

Received 11 February 2025, accepted 20 March 2025, date of publication 24 March 2025, date of current version 31 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3554045



RESEARCH ARTICLE

QRS-Trs: Style Transfer-Based Image-to-Image Translation for Carbon Stock Estimation in Quantitative Remote Sensing

ZHENYU YU^{ID1}, (Member, IEEE), JINNIAN WANG^{ID2}, HANQING CHEN^{ID3,4},

AND MOHD YAMANI IDNA IDRIS^{ID1}

¹College of Ocean Engineering and Energy, Guangdong Ocean University, Zhanjiang 524005, China

²Guangdong Provincial Key Laboratory of Intelligent Equipment for South China Sea Marine Ranching, Guangdong Ocean University, Zhanjiang 524088, China

³Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia

⁴School of Geography and Remote Sensing, Guangzhou University, Guangzhou 510002, China

Corresponding authors: Hanqing Chen (hanqing@gdou.edu.cn) and Mohd Yamani Idna Idris (yamani@um.edu.my)

This work was supported in part by the National Natural Science Foundation of China under Grant 42201029, and in part by the Program for Scientific Research Start-Up Funds of Guangdong Ocean University.

ABSTRACT Forests serve as vital carbon reservoirs, reducing atmospheric CO₂ and mitigating climate change. Monitoring carbon stocks typically combines ground-based data with satellite remote sensing, yet accuracy remains a challenge. This study analyzes Huize County, China, using GF-1 WFV and Landsat TM images and introduces the Quantitative Remote Sensing Transformer (QRS-Trs), which leverages style transfer and attention mechanisms to enhance carbon stock estimation as an image-to-image translation task. QRS-Trs demonstrates three advantages: 1) Swin-Pix2Pix effectively reduces inter-domain discrepancies caused by sensor and lighting variations while excelling in de-clouding, outperforming Pix2Pix. 2) It incorporates a median filter to eliminate anomalies and a mask module to exclude non-target areas, achieving MAE = 16.29 Mg/ha, RMSE = 29.38 Mg/ha, R² = 0.71, and SSIM = 0.75. 3) Applied to multi-year data, from 2005 to 2020, 44.04% of the area showed increased carbon stock, 10.22% decreased, and 45.74% remained unchanged. While QRS-Trs performs well, its generalization to diverse ecological conditions depends on high-quality training data. Nevertheless, this study provides a robust approach for high-resolution carbon stock estimation, contributing to improved forest carbon sink management.

INDEX TERMS Carbon stock estimation, style transfer, remote sensing, image-to-image translation, swin transformer.

I. INTRODUCTION

Forests serve as crucial carbon reservoirs, playing a significant role in reducing atmospheric CO₂ levels and mitigating climate change. Accurate monitoring and estimation of carbon stocks within forests are essential for understanding carbon distribution, dynamic ecosystem changes, and carbon sequestration capacity. Such analyses provide a scientific basis for predicting ecosystem responses to climate change and for managing and conserving carbon sinks. Additionally, carbon stocks are integral to carbon trading markets, facilitating the trade and transfer of carbon sinks and thereby

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

promoting global efforts in carbon reduction and climate change mitigation [1].

However, existing methodologies for forest carbon stock assessment face several challenges. Traditional statistical models, such as Ordinary Least Squares (OLS) and Geographically Weighted Regression (GWR), struggle to capture nonlinear relationships in complex ecological environments, leading to poor spatial generalization. Machine learning models, including Random Forest (RF) and Support Vector Regression (SVR), improve predictive capabilities but lack hierarchical feature extraction and spatial consistency. Deep learning approaches, such as Convolutional Neural Networks (CNNs), enhance feature representation but often fail to account for inter-domain discrepancies caused by sensor

variations, atmospheric interference (e.g., cloud cover), and inconsistent temporal resolutions [2], [3], [4], [5].

To address these limitations, we propose the Quantitative Remote Sensing Transformer (QRS-Trs), a novel framework that leverages Swin Transformer-based style transfer and advanced filtering techniques for robust carbon stock estimation. Our approach mitigates sensor and lighting discrepancies through Swin-Pix2Pix, ensuring improved multi-temporal consistency. Additionally, the integration of a mask module and median filter refines carbon stock estimation by eliminating anomalies and filtering out non-target regions. These enhancements significantly outperform conventional methods, achieving superior accuracy and robustness in large-scale, long-term carbon stock assessments.

