $J(\theta)$ with respect to the parameter vector θ at time step t, $E[g^2]_t$ is the exponentially-weighted moving average of the squared gradient, and ϵ is a small constant added for numerical stability. The experiments are conducted on AMD Ryzen 9 5950X 16-Core Processor 32T/16C, Nvidia NVIDIA GeForce RTX 3090, Ubuntu 20.04.2 LTS. We have used torch version 2.1.0, rasterio 1.3.9, sklearn 1.4.0, matplotlib 3.8.2, numpy 1.26.0 to implement the proposed approach.

In order to evaluate the sensitivity of the model to various parameters, we conducted controlled experiments by varying key hyperparameters, such as the learning rate, kernel size, stride and batch size, while keeping other settings constant. The results showed that the learning rate was the most sensitive parameter, with a noticeable impact on the convergence rate and model performance. For instance, reducing the learning rate below 0.001 led to slower convergence, whereas increasing it beyond 0.01 caused instability in the training process. The selected learning rate of 0.01 achieved an optimal balance between convergence speed and accuracy.

Cross-entropy loss is a commonly used loss function for classification problems, particularly for multiclass problems. It measures the dissimilarity between the predicted probability distribution and the true probability distribution of the labels. The weighted cross-entropy loss between the predicted probability distribution $\hat{y}$ and the true probability distribution y can be written as:

$$\mathcal{W}\mathrm{CE}(\gamma, \hat{\gamma}, w_h) = -\sum_{h=1}^{c} \mathcal{W}.\gamma_h \log(\hat{\gamma}_h) \tag{16}$$

where c is the number of classes, γ_h is the true probability of the h^{th} class, $\hat{\gamma}_h$ is the predicted probability of the h^{th} class and $\mathcal{W} = [w_1, w_2, w_3, ..., w_c]$ is the vector of class weights where w_h is the weight assigned to the h^{th} class. The weighted cross-entropy loss is minimized when the predicted probability distribution matches the true probability distribution and is non-negative.

The test experiments were conducted on the open-source DFC2020 dataset [70] consisting of a quadruple of Sentinel-1 SAR data, Sentinel-2 multispectral imagery, Modis low-resolution land-cover and the corresponding high-resolution land-cover maps. Each patch size is 256×256 pixels. We have used 2986 quadruple samples from the DFC2020 test set to predict the efficacy of the proposed method.

C. Qualitative and Quantitative Evaluation

TABLE V
ANALYSIS OF THE EFFECTIVENESS OF THE PROPOSED MULTI-SOURCE DATA FUSION CLASSIFICATION APPROACH RESULTS ACHIEVED ON THE TEST SET OF THE DFC2020 DATA FOR DIFFERENT COMBINATIONS OF INPUT DATA.

Dataset			Metrics		
Sentinel-2	**Sentinel-1**	**WRM**	**OA**	**AA**	κ
✓			50.47	36.06	37.32
✓	✓		63.19	40.97	50.77
✓	✓	✓	69.68	51.96	61.55

TABLE VI
RESULTS OF THE ABLATION STUDY OF THE PROPOSED METHOD ON THE TEST SET OF THE DFC2020 DATA. IN THE PROPOSED APPROACH, THE CBAM-UNET IS CONSIDERED AS A BASELINE FOR THE CLASSIFICATION EXPERIMENT

Metrics	**OA**	**AA**	κ
Baseline	63.19	40.97	50.77
Proposed (Baseline + WRM)	69.68	51.96	61.55
Proposed (Baseline + WRM + $\mathcal{W}$CE)	**77.34**	**55.28**	**70.79**

