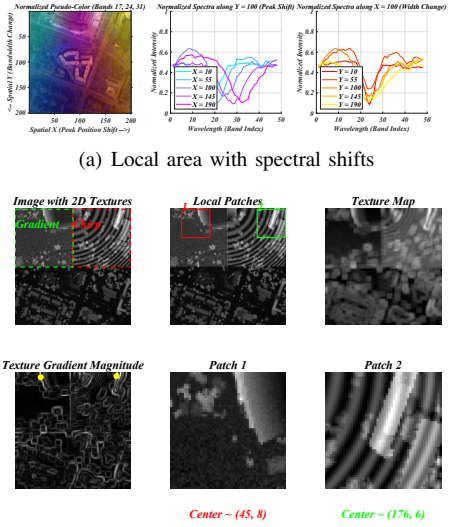
Chirplet Fourier Analysis Network for Cross-Scene Classification of Multisource Remote Sensing Data

Xudong Zhao, Qi Ming, Yixiao Yang, Wenshuai Hu, Wei Li, *Senior Member, IEEE*, and Ran Tao, *Senior Member, IEEE*

***Abstract*—The joint application of multisource remote sensing (MSRS) data, such as hyperspectral image (HSI) and light detection and ranging (LiDAR), offers significant potential for accurate land cover classification. However, the existing applications often struggle with domain shifts across scenes caused by sensor, illumination, and phase variations. Focusing on this domain adaptation problem, a Chirplet Fourier analysis network (ChirpFAN) is proposed for cross-scene classification of MSRS data in this paper. Firstly, a fractional spatial-frequency-phase feature extraction module including the fractional Fourier transform and a learnable phase-aware weighting block is proposed to capture multi-domain features. Secondly, a Chirplet swin transformer (ChirpST) block integrates a Chirplet Fourier analysis (ChirpFA) layer within a Swin transformer is designed to analyze multi-scale textural and oscillatory patterns. Finally, a modality-shared network including ChirpST blocks is designed for inter-modal fusion and alignment. Extensive experiments demonstrate that the ChirpFAN framework achieves state-of-the-art performance with 3% average improvements on three challenging cross-scene MSRS datasets. Code will be released on GitHub.**

***Index Terms*—Chirplet Fourier analysis (ChirpFA), fractional Fourier transform (FrFT), hyperspectral image (HSI), light detection and ranging (LiDAR), cross-scene classification.**



(a) Local area with spectral shifts

(b) Texture gradients in MSRS

Fig. 1. HSI/LiDAR patches with non-stationary, frequency-modulated features. (a) Spectral shifts example in HSI patches, (b) Texture gradients and Synthetic Chirplet dispersion in MSRS data.

I. Introduction

Recent progress in remote sensing technologies leads to the acquisition of large-scale and high-quality earth observation (EO) data [1], [2]. The exponential data driven by an increasing number of satellites present both opportunities and analytical challenges. To enhance the accuracy and robustness of land cover classification, the focus of researchers has turned from single data sources, such as hyperspectral images (HSIs), to the integration of multi-source remote sensing (MSRS) data [3], [4]. Among these sensor types, HSIs can provide detailed spectral information while light detection and ranging (LiDAR) data offers precise three-dimensional structural information. The joint utilization of HSI and LiDAR has demonstrated their complementary features and shown improvement of classification performance [5], [6].

Conventional MSRS classification methods typically operate under the assumption that training and testing samples are with same statistical distributions and pertain to a localized scene [7], [8]. However, in practical scenarios, there are variations in sensor, light, seasonal changes, and atmospheric conditions, which lead to spectral and spatial texture discrepancies between different scenes [9], [10]. Therefore, domain shifts exist between the source data used for training and the target data where the classification models are applied [11]. Moreover, the acquisition of extensive and accurately labeled MSRS datasets for both the source and target scenes is labor-intensive and even impossible in real tasks, which degrades the generalization capability and performance.

Domain shifts in MSRS data can be seen as modulations of the standard signals due to physical processes such as atmospheric effects, illumination changes, and sensor responses. As shown in Fig. 1, spectral shifts can be seen as frequency modulations of spectral curves and texture gradients include spatial modulations in textures and intensity variations. These modulations are often non-linear and non-stationary, which varies across the images. These facts require techniques that can model such intricate and non-stationary modulations and preserve fine-grained phase information, which is a challenge for many existing domain adaptation methods [12]–[14].

Various domain adaptation methods are proposed focus-

This work was supported in part by the National Natural Science Foundation of China under Grant 62401049, in part by the China Postdoctoral Science Foundation under Grant 2023M740268 and Grant 2024M754088, in part by the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation under Grant GZC20242184. (Corresponding author: Qi Ming, Yixiao Yang.)

Xudong Zhao, Wen-Shuai Hu, Wei Li, and Ran Tao are with the School of Information and Electronics, Beijing Institute of Technology, Beijing, 100081, China, and also with the National Key Laboratory of Science and Technology on Space-Born Intelligent Information Processing, 100081, China (e-mail: zhaoxudong@bit.edu.cn; wshuswjtu@163.com; liwei089@ieee.org; rantao@bit.edu.cn)

Qi Ming is with the College of Computer Science, Beijing University of Technology, Beijing, 100124 , China (e-mail: chaser.ming@gmail.com)

Yixiao Yang is with the Centre for Bioimaging Sciences, National University of Singapore, Singapore 117551, Singapore (e-mail: yixiao24@nus.edu.sg).

ing on domain shift [15], [16], which can be classified into three kinds. 1) Statistical methods aim to minimize domain shifts by aligning statistical distributions [17]–[20]. 2) Geometric approaches include subspace alignment and optimal transport [21]. 3) Deep domain adaptation methods include adversarial training and self-supervised learning [22], [23]. However, when applying domain adaptation methods to MSRS data, it remains challenging due to the non-linear inter-modal relations across sources [24]–[27]. Shared-private feature alignment techniques were proposed to obtain domain-invariant features and more robust alignment [28]. However, combining sources without alignments and addressing individual shifts may increase heterogeneity, which degrades performance. Therefore, more efficient domain adaptation methods are required to effectively fuse features while addressing domain shifts between cross-scene MSRS data.

The above supervised learning methods usually depend on high-quality labeled MSRS datasets, limiting their robustness [29], [30]. In contrast, the self-supervised learning (SSL) methods use feature augmentation in transform domains has demonstrated effectiveness in utilizing the multi-dimensional features of labeled data while learning fine-grained features from unlabeled data [31]–[33]. These feature augmentation and unlabeled feature extraction methods improve the generalization by learning more intrinsic representations in transformed domain [34]. Further, recent advances in unsupervised hyperspectral classification have also shown performance in learning discriminative features from unlabeled data [35]–[38].

The signal-level modulations of domain shift is a core physical property. While existing adaptation methods attempt to solve this problem at the statistical or geometric level, our work proposes to tackle the signal modulation directly. However, research on transform domain-based feature reconstruction and augmentation for MSRS data still confronts challenges [39]–[43]. Firstly, conventional spatial-frequency domain analyses usually miss the phase feature, which is known to include textural and structural information. Then it leads to the loss of details during feature reconstruction and transfer, limiting the discrimination in complex scenes. Secondly, the analysis of patches with small-sized local spatial-frequency windows makes it difficult to establish reliable global and local relationships for cross-scene data [44], [45]. Local windows possess a limited receptive field, which fail to incorporate contextual information to reduce feature shifts and maintain consistency. Thirdly, it is significant to analyze the multi-dimensional, multi-modal, and multi-domain information of cross-scene MSRS data. A comprehensive understanding is essential for sufficient training networks using limited training samples.

To address these signal-level domain shifts in MSRS data, we turn to Chirplet analysis, which is designed for non-stationary signals exhibiting varying frequencies. Compared with traditional Fourier and wavelet transforms, its parametric flexibility (controlling time/frequency localization, scale, and chirp rate) allows it to create adaptive space-frequency matched filters. Domain shifts often manifest as frequency modulations in observed signals, e.g., spectral features shift, and textures exhibit spatially varying frequencies due to geometric effects. Chirplets can model such linear frequency modulations, preserving phase dynamics associated with these modulations for textural and structural feature extraction. This intrinsic ability to adaptively localize and model frequency-modulated components makes Chirplets suit to represent the complex signal distortions caused by domain shifts in MSRS data. Focusing on the above issues, a Chirplet Fourier analysis network (ChirpFAN) is proposed for cross-scene classification of MSRS data. The main contributions are summarized as follows.

1) A learnable Chirplet Fourier analysis (ChirpFA) method is proposed for the first time for MSRS cross-scene classification. Its properties enable an efficient and effective analysis of the non-stationary and frequency-modulated representations of MSRS textural and structural features.
2) A fractional spatial-frequency-phase (FrSFP) feature extraction module is designed to jointly analyze MSRS data. The FrSFP can extract robust and discriminative features in an optimal domain between space and frequency, while preserving phase information.
3) A Chirplet swin transformer (ChirpST) is designed to analyze texturally rotated and frequency-modulated patterns in MSRS data. Compared with spatial-frequency analysis using fixed frequency components, it models local frequency variations and enables contexture extraction which are robust to domain shifts.

The remainder of this paper is organized as follows. Section II presents the related work. The Chirplet Fourier analysis theory is proposed in Section III. The proposed ChirpFAN framework is introduced in Section IV. In Section V, we present the experimental results and corresponding analysis. Finally, Section VI summarizes with some concluding remarks.

II. Related Work

A. Time-Frequency Analysis in Deep Learning

Deep learning has seen various integrations of time-frequency analysis, such as wavelet-based networks [46] and scattering transforms [47], which provide multi-resolution analysis using fixed or learnable wavelets. However, they cannot represent signals with frequency modulations unless specialized wavelets are employed. This is a limitation when dealing with MSRS data, where domain shifts induce such complex modulations, e.g., spectral warping, geometric texture distortions. Gabor filters [48] have been used in CNNs to extract textural features, but they typically rely on predefined filters or learnable filters with fixed tiling. Existing Fourier analysis networks (FAN) [49] incorporate the Fourier series to model periodicity, but focus on fixed frequencies and cannot model time-varying frequencies. The proposed Chirplet Fourier analysis (ChirpFA) extends this by incorporating learnable chirp rates and enabling the modeling of linear frequency modulations, which are significant for representing domain shifts found in MSRS data.

B. Fractional Fourier Analysis

The fixed-basis signal analysis approaches are unable to adapt to the diverse and localized time-frequency structures.

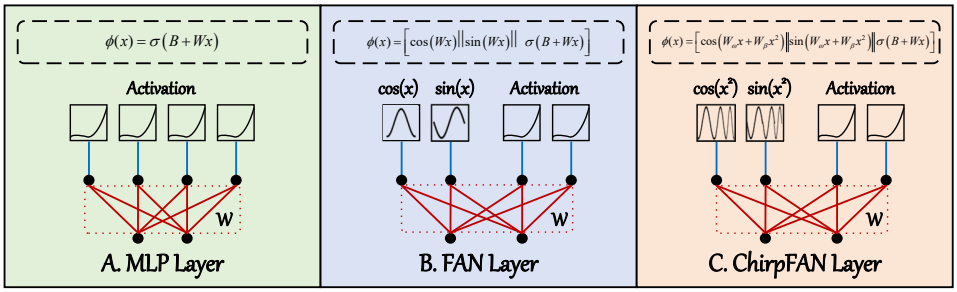


Fig. 2. The proposed Chirp-FAN layer compared with MLP and FAN.

Thus, Chirplets are designed for local adaptation by explicitly parameterizing various geometric transformations for better localized feature representation. To address the interference caused by multi-sensor imaging, frequency-domain methods such as the Fourier transform and the fractional transforms are utilized to capture image structures. By utilizing amplitude and phase information, the multi-domain features can discriminate land covers while suppressing noise and finally provide enhanced features for the classification of MSRS data.