Among the most promising sources for ongoing forest monitoring are medium- to high-resolution (10–30 m) optical datasets. Earlier studies primarily estimated forest biomass and stock volume before calculating carbon stocks, with relatively few directly estimating carbon stocks from remote sensing imagery [6], [7]. The introduction of deep learning methods has significantly improved feature extraction accuracy, enhancing their effectiveness for carbon stock estimation [8]. Style transfer methods, leveraging transfer learning, address domain shifts in multi-temporal images, rectify spectral discrepancies, and enhance generalizability across extended time series. Unlike conventional deep learning approaches, QRS-Trs integrates style transfer with a Swin Transformer backbone, improving domain adaptation and global feature extraction. This innovation enables superior performance in de-clouding and carbon stock estimation across diverse temporal and spatial conditions, outperforming existing models like Pix2Pix in terms of MAE, RMSE, and SSIM.

Despite advancements in remote sensing and machine learning for carbon stock estimation, several limitations persist. The accuracy of satellite-derived data can be compromised by atmospheric conditions, such as cloud cover, and the temporal resolution of satellite passes may not align perfectly with data acquisition needs. Additionally, applying machine learning models across different geographic regions without retraining can lead to inaccuracies due to ecological and climatic variations.

To overcome these challenges, we propose the QRS-Trs model. Our main **contributions** are:

- **Domain Alignment and Image De-clouding.** We introduce the Swin-Pix2Pix model to improve domain alignment by reducing discrepancies in multi-temporal images caused by temporal shifts, sensor variations, and lighting conditions. Furthermore, its enhanced de-clouding capability outperforms traditional Pix2Pix, ensuring more reliable and temporally consistent carbon stock estimation.
- **Feature Extraction and Model Performance.** By integrating the Swin Transformer with Pix2Pix, our model effectively captures global spatial dependencies, improving feature extraction and overall model stability.

The median filter eliminates anomalies, while the mask module excludes non-target areas, significantly enhancing carbon stock estimation accuracy. The QRS-Trs model achieves a MAE of 16.29 Mg/ha, RMSE of 29.38 Mg/ha, R² of 0.71, and SSIM of 0.75.

- **Spatial and Temporal Dynamics of Carbon Stock.** Our model enables large-scale assessments of carbon stock dynamics, revealing that 44.04% of the study area experienced an increase in carbon stock, 10.22% showed a decrease, and 45.74% remained unchanged. From 2005 to 2020, forest coverage expanded from 39.9% to 50.38%, reflecting a positive trend in carbon storage. This improved estimation contributes to a deeper understanding of carbon cycle variations and supports environmental conservation and policymaking.

II. MATERIALS

A. STUDY AREA

Huize County, located in northeastern Yunnan Province and northwestern Qujing City, spans 5,889 km² (Figure 1). The county's elevation varies significantly, ranging from 695 m to 4,017 m above sea level, creating diverse microclimates that influence vegetation growth and biomass distribution. According to the Third National Land Survey (2019), forest land covers 3,080.53 km² (~4,620,800 mu), with arboreal forest comprising 82.39% of the total forested area. The complex topography and varied forest densities introduce challenges in carbon stock estimation, but also provide a valuable opportunity to analyze the impact of geographical features on biomass distribution.

B. DATA SOURCES

1) GROUND-BASED SURVEY DATA

This study utilized data from the Third National Land Survey (Third Survey), conducted by the Forestry and Grassland Bureau of Huize County. The survey, conducted between October 8, 2017, and December 31, 2019, recorded over 70 attributes, including tree accumulation, dominant species, small group area, and species structure. These data serve as ground truth for validating remote sensing-based carbon stock estimates, offering a detailed representation of the region's forest resources.

2) REMOTE SENSING DATA

To complement the ground-based survey, we employed GF-1 WVF image data from China's High-resolution Earth Observation System. GF-1 WVF images, with a spatial resolution of 16 meters, provide temporally consistent coverage of the study area, ensuring compatibility with carbon stock estimation models. Two images recorded on August 27, 2020, were selected for analysis (Table 1).