1) Refined Super-Resolved Weak Reference Map Generation: The generator is trained to generate output on a finite set of latent vectors as well as unseen latent vectors sampled during training. Moreover, to keep diversity we restrict the training up to a certain level in order to generate diverse feature maps and avoid overfitting. Table III shows the quantitative comparison of the generated results on DFC2020 and Chesapeake NewYork data obtained by comparing the proposed WRGAN with the general cGAN. In terms of key evaluation metrics: Mean Squared Error (MSE) and Structural Similarity Index (SSIM). These metrics were computed on the test set against the generated refined super-resolved WRM outputs. The results demonstrate that the proposed WRGAN obtained a significant improvement in MSE and SSIM, indicating better reconstruction quality and structural fidelity. These results highlight the effectiveness of the WRGAN in exploiting wavelet-based features to improve the quality of generated Weak Reference Map, while controlling the effectiveness of decomposition process and generated feature maps. The proposed method is successful in generating shape information of the principal component features. Fig. 3 shows the refined super-resolved WRM generated against the principal component using the proposed WRGAN. The results are plotted with weak labels and real labels for visualization purposes.

To further evaluate the effectiveness of the generated data, we employed Paraformer (L2HNet-v2) [69], for semantic segmentation as a downstream task. Specifically, we used the generated WRM data to train the semantic segmentation model on the training set of the Chesapeake New York dataset. We then evaluated the model on the validation set using 1m high-resolution reference ground truth labels obtained from the Chesapeake Bay Conservancy Land Cover (CCLC) project. Table IV presents the performance of the proposed WRGAN in terms of Mean Intersection over Union (mIoU), Overall Accuracy (OA), Kappa Coefficient (κ), and class-wise scores. The results demonstrate the impact of using the WRM generated by the proposed WRGAN, compared to both the

TABLE VII
COMPARISON OF THE CLASSIFICATION RESULTS OF THE PROPOSED METHOD AND THE REFERENCE METHODS ON THE TEST SET OF THE DFC2020 DATA.

Metrics (%)	Deeplabv3+ [73]	UNet [19]	Attention-UNet [74]	CBAM-UNet [75]	Segformer [76]	ViT-Encoder [77]	TransUNet [78]	Proposed
Overall Accuracy (OA)	57.51	57.47	62.12	63.19	55.82	57.08	59.24	**69.68**
Kappa Coefficient (κ)	47.38	44.63	50.31	50.77	44.00	45.95	45.84	**61.55**
Average Accuracy (AA)	44.08	41.61	39.91	40.97	42.61	44.80	43.30	**51.96**
Computational Complexity (iterations/s)	3.18	2.87	1.63	1.57	6.75	1.62	1.57	**1.587**

TABLE VIII
COMPARISON OF CLASS-WISE ACCURACY METRICS OBTAINED BY UNET, DEEPLABV3+ , ATTENTION-UNET, CBAM-UNET, SEGFORMER, VIT-ENCODER, TRANSUNET AND THE PROPOSED METHOD ON THE TEST SET OF THE DFC2020 DATA.

	Class-Wise Accuracy (%)							
LULC Classes	**Deeplabv3+ [73]**	**UNet [19]**	**Attention-UNet [74]**	**CBAM UNet [75]**	**Segformer [76]**	**ViT-Encoder [77]**	**TransUNet [78]**	**Proposed**
Forest	64.51	86.29	70.74	63.90	60.56	68.35	34.25	**84.59**
Shrubland	26.55	0.02	0.02	0.04	0.0	0.0	0.0	**37.05**
Grassland	26.81	12.66	7.56	21.76	18.30	17.42	3.23	**51.69**
Wetland	-	-	-	-	-	-	-	-
Cropland	47.13	57.35	68.88	71.05	56.65	59.59	80.54	**59.55**
Urban/Builtup	55.89	17.74	66.57	27.46	33.36	32.51	5.86	**36.62**
Barren	0.0	0.16	0.19	0.16	0.0	0.0	0.0	**77.30**
Water	96.37	98.33	98.32	98.41	96.34	96.42	96.37	**99.44**

cGAN and the scenario without any GAN-based enhancement. One can observe that WRGAN improves the classification of minority classes like Impervious and the class-wise scores for Water, reflecting in a better OA. Fig. 4 highlights the impact of the inclusion of WRM in the land cover maps.