The eigen-decomposition definition of the fractional Fourier transform (FrFT) [50] for a signal $f(x)$ is given by

$$\mathcal{F}_{\alpha}[f](u)=\sum_{n=0}^{\infty}c_n e^{-in\alpha}\psi_n(u), \tag{1}$$

where $c_n=\int_{-\infty}^{\infty}f(x)\psi_n^*(x)dx$, α represents the rotation angle, $\psi_n(x)$ are the Hermite-Gauss functions, which form an orthogonal basis.

The FrFT can also be expressed in an integral kernel form

$$\mathcal{F}_{\alpha}[f](u)=\int_{-\infty}^{\infty}K_{\alpha}(u,x)f(x)dx, \tag{2}$$

where the kernel function is

$$K_{\alpha}(u,x)=\sqrt{\frac{1-i\cot\alpha}{2\pi}}\exp\left(i\pi\frac{(u^2+x^2)\cos\alpha-2ux}{\sin\alpha}\right). \tag{3}$$

For a signal $f(x)$, there exists an optimal order α_{opt} that maximizes the energy concentration in the FrFT domain

$$\alpha_{opt}=\arg\max_{\alpha}\int_{\Omega_k}|\mathcal{F}_{\alpha}[f](u)|^2du, \tag{4}$$

where Ω_k represents the frequency band region containing the top $k\%$ of the energy. This theorem guarantees the convergence of learnable parameters.

C. Fourier Analysis Networks

Multi-layer perceptron (MLP) is a foundational neural network architecture [51]. As shown in Fig. 2-A, an MLP layer $\Phi(x)$ is defined as $\Phi(x)=\sigma(B_m+W_mx)$, where σ is an activation function, B_m and W_m are learnable parameters (bias and weights). MLPs can approximate functions and serve as fundamental components in deep learning models. However, MLPs struggle with understanding periodicity signals.

Focusing on this problem, Fourier analysis network (FAN) was proposed to efficiently model and reason about periodic signals [49]. As shown in Fig. 2-B, an FAN layer $\phi(x)$ is defined as

$$\phi(x)=[cos(W_px)||sin(W_px)||\sigma(B_{\overline{p}}+W_{\overline{p}}x)], \tag{5}$$

where W_p, $W_{\overline{p}}$, and $B_{\overline{p}}$ are learnable parameters. The design decouples the frequency (W_p) and coefficient ($W_{\overline{p}}, B_{\overline{p}}$) as learning components. FAN integrates periodicity into its structure by using Fourier Series. It model the data periodicity rather than just memorizing patterns, which serves as a promising substitute for MLP. Compared with MLP, FAN generally requires fewer parameters. However, FAN relys on fixed frequencies $W_{\overline{p}}$, which limits it to model signals whose frequencies change over time/space, e.g., domain shifts in MSRS data as Fig. 1 shows. Aiming at addressing the above problems, we propose Chirplet Fourier Analysis in Section III by introducing learnable chirp rates.

III. Proposed Chirplet Fourier Analysis

The Chirplet Fourier analysis (ChirpFA) is proposed to extract and analyze the multi-domain features. In this section, we define the ChirpFA layer, analyze its properties, and its connection to MSRS domain shifts.

A. Definition

For a square-integrable function $f(x)\in L^2(\mathbb{R})$, it can be represented as a linear combination of basis functions. Chirplet transforms provide such basis where functions are localized in time-frequency and can model frequency modulation [52]. For a local signal, its representation can be approximated by

$$f(x)\approx\textstyle\sum_k(\alpha_k cos(2\pi\omega_k x+\pi\beta_k x^2)+\gamma_k sin(2\pi\omega_k x+\pi\beta_k x^2)), \tag{6}$$

where α_k and γ_k are the linear coefficients of k-th Chirplet component. ω_k is the center frequency while β_k is the chirp rate.

Within a windowed attention framework, local image features can be interpreted as 1D signals along a spatial dimension x. Chirp-FAN employs chirp basis functions to model the space-frequency features of these local signals

$$\phi_{chirp}(x)=\cos(2\pi\omega x+\pi\beta x^2)||\sin(2\pi\omega x+\pi\beta x^2), \tag{7}$$

where ω represents the initial frequency and β represents the chirp rate. To simplify the notation for derivations, we introduce the angular frequency $\omega_k'=2\pi\omega_k$ and the angular chirp rate $\beta_k'=\pi\beta_k$. Then, ω and β are used to denote these simplified angular parameters.

Firstly, we consider a network using input-output pairs $\{x_i,y_i\}$ and aims to identify function $f(x):\mathbb{R}^{d_x}\rightarrow\mathbb{R}^{d_y}$. A simple neural network $f_C(x)$ representing the Chirplet expansion is:

$$
\begin{aligned}
& f_C(x) \\
& \triangleq B + \sum_{k=1}^{K} (\alpha_k cos(\omega_k x + \beta_k x^2) + \gamma_k sin(\omega_k x + \beta_k x^2)) \\
& = B + [\alpha_1, \ldots, \alpha_N] cos([\omega_1 || \ldots || \omega_N] x + [\beta_1 || \ldots || \beta_N] x^2) \\
& + [\gamma_1, \ldots, \gamma_N] sin([\omega_1 || \ldots || \omega_N] x + [\beta_1 || \ldots || \beta_N] x^2) \\
& = B + W_\alpha cos(W_\omega x + W_\beta x^2) + W_\gamma sin(W_\omega x + W_\beta x^2) \\
& = B + W_{out}[cos(W_\omega x + W_\beta x^2) || sin(W_\omega x + W_\beta x^2)],
\end{aligned} \tag{8}
$$

where $B \in \mathbb{R}^{d_y}$, $W_\omega, W_\beta \in \mathbb{R}^{K \times d_x}$ representing the learnable angular frequencies and chirp rates, and $W_{out} \in \mathbb{R}^{d_y \times 2K}$ are learnable parameters.

The ChirpFAN layer is designed as $\phi_{chirp}(x)$ by first decoupling $f_C(x)$ into $f_{in}(x) = [cos(W_\omega x + W_\beta x^2) || sin(W_\omega x + W_\beta x^2)]$ and $f_{out}(x) = B + W_{out} x$. As shown in Fig. 2-C, the ChirpFAN layer combines this basis as

$$
\phi_{chirp}(x) \triangleq [cos(W_\omega x + W_\beta x^2) || sin(W_\omega x + W_\beta x^2) || \sigma(B_{\overline{p}} + W_{\overline{p}} x)], \tag{9}
$$

where $W_\omega \in \mathbb{R}^{d_p \times d_x}$, $W_\beta \in \mathbb{R}^{d_p \times d_x}$, $W_{\overline{p}} \in \mathbb{R}^{d_{\overline{p}} \times d_x}$, and $B_{\overline{p}} \in \mathbb{R}^{d_{\overline{p}}}$ are learnable parameters.

B. Properties

1) Property: For a 1D signal coordinate x, the instantaneous frequency associated with the Chirp phase $\Phi(x) = 2\pi\omega x + \pi\beta x^2$ is

$$
f_{inst}(x) = \frac{1}{2\pi} \frac{d\Phi(x)}{dx} = \omega + \beta x. \tag{10}
$$

The signal energy concentrates along this linear track in the spatial-frequency plane. The Wigner-Ville distribution (WVD) [53] confirms this concentration

$$
WV(x, f) = \delta(f - (\omega + \beta x)). \tag{11}
$$

In contrast, Fourier Analysis Network (FAN) basis functions use fixed frequencies ω, with its WVD concentrated at a single frequency line $WV(x, f) = \delta(f - \omega)$. Chirp-FAN's ability to model varying frequencies provides enhanced resolution and can represent detailed textures such as edges and Chirp blurs.

Theorem 1: The gradient of a loss function $\mathcal{L}$ with respect to a chirp rate β_k is

$$
\frac{\partial \mathcal{L}}{\partial \beta} = -2x^2 \sum_k (\alpha_k \sin(\theta_k) - \gamma_k \cos(\theta_k)), \tag{12}
$$

where $\theta_k = 2\pi\omega_k x + \pi\beta_k x^2$. The presence of the x^2 term indicates that the gradient magnitude is sensitive to x. The x^2 term increases the magnitudes $|x|$ and the Chirp rate gradients, which can accelerate the learning of edges or high-frequency noise. For inputs with small $|x|$, this term can suppresses gradient noise and results in training robustness.

2) Connection to MSRS Challenges: The properties of ChirpFAN are directly relevant to addressing domain shifts in MSRS data. MSRS data often exhibit non-stationarities such as shifts or broadening in spectral signatures due to atmospheric effects or material variations. Similarly, spatial textures vary because of geometric distortions or changes in resolution. ChirpFAN can model varying frequencies, which allows it to adaptively capture these linearly frequency-modulated components. This property of ChirpFAN provides a more accurate representation than the fixed-frequency bases, such as FAN and wavelets. Furthermore, as Theorem 1 shows, the gradients of ChirpFAN are sensitive to input magnitude via the x^2 term, which benefits learning from complex MSRS data. Because edges, textural boundaries, or sharp spectral features often show larger values in processed feature maps, the amplified gradients in these regions can accelerate learning of these discriminative features.

Theoretically, the x^2 term could lead to gradient instability with large input values. To solve this, the standard Layer Normalization is applied to mitigate the gradient instability in our ChirpST architecture. The normalization bounds the input x and ensures training stability. Then, the x^2 term is a data-driven and adaptive learning rate for the chirp parameter.

While the underlying physical processes may involve higher-order non-linearities, the Chirplet transform provides a powerful first-order approximation. Distinguished from fixed-frequency or purely data-driven models, the following experimental results suggest that the Chirplet linear model is effective to capture the main signal-level distortions.

IV. Proposed ChirpFAN Framework

In this paper, a Chirplet Fourier analysis network (ChirpFAN) is proposed for cross-scene classification of MSRS data. Its overall flowchart is shown in Fig. 3. The proposed ChirpFAN consists of 1) a fractional spatial-frequency-phase (FrSFP) feature extraction part to extract multi-domain features in Section IV-A, and 2) a ChirpFAN based attention module is designed for multisource data analysis in Section IV-B. Then, an End-to-End network for adaptation is designed in Section IV-C. Finally, the classifier, discriminators, and the training procedures are introduced in Section IV-D.

The labeled source domain (SD) data is denoted as $\mathcal{S} = \{(\mathcal{X}^s_{\mathcal{H}}, \mathcal{X}^s_{\mathcal{L}}), Y^s\}$, and the unlabeled target domain (TD) data is denoted as $\mathcal{T} = \{(\mathcal{X}^t_{\mathcal{H}}, \mathcal{X}^t_{\mathcal{L}})\}$. Here, $\mathcal{X}_{\mathcal{H}} \in \mathbb{R}^{C_h \times W \times H}$ represents the HSI data, and $\mathcal{X}_{\mathcal{L}} \in \mathbb{R}^{C_l \times W \times H}$ represents the LiDAR data. H and W are the spatial dimensions, and C_h and C_l are the number of channels. Y^s is the source label (containing N classes), while the target domain label space Y^t is unknown. The superscripts s and t denote the source and target domains, respectively.

A. Fractional Spatial-Frequency-Phase Feature Extraction

To address the redundancy and interference introduced during multi-sensor imaging, frequency-domain methods, represented by the Fourier transform and fractional Fourier transform (FrFT), are utilized. They can remove interference in the transform domain and capture image structures (such as edges, textures, shapes) and variational features shown in Fig. 1. Inspired by this, we designed a FrSFP block as part A of Fig. 3 shows. This block acts as a transform-domain filter and feature pre-extractor. By jointly using magnitude and phase, it enhances the discrimination of land covers while suppressing noise, providing augmented features for cross-domain analysis.