For long-term monitoring, pre-2013 Landsat TM images with a 30-meter spatial resolution were utilized. To maintain data consistency, we selected only the Red, Green, Blue, and NIR bands from both GF-1 WVF and Landsat TM images.

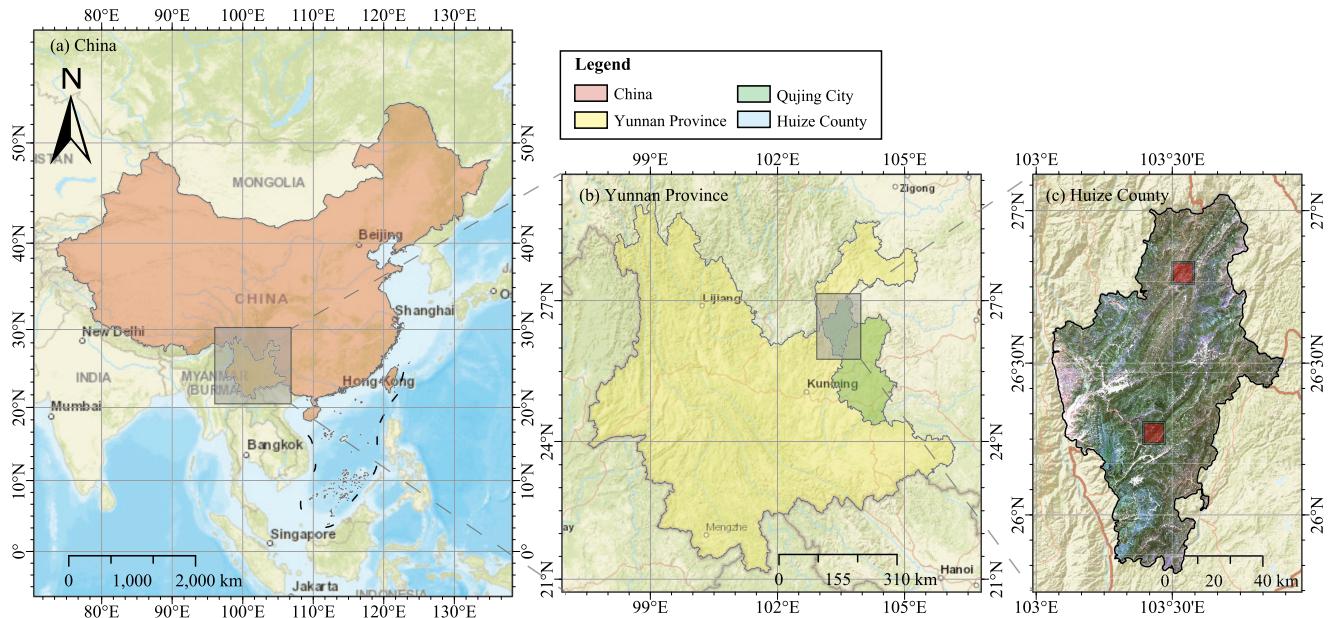


FIGURE 1. Study area. (a) China, (b) Yunnan Province, and (c) Huize County. The red areas in Figure 1(c) indicate the locations shown in Figure 5.

TABLE 1. Image band information of GF-1 WFW and Landsat TM. Note: Desp = Description, No. = Band Number, B = Band, WL = Wavelength, and Res = Resolution.

Desp	GF-1 WFW			Landsat TM		
	No.	WL (μm)	Res (m)	No.	WL (μm)	Res (m)
Blue	B1	0.45~0.52	16	B1	0.45~0.52	30
Green	B2	0.52~0.59	16	B2	0.52~0.60	30
Red	B3	0.63~0.69	16	B3	0.63~0.69	30
NIR	B4	0.77~0.89	16	B4	0.76~0.90	30

The Landsat images, identified by Path 129 and Rows 41 and 42, were mosaicked using the ENVI mosaic tool, with color correction applied via histogram matching.

Additionally, an ALOS PALSAR DEM with a spatial resolution of 12.5 meters was used as a morphological reference, supporting terrain-related adjustments in carbon stock estimation. The DEM dataset, derived from Japan's Advanced Land Observing Satellite (ALOS), provides precise elevation data, aiding in biomass and carbon distribution modeling. A detailed description of data preprocessing is provided in Section I-A.