The proposed WRGAN has a broad application prospect in the semi-labeling tasks. It is worth noting that the refined super-resolved WRM obtained by the proposed method model the principal component feature by utilizing multiscale wavelets. The proposed generator learns the mapping from principal component features to wavelet features with convolutional features and then collectively generates the refined super-resolved WRM.

2) Multi-Source Data Segmentation Accuracy: The generator of the proposed WRGAN has been used to generate refined super-resolved WRM by considering the principal component samples derived from the multi-spectral samples of SEN12MS into refined super-resolved WRM. Then we have used refined super-resolved weak reference maps with weak/low-resolution labels for the classification of Sentinel-1 and Sentinel-2 images. We considered CBAM-UNet [75] as a baseline in the proposed classification experiment. Table V shows the accuracy achieved by integrating Sentinel-1, Sentinel-2 and the WRM generated by WRGAN. The results highlight the impact of fusing the structural information of Sentinel-1 with the spectral information of Sentinel-2 data, which involves a significant improvement across all metrics. Furthermore, the addition of the WRM further increases the performance pointing out the effectiveness of WRM in enhancing feature representation by preserving multi-scale details and improving the fusion process. The ablation study presented in Table VI shows that the integration of the WRM with the baseline significantly increase the accuracy highlighting its ability to improve feature representations by capturing details from fused multi-source data. When the $\mathcal{W}$CE is added to the WRM-enhanced baseline, the accuracy obtained outperforms the other combinations. This enhancement demonstrates the effectiveness of $\mathcal{W}$CE in addressing class imbalance, further refining model performance.

We have conducted a comprehensive evaluation of our proposed method against several well established literature techniques, including Segformer [76], ViT-Encoder [77] and TransUNet [78], along with widely used architectures such as Deeplabv3+ [73], UNet [19], Attention-UNet [74] and CBAM-UNet [75]. Table VII shows the classification results achieved on the DFC2020 dataset which contains 2986 quadruples of Sentinel-1, Sentinel-2 images, MODIS low-resolution land-cover labels and corresponding high-resolution land-cover maps. It shows the classification accuracies of the proposed method in terms of OA, κ and AA. The proposed method after the addition of WRM and weighted cross-entropy $\mathcal{W}$CE loss shows better results compared to the reference methods. These results confirms the robustness and reliability of our approach in capturing spatial and spectral information. Additionally, while the ccomputational complexity in terms of iterations processed per second of the proposed approach (1.587 it/s) is comparable to that of other transformer-based models.

Table VIII shows the comparison of class-wise accuracy metrics of UNet [19], Deeplabv3+ [73], Attention-UNet [74], CBAM-UNet [75], Segformer [76], ViT-Encoder [77], TransUNet [78] and the proposed approach. The table points out the impact of refined features in identifying classes like grassland and barren, which were not identified by general UNet [19], Attention-UNet, CBAM-UNet [75], Segformer, ViT-Encoder and TransUNet. Fig. 5 visualizes the comparison of pixel-level classification of UNet, Attention-UNet and the proposed method. As one can see the proposed method is successful in creating boundaries between classes, whereas the reference methods mixes the forest, croplands and grassland classes.

The proposed method demonstrates its effectiveness in addressing multi-source image classification challenges with low-resolution labels on a large-scale and utilizing diverse remote sensing data sources for pixel-level classification.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3574396