Firstly, a 2D convolution (Conv) block is used to align the channel dimensions to $C_0 \times W \times H$

$$
\mathcal{X}^s_{\mathcal{H}_0} = f^{C_0 \to C_0}_{2D3}(f^{C_h \to C_0}_{2D3}(\mathcal{X}^s_{\mathcal{H}})), \mathcal{X}^s_{\mathcal{L}_0} = f^{C_l \to C_0}_{2D3}(\mathcal{X}^s_{\mathcal{L}}), \tag{13}
$$

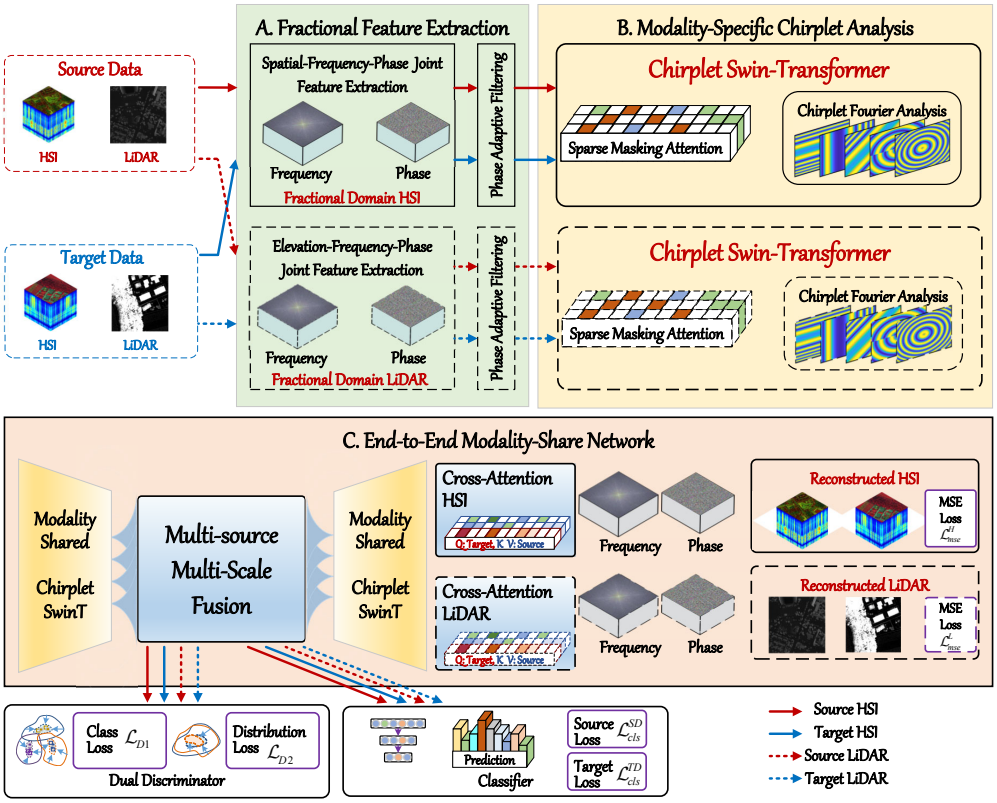


Fig. 3. The flowchart of proposed Chirp-FAN framework. It contains four parts: 1) a fractional spatial-frequency-phase (FrSFP) feature extraction part, 2) a ChirpFAN based attention module, 3) an End-to-End network for adaptation, and 4) the classifier and discriminators.

where $f_{2D3}^{*\to C_0}(\cdot)$ represents a 2D convolution block (Conv+BN+ReLU) with 3×3 window size, C_0 output channels, and a stride of 1.

Then, the FrFT is applied to the preprocessed signals $(\mathcal{X}^s_{\mathcal{H}_0}, \mathcal{X}^s_{\mathcal{L}_0})$, which transforms them into fractional domain. For the HSI data $\mathcal{X}^s_{\mathcal{H}_0} \in \mathbb{R}^{C_0\times H\times W}$, the complex-domain features after applying FrFT are

$$X_\alpha = \mathcal{F}_\alpha(\mathcal{X}^s_{\mathcal{H}_0}) = M_\alpha \odot e^{j\phi_\alpha}, \tag{14}$$

where $\mathcal{F}_\alpha$ represents the 2D FrFT operator of order α, applied independently to each spatial slice $(H \times W)$ for every channel and batch element. $M_\alpha = |X_\alpha|$ is the element-wise magnitude spectrum. $\phi_\alpha = \arg(X_\alpha)$ is the phase spectrum with the principal value in $(-\pi, \pi]$. $\odot$ denotes the Hadamard product.

Then a phase-aware weight module is designed for phase feature $\phi_\alpha \in \mathbb{R}^{C_0\times H\times W}$, which is usually discarded in multi-domain analysis. Phase spectrum reflects the structural and positional details of MSRS data. Let f_{PW1} (output channels C_0/k) and f_{PW2} (output channels $2C_0$) denote these layers

$$\begin{aligned} Z &= \mathrm{GELU}(f_{PW1}(\phi_\alpha)) \quad \in \mathbb{R}^{(C_0/k)\times H\times W}, \\ \phi_{out} &= \sigma(f_{PW2}(Z)) \quad \in \mathbb{R}^{2C_0\times H\times W}. \end{aligned} \tag{15}$$

The output ϕ_{out} is then split along the channel dimension into two tensors, ϕ_1 and ϕ_2, each of size $\mathbb{R}^{C_0\times H\times W}$. GELU is the activation function [54].

For the magnitude spectrum, two Conv blocks $T_1 = \mathcal{C}_{5\times5}(M_\alpha)$ and $T_2 = \mathcal{C}_{7\times7}(T_1)$ are applied, where $\mathcal{C}_{k\times k}$ denotes convolution operations with $k \times k$ kernels. The multi-domain phase-aware feature $\mathcal{X}^\phi_{\mathcal{H}} \in \mathbb{R}^{C_0\times H\times W}$ of HSI is

$$\mathcal{X}^\phi_{\mathcal{H}} = \phi_1 \odot T_1 + \phi_2 \odot T_2. \tag{16}$$

Under the conditions maximizing the variance of $\mathcal{X}^\phi_{\mathcal{H}}$ component-wise, the optimal weights are related to the power of the features

$$\phi_1^* \propto \|T_1\|_F^2, \quad \phi_2^* \propto \|T_2\|_F^2, \tag{17}$$

where $\|\cdot\|_F^2$ denotes the squared Frobenius norm (sum of squares of all elements). Then, normalizing operation is

$$\phi_1^* = \frac{\|T_1\|_F^2}{\|T_1\|_F^2+\|T_2\|_F^2}, \phi_2^* = \frac{\|T_2\|_F^2}{\|T_1\|_F^2+\|T_2\|_F^2}. \tag{18}$$

Similarly, the enhanced features for the LiDAR modality, derived through the same process applied to $\mathcal{X}^s_{\mathcal{L}}$, are denoted as $\mathcal{X}^\phi_{\mathcal{L}} \in \mathbb{R}^{C_0\times W\times H}$.

Theorem 2: If the phase weight generator sub-network (from ϕ_α to ϕ_1, ϕ_2) has a bounded Lipschitz constant, specifically if the operator norm of the linear layers and activations satisfy appropriate bounds (summarized as a constant L for the map $\phi_\alpha \mapsto (\phi_1, \phi_2)$), then the change in the output feature T is bounded with respect to changes in the input phase $\Delta\phi$

$$\|T(\phi_\alpha) - T(\phi_\alpha + \Delta\phi)\|_F \le K(\|\Delta\phi\|_F), \tag{19}$$

where the bound K depends on the Lipschitz constant L of the weight generator and on bounds for the feature norms $\|T_1\|_F, \|T_2\|_F$.

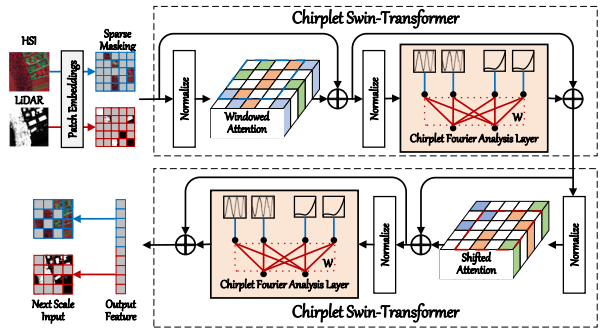


Fig. 4. The proposed Chirplet Swin-Transformer.

This indicates that small perturbations in the phase input lead to controlled changes in the fused output feature, ensuring stability. The stability follows from the Lipschitz properties of the constituent operations and the bounded nature of the features T_1, T_2 being modulated.

In this section, the proposed block provide 1) an adaptive fractional order parameters α enabling frequency feature selection, and 2) a phase-aware weighting module for optimal multi-domain feature fusion.

B. Chirplet Fourier Swin-Transformer

Using the multi-domain amplitude-phase features extracted in Section IV-A, the enhanced features are separated into blocks via patch embedding. For each input block, a linear embedding layer projects the channel dimension from C_0 to C_1. Then, a window masking step is used in the multi-modal masked learning step. With a masking ratio $\theta = 0.75$, it generates input features comprising only this subset of visible blocks, which enables the model to learn more representative multi-modal features from the MSRS data. For the input HSI features $\mathcal{X}_{\mathcal{H}}^{\phi} \in \mathbb{R}^{C_0 \times W \times H}$, the unmasked blocks retained after masking are denoted as $\mathcal{M}(\mathcal{X}_{\mathcal{H}}^{\phi}) \in \mathbb{R}^{C_0 \times (1-\theta) WH}$.

With the Chirp-FAN layer proposed in Section III, the Chirplet swin transformer (ChirpST) is designed to extract multi-modal features from the masked features. As illustrated in Fig. 4, two swin transformer (ST) [55] blocks are utilized to learn intra-modal amplitude and phase information.

The encoding process can be described as

$$\mathcal{X}_{\mathcal{H}_1}^{s} = \text{ChirpST}(\mathcal{M}(\mathcal{X}_{\mathcal{H}}^{\phi})), \tag{20}$$

where the proposed Chirp-FAN layer in ChirpST block performs time-varying frequency domain decomposition on local window features using the Chirplet basis $\psi_{\beta,\omega}(t) = e^{j2\pi(\omega t + \frac{\beta}{2}t^2)}$. According to Theorem 1, Chirp-FAN can approximate complex patterns with fewer basis functions,

$$||\mathcal{X}_{\mathcal{H}}^{\phi} - \sum_{k=1}^{M} c_k \psi_{\beta_k,\omega_k}(t)||_2 < \epsilon. \tag{21}$$

Compared to the non-linear mapping of MLPs, ChirpFAN adaptively matches the inter-spectral variations of HSI data. It enhances the sparsity of the feature representation through the Chirp-rate β_k. While the windowed self-attention represents local spatial feature, the Chirp-FAN models local non-stationary patterns and enhances local-global feature interaction. The dynamic chirp rate β_k allows Chirp-FAN to adaptively represent local frequency modulations. This adaptive filtering leads to a more compact representation where signal energy is Chirp-related. The energy concentration into fewer and more descriptive coefficients contributes to an enhanced sparsity in the feature space, which benefits the following processing. Moreover, as for the LiDAR modality, the encoded features after ChirpST can be denoted as $\mathcal{X}_{\mathcal{L}_1}^{s}$.

The spectra shifts of HSIs can be modeled as non-stationary signals, which can be modeled by the proposed ChirpFAN through its quadratic phase term. This outperforms FAN using fixed frequencies and MLPs employing non-linear transforms. In classification tasks, ChirpFAN's chirplet basis functions can generalize texturally rotated and scaled patterns while MLP and FAN lead to overfitting because of fixed frequencies or purely data-driven transforms. Specifically, ChirpFAN provides superior modeling of non-stationary characteristics or shifts in spectral signatures for HSI classification.