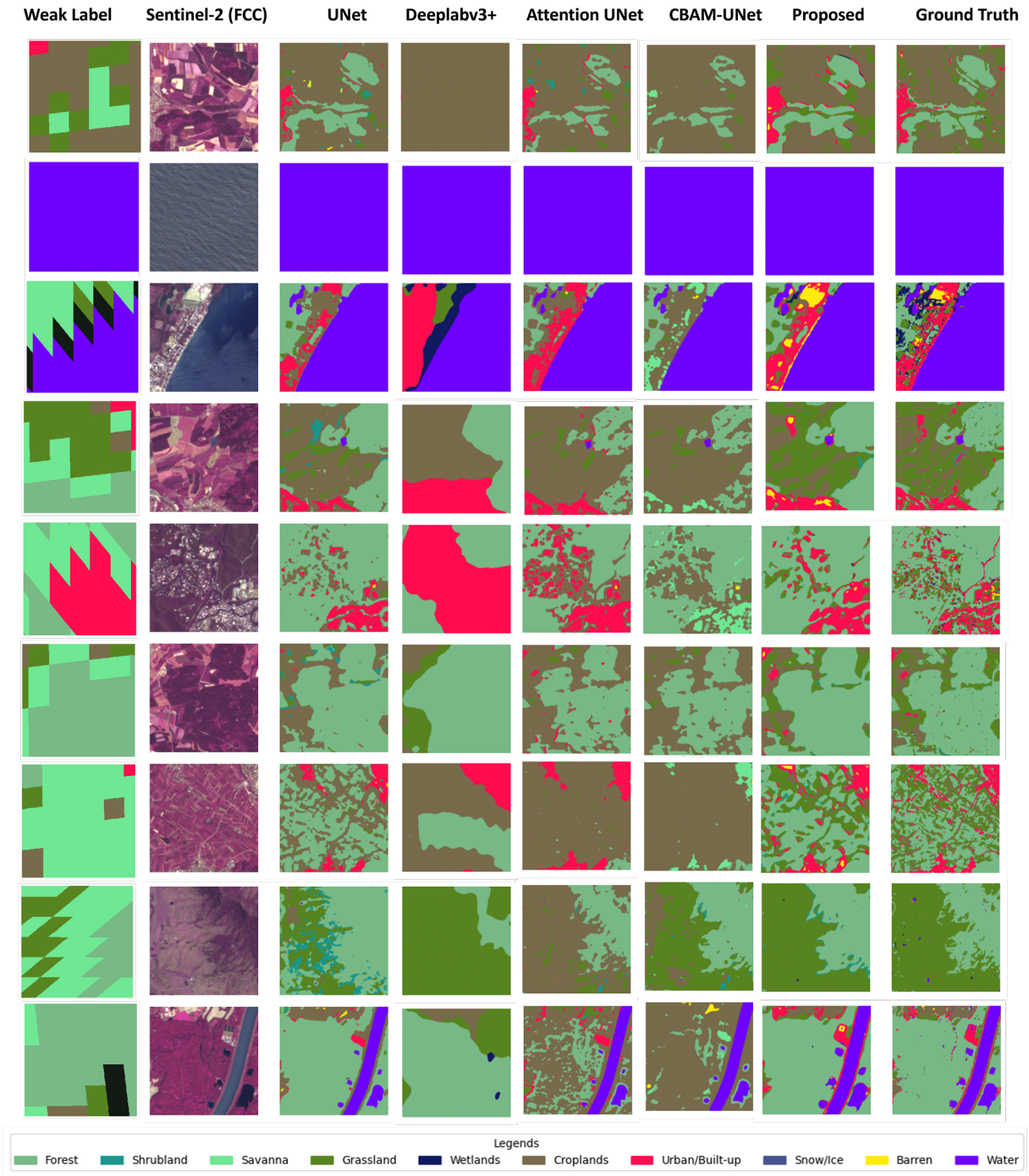

Fig. 5. Comparison of LULC maps generated through general UNet, Attention UNet, CBAM-UNet and the proposed method on the test set of the DFC2020 data.

There are still some challenges associated with the proposed method in pixel-level classification, because of ambiguity in labels. Wavelet transforms can provide localized frequency information and the transformed coefficients obtained through wavelet decomposition may not have direct and intuitive interpretations, making it difficult to understand the underlying patterns and features captured by the generative model. One limitation is that the obtained maps also contain artifacts, when fusing multiple sources of remote sensing data, uncertainties associated with each individual dataset can propagate into the fused data.

V. Conclusion

We have presented an approach which consists of two parts. First, refined super-resolved WRM are generated from the principal component using the generator of the proposed WRGAN. Then they are used along with the multispectral and SAR images for pixel-level classification. The refined super-resolved WRM generation approach is based on the exploitation of the discrete wavelet transform during the train-

ing of the generative model. The use of a frequency domain based convolutional approach to the generation task provides a specific improvement. The generation results confirm that the loss function implemented in the proposed WRGAN is more reliable compared to the general cGAN loss for generating refined samples. The fusion of decomposed multiscale wavelet features and image convolution features makes the network generate high-quality refined feature maps. In the multi-source features fusion task, we incorporate the refined feature maps as input to the classifier for weak supervised pixel-level classification. Our method is effective in solving remote sensing problems on a very large scale, which employs different available remote sensing data sources from Earth observation.

In future work, weak-supervised based semantic segmentation will be extended to a self-supervised/unsupervised approach for large-scale mapping. However, we will focus on developing specialized architectures, incorporating domain knowledge to improve model interpretability and develop lightweight models. We also aim to incorporate Point Spread Function (PSF) of the sensors in the model to better represent the spatial-context for robust modeling of spatial uncertainties.

REFERENCES

[1] Y. LeCun, Y. Bengio, *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[2] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural computation*, vol. 1, no. 2, pp. 270–280, 1989.

[3] L. Bruzzone, "Multisource labeled data: An opportunity for training deep learning networks," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4799–4802, IEEE, 2019.

[4] K. Li, L. Wang, L. Liu, Q. Ran, K. Xu, and Y. Guo, "Decoupling makes weakly supervised local feature better," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15838–15848, 2022.

[5] P. Sellars, A. I. Aviles-Rivero, and C.-B. Schönlieb, "Laplacenet: A hybrid graph-energy neural network for deep semisupervised classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[8] P. Yan, F. He, Y. Yang, and F. Hu, "Semi-supervised representation learning for remote sensing image classification based on generative adversarial networks," *IEEE Access*, vol. 8, pp. 54135–54144, 2020.

[9] A. Singh and L. Bruzzone, "Sigan: Spectral index generative adversarial network for data augmentation in multispectral remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[10] A. Singh and L. Bruzzone, "Data augmentation through spectrally controlled adversarial networks for classification of multispectral remote sensing images," in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 651–654, IEEE, 2022.

[11] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[12] R. Gal, D. C. Hochberg, A. Bermano, and D. Cohen-Or, "Swagan: A style-based wavelet-driven generative model," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–11, 2021.

[13] H. Li, C. Gu, D. Wu, G. Cheng, L. Guo, and H. Liu, "Multiscale generative adversarial network based on wavelet feature learning for sar-to-optical image translation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[14] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[15] A. Singh and L. Bruzzone, "Mono-and dual-regulated contractive-expansive-contractive deep convolutional networks for classification of multispectral remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10705–10714, 2019.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, Springer, 2015.

[20] A. Singh and L. Bruzzone, "Wianet: A wavelet-inspired attention-based convolution neural network for land cover classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[21] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.

[22] Y. He, X. Yuan, S. Chen, and X. Wu, "Online learning in variable feature spaces under incomplete supervision," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 4106–4114, 2021.

[23] W. Han, R. Feng, L. Wang, and Y. Cheng, "A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 23–43, 2018.

[24] X. Dai, X. Wu, B. Wang, and L. Zhang, "Semisupervised scene classification for remote sensing images: A method based on convolutional neural networks and ensemble learning," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 6, pp. 869–873, 2019.

[25] J. Zhang, M. Zhang, B. Pan, and Z. Shi, "Semisupervised center loss for remote sensing image scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1362–1373, 2020.