In our ChirpST, the ChirpFA layer is an in-place replacement of the standard MLP layer to impose a signal-aware inductive bias. By replacing the MLP block, the network can learn the Chirplet domain representations focusing on the signal modulations. In contrast, a parallel-block design usually weight the general-purpose MLP path, which omits the effect of Chirplet analysis in processing non-stationary spectral and spatial modulations.

C. End-to-end Multisource Network

As shown in part C of Fig. 3, a modality-shared end-to-end network is designed for inter-modal relationship. Feature consistency is enhanced through cross-modal interaction using the encoded HSI and LiDAR feature in last section $\mathcal{X}_{\mathcal{H}_1}^{s}$ and $\mathcal{X}_{\mathcal{L}_1}^{s}$. The overall shared encoding process is defined as

$$[\mathcal{X}_{\mathcal{H}_2}^{s}, \mathcal{X}_{\mathcal{L}_2}^{s}] = \text{SwinT}(\text{ChirpFAN}([\mathcal{X}_{\mathcal{H}_1}^{s}, \mathcal{X}_{\mathcal{L}_1}^{s}])). \tag{22}$$

As shown in Eq. (22), the ChirpFAN layers are applied to decompose and align the MSRS features in fractional domains. The cross-modal attention adjusts weights based on query-key similarities between modalities. It allows the proposed framework to learn shared representations related to frequency components and suppress inter-modal interference.

In this stage, the ChirpFAN enabled the frequency domain alignment. Through multi-scale decomposition using Chirp basis functions, the spectral oscillations of HSIs and the geometric structures of LiDAR data are mapped into a unified frequency domain space. This suppresses inter-modal interference. At the same time, the Chirp rates are dynamically adjusted by cross-modal attention weights, achieving frequency-adaptive fusion with the fusion weight

$$w_a = \sigma(\text{Softmax}(QK^T/\sqrt{d})), \tag{23}$$

where Q and K represent the cross-modal query and key vectors, respectively, forming a frequency-space dual attention. The encoded source domain (SD) multi-modal and multi-scale

features are $\mathcal{X}^s_{\mathcal{H}_1}, \mathcal{X}^s_{\mathcal{H}_2}$ and $\mathcal{X}^s_{\mathcal{L}_1}, \mathcal{X}^s_{\mathcal{L}_2}$. The target domain (TD) features are denoted as $\mathcal{X}^t_{\mathcal{H}_1}, \mathcal{X}^t_{\mathcal{H}_2}$ and $\mathcal{X}^t_{\mathcal{L}_1}, \mathcal{X}^t_{\mathcal{L}_2}$.

As shown in part C of Fig. 3, the decoder comprises a shared decoder and two modality-specific decoders for cross-modal reconstruction. The shared decoder $f_{\text{sha-de}}(\cdot)$ uses two ChirpST blocks and two upsampling operations to restore spatial resolution as

$$\mathcal{X}^{s'}_{\mathcal{H}_1} = f_{\text{sha-de}}(\mathcal{X}^s_{\mathcal{H}_2}), \quad \mathcal{X}^{s'}_{\mathcal{L}_1} = f_{\text{sha-de}}(\mathcal{X}^s_{\mathcal{L}_2}), \tag{24}$$

where $\mathcal{X}^s_{\mathcal{H}_2} \in \mathbb{R}^{C_3 \times W/4 \times H/4}$ is the encoder output.

Then, modality-specific decoders are designed to utilize multi-modal complementarity for cross-modal reconstruction. Taking the reconstruction of HSI multi-domain features as an example, it contain self-attention (SA) and cross-attention (CA) interaction. For SA, the reconstructed HSI modal features $\mathcal{X}^{s'}_{\mathcal{H}_1}$ are projected into query q_h, key k_h, and value v_h vectors. This module learns intra-modal frequency domain priors and obtain the self-attention map A_{sa}, which enhances the capability to model spectral continuity. For CA, the concatenated features $[\mathcal{X}^{s'}_{\mathcal{H}_1}, \mathcal{X}^{s'}_{\mathcal{L}_1}]$ are mapped to form inter-modal keys to integrate multi-modal complementarity for precise HSI reconstruction. The geometric features from LiDAR are mapped via frequency modulation projection to form cross-modal keys k_l. The frequency domain matching degree between these keys and the HSI query q_h determines the cross-modal fusion attention A_{ca}.

By fusing A_{sa} and A_{ca}, and then multiply them with the value v_h, the HSI modality $\mathcal{X}^{s'}_{\mathcal{H}_1}$ is reconstructed as

$$\mathcal{X}^{s'}_{\mathcal{H}_1} = f_o((A_{sa} \oplus A_{ca}) \otimes v_h). \tag{25}$$

Then, using the multi-domain features, the reconstructed data $\mathcal{X}^{s'}_{\mathcal{H}_0}$ can be represented as:

$$\mathcal{X}^{s'}_{\mathcal{H}_0} = \mathcal{F}^{-1}_{\alpha}(\sum_{\beta} w_\beta \cdot \mathcal{F}_{\alpha}(\mathcal{X}^{s'}_{\mathcal{H}_1}) \odot \mathcal{F}^{-1}_{\alpha}(\mathcal{X}^{v'}_{\mathcal{L}_1})), \tag{26}$$

where w_β are generated by the frequency weighting layer and implement frequency domain cross-modulation, $\mathcal{X}^{v'}_{\mathcal{L}_1}$ is the LiDAR value. In the decoder, the ChirpST blocks reconstruct detailed features while suppressing aliasing noise. The frequency modulation projection used for LiDAR keys k_l enables matching with HSI queries q_h. Within a domain sensitive to textural details encoded by frequency variations, the cross-model attention improves the reconstructed $\mathcal{X}^{s'}_{\mathcal{H}_0}$. Similarity, the reconstructed LiDAR data can be denoted as $\mathcal{X}^{s'}_{\mathcal{L}_0}$.

Based on the above design, the time-frequency localization property enable the proposed ChirpFAN to suppress aliasing noise, and the reconstruction error satisfies

$$||\mathcal{X}^{s'}_{\mathcal{H}_0} - \mathcal{X}^s_{\mathcal{H}_0}||_2 \leq \lambda \cdot \text{TV}(\mathcal{X}^s_{\mathcal{H}_0}), \tag{27}$$

where TV represents the Total Variation regularizer, and λ is controlled by the Chirp rates.

Through this multi-modal mask learning, multi-modal features and their reconstructions can be extracted using only a small part of the visible patches. Introducing ChirpFAN into the masked auto-encoder architecture mathematically optimizes the multi-modal features representation. Further, it enhances the reconstruction accuracy and generalization of the proposed network for cross-scene classification of MSRS data.

D. Network Training

In this section, we design the network with the encoded SD multi-modal multi-scale features $\mathcal{F}^s = [\mathcal{X}^s_{\mathcal{H}_1}, \mathcal{X}^s_{\mathcal{L}_1}, \mathcal{X}^s_{\mathcal{H}_2}, \mathcal{X}^s_{\mathcal{L}_2}]$ and the TD features $\mathcal{F}^t = [\mathcal{X}^t_{\mathcal{H}_1}, \mathcal{X}^t_{\mathcal{L}_1}, \mathcal{X}^t_{\mathcal{H}_2}, \mathcal{X}^t_{\mathcal{L}_2}]$.

A multi-modal feature fusion and classifier module is constructed to map the features to class using deconvolutional and convolutional layers, and then predict the final SD classification map Y^s_p as

$$Y^s_p = C(\mathcal{F}^s) = f^{128 \to N}(f^{6C_f \to 128}(\mathcal{F}^s)), \tag{28}$$

where $f^{6C_0 \to 128}(\cdot)$ is a 2D deconvolutional layer with output channels 128 and stride 2, and $f^{128 \to N}(\cdot)$ is a convolutional layer with the output N being the number of land covers. Therefore, the classifier loss is $\mathcal{L}_C = \mathcal{L}_C(Y^s, Y^s_p)$.

Then, a domain adaptation module based on dual discriminators is designed focusing the domain shifts. As shown in Fig. 3, this module includes a local fine-grained class discriminator D_1 and an entropy-constrained global discriminator D_2.

The D_1 focuses on class variations, such as the tree class that varies across seasons. Instead of traditional 0-1 domain encoding, multi-channel soft labels are used in D_1. The SD domain encoding is $[0; E^s]$ and the TD domain encoding is $[E^t; 0]$. Then, the SD per-channel calculation is $E_n = \frac{e^{(Y_j)^{N/T}}}{\sum_{j=1}^{N} e^{(Y_j)/T}}$, where the coefficient $T = 1.8$ is used to smooth the probability distribution. E_n is then thresholded with confidence threshold $\theta = 0.9$. Thus, the discriminator D_1 loss $\mathcal{L}_{dis1}$ and adversarial loss $\mathcal{L}_{adv1}$ are defined as

$$\mathcal{L}_{dis1} = -[E^s \log D_1(F^s) + E^t \log(1 - D_1(F^t))], \tag{29}$$

$$\mathcal{L}_{adv1} = -E^t \log(D_1(F^t), 0). \tag{30}$$

The global entropy-constrained discriminator D_2 is used to prevent model overfitting and improve generalization. The target is to make the entropy map of SD and TD similar, achieving cross-domain global structure consistency. Shannon entropy $H(\cdot)$ is calculated based on pixel-level prediction results (Y^s_p, Y^t_p), and then an adversarial learning strategy is used for entropy-level domain adaptation. The discriminator loss of D_2 is

$$\mathcal{L}_{dis2} = \mathcal{L}_{bce}(D_2(Y^s_p), 0) + \mathcal{L}_{bce}(D_2(Y^t_p), 1), \tag{31}$$

and the adversarial loss is

$$\mathcal{L}_{adv2} = \mathcal{L}_{bce}(D_2(H(Y^t_p)), 0), \tag{32}$$

where $\mathcal{L}_{bce}(\cdot)$ being the binary cross-entropy loss function.

Considering the dual discriminators, the total discriminator loss $\mathcal{L}_{dis}$ is

$$\mathcal{L}_{dis} = \mathcal{L}_{dis1} + \lambda_1 \mathcal{L}_{dis2}, \tag{33}$$

and total adversarial loss $\mathcal{L}_{adv}$ is

$$\mathcal{L}_{adv} = \mathcal{L}_{adv1} + \lambda_2 \mathcal{L}_{adv2}, \tag{34}$$

where the weighting factors are $\lambda_1 = \lambda_2 = 0.1$, which achieves finer-grained domain alignment and improves cross-domain classification accuracy.

The overall discriminator objective of proposed network is captured by $\mathcal{L}_D = \mathcal{L}_{dis} - \mathcal{L}_{adv}$. The data reconstruction

Algorithm 1 ChirpFAN classification framework.

Input: Source data $\{(\mathcal{X}^s_{\mathcal{H}}, \mathcal{X}^s_{\mathcal{L}})$, source labels $\mathbf{Y}^s$, target data $\{(\mathcal{X}^t_{\mathcal{H}}, \mathcal{X}^t_{\mathcal{L}})$.
Output: Best trained models G, D_1, D_2, Target labels $\mathbf{Y}^T$.
procedure TRAINING($args.epoch$)
    Data patch preparation and initialize all weights and bias terms.
    Initialize Generator G as Sec. IV-A–C and Discriminators D_1, D_2 as Sec. IV-D.
    for *ep* $= 0$ **to** *args.epoch* $- 1$ **do**
        Train Generator G Using SD data: $\mathcal{F}^s, \mathbf{P}^s, \mathcal{L}^s_{mse} \leftarrow G(\mathbf{X}^s, \text{'source'})$
        Train Generator G Using TD data: $\mathcal{F}^t, \mathbf{P}^t, \_ \leftarrow G(\mathbf{X}^t, \text{'target'})$
        Train Discriminators D_1, D_2.
    end for
end procedure
procedure TEST(G, D_1, D_2)
    $\mathbf{Y}_{\text{pred\_list}} \leftarrow [\,]$, $\mathbf{Y}_{\text{gt\_list}} \leftarrow [\,]$
    for all $(\mathbf{X}_{\text{b}}, \mathbf{Y}_{\text{gt\_b}})$ in *Dataloader* **do**
        $\_, \mathbf{P}_{\text{b}}, \_ \leftarrow G(\mathbf{X}_{\text{b}}, [D_1, D_2], \text{'target'})$
        $\mathbf{Y}_{\text{pred\_b}} \leftarrow \text{argmax}(\mathbf{P}_{\text{b}})$
        Append $\mathbf{Y}_{\text{pred\_b}}$ to $\mathbf{Y}_{\text{pred\_list}}$; $\mathbf{Y}_{\text{gt\_b}}$ to $\mathbf{Y}_{\text{gt\_list}}$
    end for
    Update the predictions $\mathbf{Y}^T, \mathbf{Y}_{\text{gt}}$
    return $\mathbf{Y}^T$
end procedure

loss is defined as $\mathcal{L}_G = \mathcal{L}_{mse}(\mathcal{X}_{\mathcal{H}}) + \mathcal{L}_{mse}(\mathcal{X}_{\mathcal{L}})$. The overall objective function is

$$\arg\min_{G,C} \arg\max_{D} (\mathcal{L}_G + \mathcal{L}_C - \mathcal{L}_D). \tag{35}$$

During the entire end-to-end training process, the 75% sparse masking is applied. This strategy is a regularizer to make the discriminators (D_1, D_2) and the classifier (C) operating on a sparse subset of visible patches. Then, the network is prevented from overfitting and can learn robust and context-aware representations, which improve the generalization capability for the unseen target domain. The detailed training and inference procedures of the proposed ChirpFAN are summarized in **Algorithm 1**.

V. Experimental Results and Comparison

A. Experimental Data

In this section, three cross-domain multimodal remote sensing (MSRS) datasets are utilized to evaluate the effectiveness of the proposed ChirpFAN.

1) The Houston13-Houston18 Scenes (H13-H18) The dataset consists of HSI and LiDAR data acquired over the University of Houston campus (USA) in 2013[1] and 2018[2], respectively. For cross-temporal analysis, the overlapping area was processed, involving resolution matching (down-sampling Houston18) and the extraction of 48 common HSI spectral bands across 7 shared land-cover classes.

2) The Nashua-Hanover Scenes (Na-Ha) Sourced from G-LiHT data[3], the dataset includes co-registered HSI and LiDAR data from Nashua (450×1050 pixels) and Hanover (700×620 pixels), USA, both at 1m Ground Sample Distance (GSD). The HSIs feature 114 spectral bands ($0.42 - 0.95\mu m$). There are 7 land-cover classes used for cross-city analysis.

3) The Houston-Trento Scenes (H13-T) The dataset combines the Houston13 HSI/LiDAR data with HSI/LiDAR data acquired over a rural area in Trento, Italy [56]. For cross-city analysis, 64 common HSI spectral bands were extracted across 4 shared land-cover classes.

B. Experimental Setup

The experimental programs are implemented using Python 3.12.0 and Pytorch 2.6.0 on a personal computer equipped with NVIDIA GeForce RTX 4060 Ti.

1) Algorithm Parameter Configuration: The proposed ChirpFAN uses two sets of parameters. The Adam optimizer is used in the training phase. No regularization term has been applied to the cost functions. The learning rate can be updated by multiplying the initial learning rate by $1 - \frac{epoch}{Epochs}$, where the weight-decay is 10^{-3} and the basic learning rate $lr = 5 \times 10^{-4}$ will be researched in the following subsection. The input batch size is 4, with the patch size being 64. In the fractional feature extraction part, $\frac{2\alpha}{\pi} = 0.4$ order FrFT is applied. In the ChirpST module, each attention block contains a $\theta = 75\%$ masked attention layer, a layer normalization, and a linear rate $g = 0.4$ of ChirpFAN activation layer. For the domain discriminator, the learning rate is $lr_D = 10^{-4}$. For the overall network, it consists of 4 modality-specific ChirpST blocks and 4 modality-shared ChirpST blocks. Each ChirpST block contains 8 attention heads.

2) Performance Analysis Configuration: To validate the proposed ChirpFAN's effectiveness quantitatively and qualitatively, classification performance on three cross-scene datasets is compared with that of other competitive classifiers. The compared state-of-the-art (SOTA) methods include fractional fusion and spatial-spectral domain adaptation (FrF-SSDA) [25], supervised contrastive learning-based unsupervised domain adaptation (SCLUDA) [32], high-resolution domain adaptation networks (HighDAN) [12], masked self-distillation domain adaptation (MSDA) [33], multilevel unsupervised domain adaptation (MLUDA) [31], and shared-private feature alignment semi-supervised learning (SASS) [28], as well as two time-frequency analysis methods: Wavelet-CNN [46] and Gabor-based CNN [48]. The experimental setups of the compared methods are optimized as suggested. All compared methods use original image patches as inputs without data augmentation. Three commonly used evaluation metrics are adopted, including overall accuracy (OA), average accuracy (AA), and Kappa coefficient (Kappa), to quantify the experimental results. To assess statistical significance, ten experiments are conducted, and the standard error is computed.

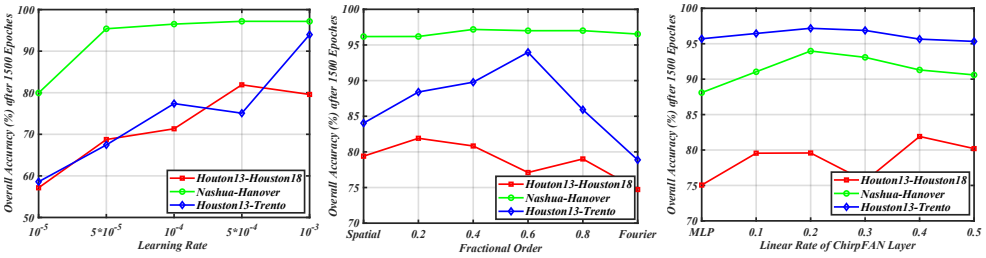
C. Parameter Analysis

To validate the effectiveness and sensitivity of parameters involved in the proposed ChirpFAN, experimental analysis using varying parameters is compared in Fig. 5.

[1] http://www.grss-ieee.org/community/technical-committees/data-fusion/2013-ieee-grss-data-fusion-contest/

[2] http://www.grss-ieee.org/community/technical-committees/data-fusion/2018-ieee-grss-data-fusion-contest/

[3] https://glihtdata.gsfc.nasa.gov



(a) (b) (c)

Fig. 5. Classification performance of the proposed ChirpFAN with different parameters. (a) Learning rate lr, (b) Fractional order $\frac{2\alpha}{\pi}$, (c) Linear rate of Chirplet layer.

1) Learning rate lr: In Fig. 5 (a), the OA after 1500 training epochs shows that the learning rate lr affects the convergence of the learning process. The search range is constrained from 10^{-5} to 10^{-3}. The algorithm using a small learning rate, such as 10^{-5}, cannot converge effectively and results in poor performance, with OA below 70% for all three datasets. The OA improves as the learning rate increases, but large learning rates like 10^{-3} can lead to large fluctuations in the objective function and result in a unstable training process. The optimal learning rates appear to be around 10^{-4} to 5×10^{-4}, where most datasets achieve their peak performance.

2) Fractional order $\frac{2\alpha}{\pi}$: Fig. 5 (b) illustrates the impact of the fractional order $\frac{2\alpha}{\pi}$ on the OA after 1500 epochs. The fractional order $\frac{2\alpha}{\pi}$ is varied from spatial domain $\frac{2\alpha}{\pi} = 0$ to the Fourier domain $\frac{2\alpha}{\pi} = 1$. For the scene sensitive to local spatial feature, e.g., H13-H18 dataset, the performance is better when using small $\frac{2\alpha}{\pi}$ with more spatial features in mixed features. The Na-Ha dataset shows a peak performance around $\frac{2\alpha}{\pi} = 0.4$. For the H13-T dataset focusing on global features, the OA shows a peak at $\frac{2\alpha}{\pi} = 0.6$ and decreases as $\frac{2\alpha}{\pi}$ approaches the Fourier domain. Generally, larger fractional orders indicate more global frequency features, while the optimal $\frac{2\alpha}{\pi}$ is decided by the resolutions of datasets. In the following researches, a learnable fractional order α in the FrFT module would enable the feature extraction module find the optimal analysis domain as Eq. (4) instead of relying on a fixed value.

3) Linear rate of Chirplet layer g: Fig. 5 (c) shows the effect of the linear rate g of the ChirpFAN layer. Larger g means more linear features are kept, while $g = 0$ means pure non-linear layer, and $g = 0.5$ means pure linear layer. The parameter g is varied from $g = 0$ to 0.5. For the H13-H18 dataset, the OA shows a peak at $g = 0.4$. The Na-Ha and H13-T datasets show the highest performance at $g = 0.2$, decreasing as g moves towards 0.5.

4) Analysis of Learned Chirplet Rate β_k: To analyze the learned Chirplet Rate β_k and validate its learning stability, we examined its distributions from the W_β weights after training. Table I shows these distributions in both the HSI and LiDAR modality-specific branches. The distributions of learned β_k parameters are stable and not collapsing to zero. Thus, the network is actively using the chirp rate to model non-stationary, frequency-modulated components. Combined with the gradient stability provided by the Layer Normalization, the proposed ChirpFA demonstrates stable learning behavior.

As Table I shows, the model learns different chirp rates for HSI and LiDAR, which suggests it learns signal-level modulations of different sensors. For the spectral chirps in HSI, the learned β_k is a positive up-chirp showing the spectral shifts. These spectral domain shifts are caused by atmospheric effects or illumination changes, which are direct forms of spectral frequency modulation. For the spatial chirps in LiDAR, the negative-biased β_k shows that the model is capturing spatial chirps related to geometric and perspective distortions. E.g., as the sensor's view angle changes between scenes, spatial textures appear compressed with high spatial frequency at one end and stretched with low spatial frequency at the other, which is represented as a classic down-chirp. The model learns specific β_k distributions for the spectral chirps in HSI and the spatial chirps in LiDAR, which can adapt to specific physical, signal-level modulations of sensors.

TABLE I
STATISTICAL ANALYSIS AND PHYSICAL PROPERTIES OF LEARNED β_k DISTRIBUTIONS

Feature	**HSI Branch**	**LiDAR Branch**
Learned Chirp	**Up-Chirp**	**Down-Chirp**
Mean (μ)	0.182	-0.179
Std. Dev. (σ)	0.051	0.063
Range [Min, Max]	$[0.02, 0.40]$	$[-0.25, -0.02]$

D. Ablation Analysis

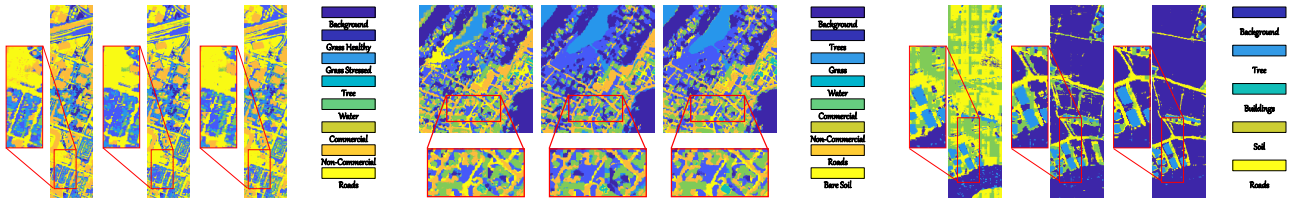
To measure the contributions of the FrFT feature extractor and ChirpFAN modules to the performance of the proposed ChirpFAN, a detailed ablation study is conducted. The Baseline models use only spatial-spectral features and use MLP layers for analysis. The proposed ChirpFAN with and without each component is denoted as ($\surd$) and ($\times$). Table II reports the experimental results for the three cross-scene MSRS datasets. For the sake of comparative fairness, no pretrained models or processes are provided for all experiments.