[26] Y.-F. Li, L.-Z. Guo, and Z.-H. Zhou, "Towards safe weakly supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 334–346, 2019.

[27] S. Liao, X. Jiang, and Z. Ge, "Weakly supervised multilayer perceptron for industrial fault classification with inaccurate and incomplete labels," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 2, pp. 1192–1201, 2020.

[28] E. Saralioglu and O. Gungor, "Crowdsourcing-based application to solve the problem of insufficient training data in deep learning-based classification of satellite images," *Geocarto International*, vol. 37, no. 18, pp. 5433–5452, 2022.

[29] S. Wang, S. Di Tommaso, J. Faulkner, T. Friedel, A. Kennepohl, R. Strey, and D. B. Lobell, "Mapping crop types in southeast india with smartphone crowdsourcing and deep learning," *Remote Sensing*, vol. 12, no. 18, p. 2957, 2020.

[30] W. Miao, J. Geng, and W. Jiang, "Semi-supervised remote-sensing image scene classification using representation consistency siamese network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[31] P. Gómez and G. Meoni, "Msmatch: Semisupervised multispectral scene classification with few labels," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 11643–11654, 2021.

[32] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. X. Zhu, "Joint and progressive subspace analysis (jpsa) with spatial–spectral manifold alignment for semisupervised hyperspectral dimensionality reduction," *IEEE Transactions on Cybernetics*, vol. 51, no. 7, pp. 3602–3615, 2020.

[33] A. Montanaro, D. Valsesia, G. Fracastoro, and E. Magli, "Semi-supervised learning for joint sar and multispectral land cover classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[34] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive svm for semisupervised classification of remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 11, pp. 3363–3373, 2006.

[35] G. Camps-Valls, T. V. B. Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE transactions on Geoscience and Remote Sensing*, vol. 45, no. 10, pp. 3044–3054, 2007.

[36] X. Sun, A. Shi, H. Huang, and H. Mayer, "BAS^4net: Boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5398–5413, 2020.

[37] 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 Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[38] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-modalnet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 167, pp. 12–23, 2020.

[39] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. X. Zhu, "Learning to propagate labels on graphs: An iterative multitask regression framework for semi-supervised hyperspectral dimensionality reduction," *ISPRS journal of photogrammetry and remote sensing*, vol. 158, pp. 35–49, 2019.

[40] H. Wu and S. Prasad, "Semi-supervised dimensionality reduction of hyperspectral imagery using pseudo-labels," *Pattern Recognition*, vol. 74, pp. 212–224, 2018.

[41] C. Tao, J. Qi, W. Lu, H. Wang, and H. Li, "Remote sensing image scene classification with self-supervised paradigm under limited labeled samples," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2020.

[42] H. Ji, Z. Gao, Y. Zhang, Y. Wan, C. Li, and T. Mei, "Few-shot scene classification of optical remote sensing images leveraging calibrated pretext tasks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[43] G. E. Hinton and R. Zemel, "Autoencoders, minimum description length and helmholtz free energy," *Advances in neural information processing systems*, vol. 6, 1993.

[44] P. Zhang, M. Gong, L. Su, J. Liu, and Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 116, pp. 24–41, 2016.

[45] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 9, pp. 5148–5157, 2017.

[46] R. Kemker and C. Kanan, "Self-taught feature learning for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2693–2705, 2017.

[47] L. Mou, P. Ghamisi, and X. X. Zhu, "Unsupervised spectral–spatial feature learning via deep residual conv–deconv network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 1, pp. 391–406, 2017.

[48] X. Wang, Z. Luo, W. Li, X. Hu, L. Zhang, and Y. Zhong, "A self-supervised denoising network for satellite-airborne-ground hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[49] B. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Blind hyperspectral unmixing using autoencoders: A critical comparison," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1340–1372, 2022.