1) Analysis of the fractional feature extraction: In Part I of Table II, the feature extractor's effectiveness is enhanced when jointly using spatial, frequency, and phase features as input to the ChirpST. The combination yields the highest OA and Kappa scores across all datasets. Among these features, spatial features describe the basic geometric and structural properties, while frequency features and phase features align the spatial patterns. By integrating these varied perspectives, the ChirpFAN module receives a discriminative feature set and improves the classification results. To evaluate the effect

TABLE II
ABLATION STUDY OF CHIRPFAN ON THE THREE CROSS-SCENE MSRS DATASETS

	Modules									H13-18		Na-Ha		H13-T	
	Spatial	Frequency	Phase	FrFT	MLP	FAN	GaborST	WaveletST	ChirpFAN	OA	Kappa	OA	Kappa	OA	Kappa
I	√	×	×	×	×	×	×	×	√	76.40	64.84	90.18	90.12	84.04	71.33
	×	√	×	×	×	×	×	×	√	74.72	60.57	92.53	91.11	78.88	62.74
	√	√	√	×	×	×	×	×	√	78.83	66.6	93.81	91.01	89.42	79.82
II	√	√	√	√	√	×	×	×	×	75.05	61.47	91.41	88.02	88.10	77.98
	√	√	√	√	×	√	×	×	×	79.71	67.45	94.87	92.76	91.41	83.45
	√	√	√	√	×	×	√	×	×	80.56	69.12	95.14	93.15	91.87	84.21
	√	√	√	√	×	×	×	√	×	80.14	68.53	95.07	92.95	91.53	83.74
	√	√	√	√*	×	×	×	×	√	79.25	67.15	94.56	92.41	91.37	83.63
	√	√	√	√	×	×	×	×	√	**81.91**	**71.23**	**97.16**	**96.01**	**93.97**	**87.75**

*This row was conducted using the Simple Fusion: a concatenation of magnitude and phase.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

Fig. 6. The ablation analysis based on classification maps of different analysis layers. (a) MLP on H13-18 dataset, (b) FAN on H13-18 dataset (c) ChirpFAN on H13-18 dataset, (d) Legend of H13-18 dataset, (e) MLP on Na-Ha dataset, (f) FAN on Na-Ha dataset, (g) ChirpFAN on Na-Ha dataset, (h) Legend of Na-Ha dataset, (i) MLP on H13-T dataset, (j) FAN on H13-T dataset, (k) ChirpFAN on H13-T dataset, (l) Legend of H13-T dataset.

of phase-aware weighting module, a simple fusion variant is tested by replacing the module with a simple concatenation of magnitude and phase, followed by a 1×1 convolutional layer for fusion. As shown in Table II, the concatenation resulted in a $2-3\%$ drop in OA, which further proves the effectiveness of proposed phase-aware weighting module.

2) Analysis of the ChirpFAN layer: As listed in Part II of Table II, the proposed ChirpFAN analysis module is compared with traditional MLP and recently developed FAN. With the joint extraction of spatial-frequency-phase features, simply analyzing the non-linear spatial features or linear frequency features is highly susceptible to redundant and incomplete information stacking. Thus, the proposed ChirpFAN can analyze the extracted results in Part I and consistently yields the higher OA and Kappa scores than the other two single-domain methods across all three datasets. As shown in Fig. 6, we further compare MLP, FAN, and ChirpFAN by visualizing the classification maps. As expected, both the FAN and ChirpFAN combine frequency and phase features and obtain clear boundaries. Particularly, the proposed ChirpFAN layer can better represent the sharp edge region of patches, which improves the feature discrimination of mixed land covers. Compared with the baseline models, the proposed ChirpFAN network fuses linear transform and non-linear learning contexts to acquire both global and local dependencies. Compared with the single domain models, the proposed ChirpFAN model gains $6\% - 10\%$ improvements quantitatively.

To further validate the effectiveness of the proposed adaptive Chirplet layer, WaveletST and GaborST are used to replace the ChirpFA layer with 2D Haar Wavelets and 2D Gabor filters. As shown in Table II, WaveletST and GaborST outperform the standard MLP and FAN, but are outperformed by the proposed ChirpFAN. The learnable chirp rate enables the proposed method to model the complex frequency modulations of cross-scene domain shifts.

E. Classification Performance Analysis

To evaluate the effectiveness of proposed ChirpFAN, the classification performance of compared SOTA methods are shown in this section. The number of training samples are detailed in Tables III-V. The same training and testing samples are used for a fair comparison.

1) Performance on the H13-H18 Scenes: The qualitative results of competitive methods and the proposed ChirpFAN on the H13-H18 dataset are shown in Fig. 7 (f)-(l), with corresponding accuracies presented in Table III. From the classification maps, methods like FrF-SSDA with an OA of 66.69% show misclassification in distinguishing Grass Healthy, Grass Stressed, and Tree. Semi-supervised domain adaptation methods like SCLUDA (69.96% OA) and MLUDA (71.38% OA) also rely on training sample numbers and exhibit visual misclassifications. For instance, from the SCLUDA map in Fig. 7 (g), there are noticeable areas where Roads are incorrectly classified. Based on fractional multi-domain feature fusion and alignment, the proposed ChirpFAN can align source and target domains with joint spectral-spatial-frequency-phase features. By reconstructing multi-scale multi-domain features, the fused Chirplet features can enhance the classification of textural details and targets affected by environmental variations. The proposed ChirpFAN in Fig. 7 (l), with an OA of 81.91%, shows better delineation between different land cover types, especially in complex areas with mixed classes. For instance, the distinction between Commercial and Non-Commercial areas, and the accurate mapping of Roads. Quantitatively, as shown in Table III, the proposed ChirpFAN achieves the highest OA, outperforming all listed competitive

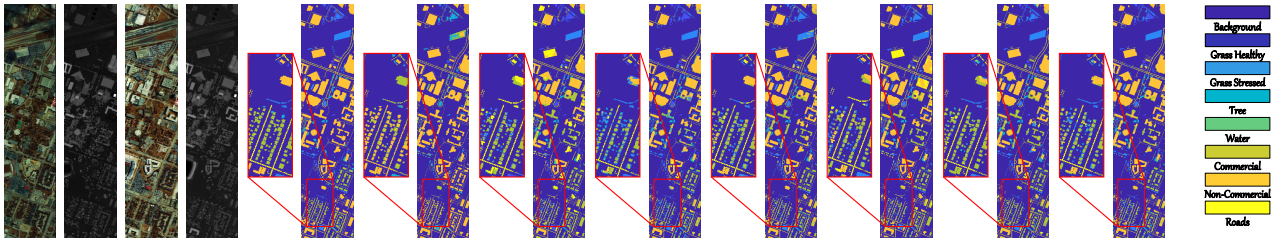


Fig. 7. The H13-H18 cross-domain classification maps. (a) Source HSI (Rband:25, Gband:14, Bband:6), (b) Source LiDAR, (c) Target HSI (Rband:25, Gband:14, Bband:6), (d) Target LiDAR, (e) Ground truth, (f) FrF-SSDA [25] (66.69%), (g) SCLUDA [32] (69.96%), (h) HighDAN [12] (75.59%), (i) MSDA [33] (77.66%), (j) MLUDA [31] (71.38%), (k) SASS [28] (78.20%), (l) ChirpFAN (81.91%).

TABLE III
COMPARISON OF THE CLASSIFICATION ACCURACY (%) USING THE H13-H18 SCENES.

Class	Training/All Houston 13	Test Houston 18	FrF-SSDA 2022 [25]	SCLUDA 2023 [32]	HighDAN 2023 [12]	MSDA 2024 [33]	MLUDA 2024 [31]	SASS 2024 [28]	Gabor-CNN 2021 [48]	Wavelet-CNN 2018 [46]	ChirpFAN
1. Grass Healthy	50/345	1353	53.22±1.54	87.65±0.88	81.01±1.12	4.29±0.50	65.27±2.10	73.45±1.30	70.15±1.61	72.33±1.45	76.94±0.95
2. Grass Stressed	50/365	4888	3.17±0.75	69.67±1.95	74.30±1.60	88.46±0.70	72.62±1.88	64.28±2.20	60.22±2.11	65.18±1.96	80.63±1.15
3. Trees	50/365	2007	28.85±3.10	78.18±1.35	91.13±0.65	81.27±1.25	59.31±3.50	89.91±0.80	65.13±3.04	71.06±2.53	93.87±0.45
4. Water	50/285	22	86.36±2.50	45.45±4.10	90.91±1.00	100.00±0.00	90.91±1.50	97.11±0.60	91.05±1.82	91.53±1.51	95.45±0.85
5. Commercial	50/319	5347	24.61±2.80	74.68±1.40	69.74±1.75	82.12±1.10	78.74±1.30	100.00±0.00	68.48±2.21	70.23±1.91	77.18±1.20
6. Non-commercial	50/408	32459	95.19±0.60	71.19±1.80	77.43±1.30	80.09±1.05	71.81±1.90	91.00±0.75	75.17±1.10	78.68±0.93	84.42±0.90
7. Roads	50/443	6565	24.74±3.20	53.94±2.50	66.00±2.10	67.57±1.90	67.41±2.30	66.10±2.00	55.37±2.52	61.52±2.14	71.28±1.50
OA			66.69 ± 1.10	69.96 ± 1.25	75.59 ± 0.90	77.66 ± 0.85	71.38 ± 1.40	78.20 ± 0.80	70.40 ± 1.30	73.15 ± 1.10	**81.91 ± 0.65**
AA			45.16 ± 1.80	68.68 ± 1.50	78.65 ± 1.15	71.97 ± 1.00	72.30 ± 1.70	**83.80 ± 0.95**	69.37 ± 1.70	72.93 ± 1.50	82.83 ± 0.70
Kappa×100			36.93 ± 2.00	55.45 ± 1.75	62.73 ± 1.35	64.95 ± 1.20	56.78 ± 1.95	61.62 ± 1.55	55.10 ± 2.10	58.20 ± 1.80	**71.23 ± 0.85**

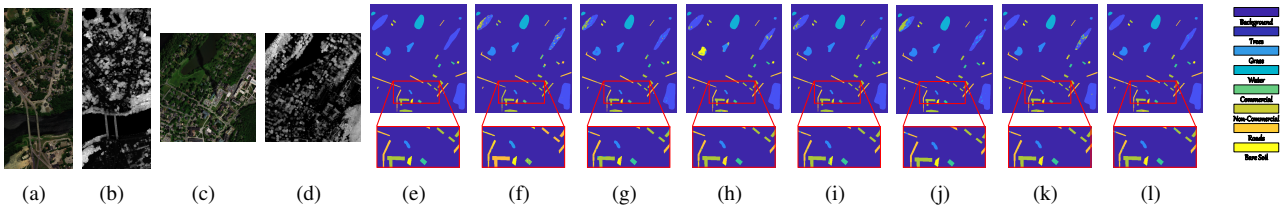


Fig. 8. The Nashua-Hanover cross-domain classification maps. (a) Source HSI (Rband:54, Gband:33, Bband:15), (b) Source LiDAR, (c) Target HSI (Rband:54, Gband:33, Bband:15), (d) Target LiDAR, (e) Ground truth, (f) FrF-SSDA [25] (83.31%), (g) SCLUDA [32] (92.48%), (h) HighDAN [12] (90.65%), (i) MSDA [33] (94.39%), (j) MLUDA [31] (90.24%), (k) SASS [28] (94.28%), (l) ChirpFAN (97.16%).

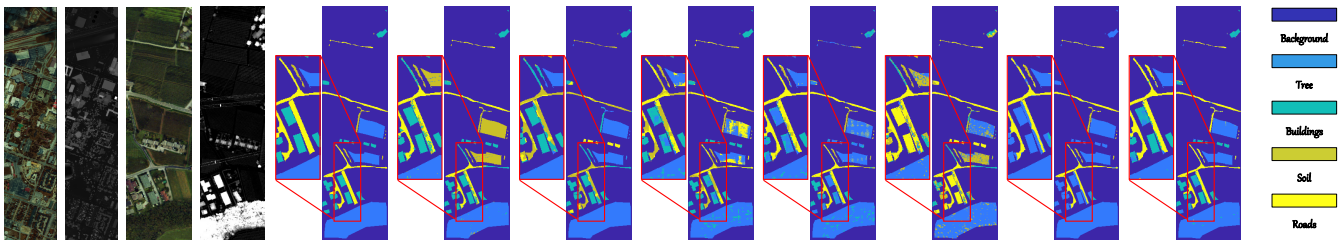
methods on the H13-H18 task. For land covers with similar spectral features, such as Commercial and Non-Commercial areas, the proposed ChirpFAN obtains accuracies of 97.11% and 91.00%, respectively. It is a significant improvement over methods like FrF-SSDA, which score 24.61% and 71.19%. The classification of these spectrally similar classes is challenging due to disparities between the source (Houston13) and target (Houston18) domains. Even for classes that are typically difficult, ChirpFAN demonstrates robust performance. For example, Grass Stressed is classified with 76.94% accuracy by ChirpFAN. The improved discrimination of Grass Healthy and Grass Stressed in Table III relies on the ChirpFAN's ability to model spectral modulations differentiating these classes. The better delineation of Commercial and Non-Commercial areas also benefits from ChirpFAN's capacity to capture spatially varying frequency patterns in both HSI and LiDAR data. The similar kinds of land covers often differ in textural complexity and structural layout, while ChirpFAN adapts to the geometric features even under domain shifts.

2) Performance on the Na-Ha Scenes: The qualitative results of competitive methods and the proposed ChirpFAN on the Na-Ha dataset are shown in Fig. 8 (f)-(l), with corresponding accuracies presented in Table IV. From the classification maps, methods like FrF-SSDA with an OA of 83.31% show some misclassifications, for example, in accurately delineating Bare Soil from other classes like Grass, leading to its 0.00% accuracy. Other methods like SCLUDA (92.48% OA) and HighDAN (90.65% OA) exhibit generally good performance but still show visual discrepancies. For instance, there is confusion between Trees and Commercial areas in certain parts of the HighDAN map. The proposed ChirpFAN in Fig. 8 (l) shows superior performance for the various land covers, whose visual representation is cleaner with more defined boundaries for classes like Roads and Water. Quantitatively, as shown in Table IV, the proposed ChirpFAN achieves the highest OA of 97.16%, outperforming competitive methods on the Na-Ha cross-city classification task. Notably, for the Bare Soil class, ChirpFAN achieves a significantly better accuracy of 83.77% than other SOTA methods. The consistently high performance across almost all classes proves the effectiveness of the proposed ChirpFAN.

3) Performance on the H13-T Scenes: The qualitative results of competitive methods and the proposed ChirpFAN on the H13-T dataset are shown in Fig. 9 (f)-(l), with corre-

TABLE IV
COMPARISON OF THE CLASSIFICATION ACCURACY (%) USING THE NA-HA SCENES.

Class	Training/All Nashua	Test Hanover	FrF-SSDA 2022 [25]	SCLUDA 2023 [32]	HighDAN 2023 [12]	MSDA 2024 [33]	MLUDA 2024 [31]	SASS 2024 [28]	Gabor-CNN 2021 [48]	Wavelet-CNN 2018 [46]	ChirpFAN
1. Trees	50/12713	18605	81.92±1.35	88.73±0.90	96.09±0.50	93.27±0.75	85.05±1.50	91.69±0.80	89.13±1.12	90.51±0.95	98.50±0.30
2. Grass	50/7528	4410	99.18±0.40	100.00±0.00	55.96±3.50	99.95±0.05	97.98±0.60	100.00±0.00	99.16±0.32	99.52±0.36	100.00±0.00
3. Water	50/16523	5425	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
4. Commercial	50/698	860	63.84±4.10	100.00±0.00	100.00±0.00	100.00±0.00	99.98±0.02	100.00±0.00	98.35±0.42	99.19±0.27	100.00±0.00
5. Non-commercial	50/3417	4338	93.64±0.85	92.52±1.10	93.71±0.95	97.65±0.45	98.35±0.40	89.46±1.25	90.75±1.04	91.84±0.98	93.15±0.70
6. Roads	50/17476	4955	66.52±3.20	92.04±1.30	91.52±1.40	88.27±1.60	92.33±1.20	97.68±0.65	85.67±1.52	88.74±1.25	92.11±1.00
7. Bare Soil	50/0821	764	0.00±0.00	97.11±0.70	69.50±2.80	71.47±2.50	95.44±0.90	92.65±1.15	71.06±2.83	75.01±1.86	83.77±1.80
OA			83.31±1.05	92.48±0.80	90.65±0.95	94.39±0.60	90.24±1.10	94.28±0.70	90.15 ± 0.90	91.82 ± 0.85	**97.16±0.40**
AA			72.13±1.90	95.77±0.55	86.68±1.25	92.94±0.85	95.30±0.75	**95.93±0.65**	90.58 ± 1.14	92.11 ± 0.96	95.36±0.50
Kappa×100			77.38±1.40	89.87±0.90	87.03±1.15	92.23±0.75	86.99±1.30	92.18±0.85	87.04 ± 1.20	89.13 ± 1.23	**96.01±0.45**



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

Fig. 9. The Houston13-Trento cross-domain classification maps. (a) Source HSI (Rband:25, Gband:14, Bband:6), (b) Source LiDAR, (c) Target HSI (Rband:25, Gband:14, Bband:6), (d) Target LiDAR, (e) Ground truth, (f) FrF-SSDA [25] (74.43%), (g) SCLUDA [32] (82.89%), (h) HighDAN [12] (79.39%), (i) MSDA [33] (84.53%), (j) MLUDA [31] (73.69%), (k) SASS [28] (89.97%), (l) ChirpFAN (93.97%).

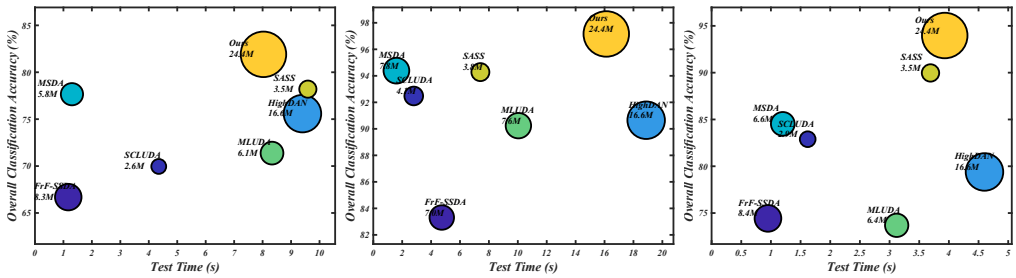
TABLE V
COMPARISON OF THE CLASSIFICATION ACCURACY (%) USING THE H13-T SCENES.

Class	Training/All Houston 13	Test Trento	FrF-SSDA 2022 [25]	SCLUDA 2023 [32]	HighDAN 2023 [12]	MSDA 2024 [33]	MLUDA 2024 [31]	SASS 2024 [28]	Gabor-CNN 2021 [48]	Wavelet-CNN 2018 [46]	ChirpFAN
1. Trees	50/365	9123	68.71±2.15	85.85±1.20	82.10±1.60	93.02±0.80	87.75±1.40	93.13±0.75	88.17±1.34	90.22±1.15	99.33±0.25
2. Buildings	50/319	2903	99.52±0.30	75.94±2.50	92.35±1.00	53.74±4.50	15.28±5.00	3.96±2.00	71.04±2.83	75.47±2.29	86.50±1.80
3. Ground	50/650	479	95.62±0.90	1.25±1.00	82.46±1.90	71.61±3.10	59.82±3.80	74.74±2.90	61.28±3.14	65.19±2.94	0.00±0.00
4. Roads	50/443	3174	72.02±2.80	15.41±3.50	55.86±4.00	79.46±2.20	70.97±3.00	78.85±2.40	71.31±2.52	72.16±2.24	92.75±1.10
OA			74.43±1.70	82.89±1.50	79.39±1.85	84.53±1.30	73.09±2.60	89.97±1.00	81.51 ± 1.94	84.22 ± 1.76	**93.97±0.60**
AA			**83.97±1.10**	44.61±2.80	78.19±2.05	74.45±2.30	58.45±3.15	62.67±2.55	72.95 ± 2.42	75.76 ± 2.16	69.65±1.30
Kappa×100			60.16±2.30	30.15±3.90	63.90±2.70	68.48±2.90	54.02±4.10	34.11±3.30	65.46 ± 2.82	68.85 ± 2.57	**87.75±1.50**

sponding accuracies presented in Table V. The target rural scene consists of Trees, Buildings, Ground, and Roads. From the classification maps of methods SCLUDA (82.89% OA) and HighDan (79.39% OA), there is visual confusion, such as Ground appears to be misclassified as Trees or Buildings in HighDAN's map in Fig. 9 (h). The proposed ChirpFAN in Fig. 9 (l) presents a classification map with an OA of 93.97%, which shows improved differentiation between classes like Buildings and Ground. Quantitatively, as shown in Table V, the proposed ChirpFAN achieves the highest OA on this H13-T cross-scene classification task. For instance, it achieves 99.33% accuracy for Trees and 86.50% for Buildings. In this dataset, the misclassification of the Ground class is obvious, e.g., FrF-SSDA (1.25%) and SASS (0.00%). Due to the spatial position, similar elevation and same composition of Ground and Roads in rural areas, these class is almost entirely misclassified as the classification map in Fig. 9(l) shows. Because these two classes in the rural Trento target scene share similar spectral and spatial-textural features, classification can fail even with robust domain adaptation methods when the inter-class variance is low.

F. Computational Analysis

To analyze the computational cost of the proposed ChirpFAN compared with SOTA networks, both the model sizes and inference time are analyzed with their performance on the three datasets used. As shown in Fig. 10, the parameter number (M) and the inference times (s) are used to evaluate the model sizes and the efficiency of methods. For all the compared methods, the same hardware and software configurations are used. As shown in Fig. 10, the model size of the proposed ChirpFAN is larger than the compared methods because of the linear matrix used in fractional frequency-phase joint extraction layers and initialization of Chirplet layers. Even though the testing time of the proposed ChirpFAN is acceptable because of the fast computations of linear transforms and stable filter design. Overall, the proposed ChirpFAN obtains the best performance with similar inference time. As shown in Table VI, while ChirpFAN is one of the largest models, its training and memory costs are comparable to other SOTA methods and its inference time remains practical.



(a) (b) (c)

Fig. 10. Computation analysis of the proposed ChirpFAN on different datasets. (a) H13-H18, (b) Na-Ha, (c) H13-T.

TABLE VI
COMPUTATIONAL COST ANALYSIS ON THE H13-H18 DATASET.