[50] A. Singh and L. Bruzzone, "Refining land-cover weak labels using a discrete wavelet transform inspired deep adversarial model," in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, 2023.

[51] P. Naik, M. Dalponte, and L. Bruzzone, "Generative feature extraction from sentinel 1 and 2 data for prediction of forest aboveground biomass in the italian alps," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 4755–4771, 2022.

[52] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[53] Q. Shi, X. Liu, and X. Li, "Road detection from remote sensing images by generative adversarial networks," *IEEE access*, vol. 6, pp. 25486–25494, 2017.

[54] Y. Li, B. Peng, L. He, K. Fan, and L. Tong, "Road segmentation of unmanned aerial vehicle remote sensing images using adversarial network with multiscale context aggregation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2279–2287, 2019.

[55] M. Gong, X. Niu, P. Zhang, and Z. Li, "Generative adversarial networks for change detection in multispectral imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2310–2314, 2017.

[56] D. Ma, P. Tang, and L. Zhao, "Siftinggan: Generating and sifting labeled samples to improve the remote sensing image scene classification baseline in vitro," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 7, pp. 1046–1050, 2019.

[57] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "A new deep generative network for unsupervised remote sensing single-image super-resolution," *IEEE Transactions on Geoscience and Remote sensing*, vol. 56, no. 11, pp. 6792–6810, 2018.

[58] Y. Li, S. Tang, R. Zhang, Y. Zhang, J. Li, and S. Yan, "Asymmetric gan for unpaired image-to-image translation," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5881–5896, 2019.

[59] J. Park, D. K. Han, and H. Ko, "Fusion of heterogeneous adversarial networks for single image dehazing," *IEEE Transactions on Image Processing*, vol. 29, pp. 4721–4732, 2020.

[60] P. Ghamisi and N. Yokoya, "Img2dsm: Height simulation from single imagery using conditional generative adversarial net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 794–798, 2018.

[61] C. Zhang, X. Yang, Y. Tang, and W. Zhang, "Learning to generate radar image sequences using two-stage generative adversarial networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 3, pp. 401–405, 2019.

[62] Y. Zhan, D. Hu, Y. Wang, and X. Yu, "Semisupervised hyperspectral image classification based on generative adversarial networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 2, pp. 212–216, 2017.

[63] S. Ozkan and G. B. Akar, "Spectral unmixing with multinomial mixture kernel and wasserstein generative adversarial loss," *arXiv preprint arXiv:2012.06859*, 2020.

[64] G. Cheng, X. Sun, K. Li, L. Guo, and J. Han, "Perturbation-seeking generative adversarial networks: A defense framework for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[65] Z. Ren, B. Hou, Q. Wu, Z. Wen, and L. Jiao, "A distribution and structure match generative adversarial network for sar image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 6, pp. 3864–3880, 2020.

[66] J. Chen, L. Wang, R. Feng, P. Liu, W. Han, and X. Chen, "Cyclegan-stf: Spatiotemporal fusion via cyclegan-based image generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5851–5865, 2020.

[67] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[68] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[69] Z. Li, W. He, J. Li, F. Lu, and H. Zhang, "Learning without exact guidance: Updating large-scale high-resolution land cover maps from low-resolution historical labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27717–27727, June 2024.

[70] M. Schmitt, L. Hughes, P. Ghamisi, N. Yokoya, and R. Hansch, "2020 ieee grss data fusion contest," 2019.

[71] C. Robinson, L. Hou, K. Malkin, R. Soobitsky, J. Czawlytko, B. Dilkina, and N. Jojic, "Large scale high-resolution land cover mapping with multi-resolution data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12726–12735, 2019.

[72] M. Schmitt, L. Hughes, C. Qiu, and X. Zhu, "Sen12ms-a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, pp. 153–160, 2019.

[73] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

[74] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

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

[76] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation

with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12077–12090, 2021.

[77] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[78] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.