Method	Params (M)	Test (s)	Training (hrs)	GPU VRAM (GB)
FrF-SSDA	8.3	1.3	1.5	4.1
SCLUDA	2.6	4.4	2.0	5.3
HighDAN	16.6	9.6	3.5	8.2
MSDA	5.8	1.4	3.1	7.5
MLUDA	6.1	8.5	1.8	4.9
SASS	3.5	9.7	2.7	6.8
Gabor-CNN	19.2	7.6	4.2	6.8
Wavelet-CNN	18.8	7.9	4.1	6.5
ChirpFAN (Ours)	**24.4**	**8.2**	**4.1**	**9.3**

VI. CONCLUSION

Aiming to overcome challenges in cross-scene classification of MSRS data, particularly domain shifts and the comprehensive extraction of features from HSIs and LiDAR data, a Chirplet Fourier Analysis Network (ChirpFAN) was proposed in this paper. Firstly, a fractional feature extraction module is designed to jointly analyze spatial-elevation-frequency-phase features, providing robust representations. Secondly, a Chirplet Fourier analysis layer was integrated into a multi-scale Swin Transformer network (ChirpST), and then employed to extract and analyze texturally rotated and scaled patterns. Thirdly, these components were jointly utilized within an end-to-end network with domain adaptation stages. The network not only enables robust feature extraction from MSRS data but also aligns feature distributions. Finally, experiments on three cross-scene MSRS datasets were conducted and compared with competitive methods, demonstrating its effectiveness for MSRS classification tasks. However, as computational analysis indicated a larger model size compared to some methods, future efforts will focus on exploring model compression while preserving its high accuracy. Furthermore, the model's robustness under more extreme domain shifts, where the linear chirp assumption may no longer hold (e.g., transfers across highly dissimilar sensor modalities), remains a limitation and an important direction for future investigation. Finally, the ChirpFA layer itself shows significant promise as a general-purpose replacement for standard MLP blocks in architectures designed for non-stationary signals, such as in audio, seismology, or biomedical signal processing.

REFERENCES

[1] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Hofle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen, et al., "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 1, pp. 6–39, 2019.

[2] X. Zhu, F. Cai, J. Tian, and T. K. Williams, "Spatiotemporal fusion of multisource remote sensing data: Literature survey, taxonomy, principles, applications, and future directions," *Remote Sensing*, vol. 10, no. 4, pp. 527, 2018.

[3] M. Schmitt and X. Zhu, "Data fusion and remote sensing: An ever-growing relationship," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 4, pp. 6–23, 2016.

[4] R. Tao, X. Zhao, W. Li, H. Li, and Q. Du, "Hyperspectral anomaly detection by fractional fourier entropy," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 12, pp. 4920–4929, 2019.

[5] H. Li, W. Hu, W. Li, J. Li, Q. Du, and A. Plaza, "A3 clnn: Spatial, spectral and multiscale attention convlstm neural network for multisource remote sensing data classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 747–761, 2022.

[6] Y. Cai, Z. Zhang, X. Liu, Y. Ding, F. Li, and J. Tan, "Learning unified anchor graph for joint clustering of hyperspectral and lidar data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 4, pp. 6341–6354, 2025.

[7] M. Brell, K. Segl, L. Guanter, and B. Bookhagen, "Hyperspectral and lidar intensity data fusion: A framework for the rigorous correction of illumination, anisotropic effects, and cross calibration," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2799–2810, 2017.

[8] X. Zhao, R. Tao, W. Li, H. C. Li, Q. Du, W. Liao, and W. Philips, "Joint classification of hyperspectral and lidar data using hierarchical random walk and deep cnn architecture," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7355–7370, 2020.

[9] A. Zare and K. Ho, "Endmember variability in hyperspectral analysis: Addressing spectral variability during spectral unmixing," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 95–104, 2013.

[10] S. Feng, X. Wang, R. Feng, F. Xiong, C. Zhao, W. Li, and R. Tao, "Transformer-based cross-domain few-shot learning for hyperspectral target detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[11] S. Feng, H. Zhang, B. Xi, C. Zhao, Y. Li, and J. Chanussot, "Cross-domain few-shot learning based on decoupled knowledge distillation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[12] D. Hong, B. Zhang, H. Li, Y. Li, J. Yao, C. Li, M. Werner, J. Chanussot, A. Zipf, and X. Zhu, "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sensing of Environment*, vol. 299, pp. 113856, 2023.

[13] J. Feng, T. Zhang, J. Zhang, R. Shang, W. Dong, G. Shi, and L. Jiao, "S4dl: Shift-sensitive spatialspectral disentangling learning for hyperspectral image unsupervised domain adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2025.

[14] J. Qu, W. Dong, Y. Yang, T. Zhang, Y. Li, and Q. Du, "Cycle-refined multidecision joint alignment network for unsupervised domain adaptive hyperspectral change detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 2, pp. 2634–2647, 2025.

[15] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[16] W. Hu, W. Li, H. Li, X. Zhao, M. Zhang, and R. Tao, "Unsupervised domain adaptation with hierarchical masked dual-adversarial network for end-to-end classification of multisource remote sensing data," *IEEE*

Transactions on Geoscience and Remote Sensing, vol. 63, pp. 1–17, 2025.

[17] M. Long, J. Wang, G. Ding, J. Sun, and P. S Yu, "Transfer feature learning with joint distribution adaptation," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2200–2207.

[18] J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, and P. S Yu, "Visual domain adaptation with manifold embedded distribution alignment," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 402–410.

[19] Y. Zhang, W. Li, R. Tao, J. Peng, Q. Du, and Z. Cai, "Cross-scene hyperspectral image classification with discriminative cooperative alignment," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 11, pp. 9646–9660, 2021.

[20] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Computer vision–ECCV 2016 workshops: Amsterdam, the Netherlands, October 8-10 and 15-16, 2016, proceedings, part III 14*. Springer, 2016, pp. 443–450.

[21] X. Zeng and M. Xu, "Gum-net: Unsupervised geometric matching for fast and accurate 3d subtomogram image alignment and averaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4073–4084.

[22] S. Nirmal, V. Sowmya, and K. Soman, "Open set domain adaptation for hyperspectral image classification using generative adversarial network," in *Inventive Communication and Computational Technologies*, pp. 819–827. Springer, 2020.

[23] X. Ma, X. Mou, J. Wang, X. Liu, J. Geng, and H. Wang, "Cross-dataset hyperspectral image classification based on adversarial domain adaptation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4179–4190, 2020.

[24] M. Zhang, X. Zhao, W. Li, Y. Zhang, R. Tao, and Q. Du, "Cross-scene joint classification of multisource data with multilevel domain adaption network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 11514–11526, 2024.

[25] X. Zhao, M. Zhang, R. Tao, W. Li, W. Liao, and W. Philips, "Cross-domain classification of multisource remote sensing data using fractional fusion and spatial-spectral domain adaptation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 5721–5733, 2022.

[26] T. Yang, S. Xiao, J. Qu, W. Dong, Q. Du, and Y. Li, "Graph embedding interclass relation-aware adaptive network for cross-scene classification of multisource remote sensing data," *IEEE Transactions on Image Processing*, vol. 33, pp. 4459–4474, 2024.

[27] W. Dong, J. Qu, T. Zhang, S. Xiao, and Y. Li, "Contrastive constrained cross-scene model- informed interpretable classification strategy for hyperspectral and lidar data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[28] J. Qu, L. Zhang, W. Dong, N. Li, and Y. Li, "Shared-private decoupling-based multilevel feature alignment semisupervised learning for hsi and lidar classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[29] J. Bai, Z. Zhou, Z. Chen, Z. Xiao, E. Wei, Y. Wen, and L. Jiao, "Cross-dataset model training for hyperspectral image classification using self-supervised learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.

[30] Y. Feng, X. Yi, S. Wang, J. Yue, S. Xia, and L. Fang, "Hyperedl: Spectral-spatial evidence deep learning for cross-scene hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[31] M. Cai, B. Xi, J. Li, S. Feng, Y. Li, Z. Li, and J. Chanussot, "Mind the gap: Multi-level unsupervised domain adaptation for cross-scene hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[32] Z. Li, Q. Xu, L. Ma, Z. Fang, Y. Wang, W. He, and Q. Du, "Supervised contrastive learning-based unsupervised domain adaptation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[33] Z. Fang, W. He, Z. Li, Q. Du, and Q. Chen, "Masked self-distillation domain adaptation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–20, 2024.

[34] P. Duan, T. Shan, X. Kang, and S. Li, "Spectral super-resolution in frequency domain," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–11, 2024.

[35] F. Gao, S. Liu, C. Gong, X. Zhou, J. Wang, J. Dong, and Q. Du, "Prototype-based information compensation network for multi-source remote sensing data classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[36] F. Gao, X. Jin, X. Zhou, J. Dong, and Q. Du, "Msfmamba: Multi-scale feature fusion state space model for multi-source remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[37] J. Lin, X. Jin, F. Gao, J. Dong, and H. Yu, "Boosting spatial-spectral masked auto-encoder through mining redundant spectra for hsi-sar/lidar classification," in *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2024, pp. 9744–9747.

[38] M. Wang, F. Gao, J. Dong, H. Li, and Q. Du, "Nearest neighbor-based contrastive learning for hyperspectral and lidar data classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[39] J. Zhang, C. Zhang, S. Liu, Z. Shi, and B. Pan, "Three-dimensional frequency domain transform network for cross-scene hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[40] P. Zhu, X. Zhang, X. Han, X. Cheng, J. Gu, P. Chen, and L. Jiao, "Cross-domain classification based on frequency component adaptation for remote sensing images," *Remote Sensing*, vol. 16, no. 12, pp. 2134, 2024.

[41] X. Zhao, M. Zhang, R. Tao, W. Li, W. Liao, L. Tian, and W. Philips, "Fractional fourier image transformer for multimodal remote sensing data classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 2, pp. 2314–2326, 2022.

[42] B. Tu, X. Yang, B. He, Y. Chen, J. Li, and A. Plaza, "Anomaly detection in hyperspectral images using adaptive graph frequency location," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2024.

[43] S. Feng, S. Wang, C. Xu, C. Zhao, W. Li, and R. Tao, "Fractional domain information enhanced hyperspherical prototype learning method for hyperspectral image open-set classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[44] R. Tian, D. Liu, Y. Bai, Y. Jin, G. Wan, and Y. Guo, "Swin-msp: A shifted windows masked spectral pretraining model for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[45] S. Feng, T. Lan, Y. Fan, M. Zhang, C. Zhao, W. Li, and R. Tao, "An adaptive weighted metric learning network based on fractional domain decoupling for hyperspectral change detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[46] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, "Multi-level wavelet-cnn for image restoration," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 773–782.

[47] E. Oyallon, E. Belilovsky, and S. Zagoruyko, "Scaling the scattering transform: Deep hybrid networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5618–5627.

[48] X. Zhao, R. Tao, W. Li, W. Philips, and W. Liao, "Fractional gabor convolutional network for multisource remote sensing data classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2021.

[49] Y. Dong, G. Li, Y. Tao, X. Jiang, K. Zhang, J. Li, J. Deng, J. Su, J. Zhang, and J. Xu, "Fan: Fourier analysis networks," *arXiv preprint arXiv:2410.02675*, 2024.

[50] S. Pei, M. Yeh, and C. Tseng, "Discrete fractional fourier transform based on orthogonal projections," *IEEE Transactions on Signal Processing*, vol. 47, no. 5, pp. 1335–1348, 1999.

[51] F. Murtagh, "Multilayer perceptrons for classification and regression," *Neurocomputing*, vol. 2, no. 5-6, pp. 183–197, 1991.

[52] S. Mann and S. Haykin, "The chirplet transform: Physical considerations," *IEEE Transactions on Signal Processing*, vol. 43, no. 11, pp. 2745–2761, 1995.

[53] B. Boashash and P. Black, "An efficient real-time implementation of the wigner-ville distribution," *IEEE transactions on acoustics, speech, and signal processing*, vol. 35, no. 11, pp. 1611–1618, 1987.

[54] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[55] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

[56] B. Rasti, P. Ghamisi, and R. Gloaguen, "Hyperspectral and LiDAR fusion using extinction profiles and total variation component analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3997–4007, 2017.