III. METHOD

A. SWIN TRANSFORMER BLOCK (STB)

Transformers, originally designed for natural language processing, have demonstrated superior performance over convolutional neural networks (CNNs) in tasks requiring global information modeling, such as image classification. While CNNs excel at capturing local features, Transformers handle long-distance dependencies better, although they face challenges in dense prediction tasks like pixel-level segmentation.

The Swin Transformer addresses these challenges by employing sliding windows and hierarchical feature maps, progressively down-sampling images by factors of $4\times$, $8\times$, and $16\times$, enabling adaptive processing across multiple scales, as shown in Figure 3. The Swin Transformer uses Windows Multi-Head Self-Attention (W-MSA) to partition feature maps into disjointed windows, and introduces Shifted Windows Multi-Head Self-Attention (SW-MSA) to facilitate information sharing between adjacent windows. These innovations enhance computational efficiency and improve feature integration across different image regions, making the Swin Transformer particularly effective for pixel-level vision tasks.

Justification for Swin Transformer Block. Compared to traditional Transformers, the Swin Transformer reduces computational complexity by limiting self-attention calculations to non-overlapping windows, making it more efficient for high-resolution image processing. Furthermore, the hierarchical structure of Swin Transformer facilitates multi-scale feature extraction, which is essential for accurately modeling spatial heterogeneity in remote sensing imagery.

Compared to CNNs, which are limited in capturing global dependencies, Swin Transformer incorporates global attention mechanisms that enable better contextual understanding across large spatial regions. Additionally, the use of Shifted Windows Multi-Head Self-Attention (SW-MSA) in Swin Transformer overcomes the boundary artifacts typically observed in standard CNNs, resulting in more coherent feature integration.

Given these advantages, the Swin Transformer Block was selected for this study to balance computational efficiency with improved performance in tasks requiring pixel-level

precision and global feature modeling. Its ability to process high-resolution images while maintaining both local and global feature integration makes it superior to other architectures for carbon stock estimation.

B. MASK SUNET (MSUNET)

Swin U-Net (SUNet) extracts shallow features with a convolutional kernel, then utilizes a U-Net structure with Swin Transformer Blocks (STB) to capture high-level semantic information [9], [10]. SUNet includes five STB layers and uses convolutional kernels for image reconstruction. It replaces traditional down-sampling with patch merging in the encoding stage and employs dual up-sampling in decoding to enhance feature recovery.

1) ADDRESSING BOUNDARY BLURRING AND NOISE

Despite its robustness, SUNet exhibited boundary blurring and noise issues in our experiments. To address this, we incorporated a mask to filter background and boundary-mixed elements, thereby enhancing boundary clarity. Additionally, we integrated a median filter to reduce speckle and salt-and-pepper noise, further improving image quality. The modified Mask SUNet architecture is shown in Figure 2(a).

2) PATCH MERGING

In the down-sampling module, we cascade the input features from each neighborhood block [11], [12] and apply a linear layer to achieve the desired number of output channels. This process effectively functions as an initial stage of convolution by flattening the input feature map, preparing it for subsequent transformation layers.

3) TRIPLE UP-SAMPLE MODULE

The original Swin-UNet employed transposed convolution for up-sampling, which led to blocking artifacts. Fan et al. [10] addressed this issue with dual up-sampling (bilinear and pixel shuffle) to eliminate checkerboard artifacts. However, this solution was insufficient for achieving precise boundaries. To further improve boundary definition, we introduced a triple up-sample module, which adds transposed convolution to the dual up-sample setup, effectively resolving checkerboard artifacts and significantly improving boundary definition.

4) MEDIAN FILTER APPLICATION

Despite these enhancements, residual noise around the mask still persisted due to the model's limited ability to estimate boundary transitions. While the mask was helpful, it didn't fully eliminate the noise. The median filter effectively removed this noise but introduced slight blurring. To address this, we focused on removing high-intensity elements (exceeding 240), which preserved essential image characteristics while reducing noise.

5) POTENTIAL LIMITATIONS OF MASK MISCLASSIFICATIONS

While the mask significantly improved boundary clarity, potential limitations remain. Misclassifications in the mask

can arise due to factors such as inaccurate thresholds or overlapping regions between target and background areas. These misclassifications may lead to incorrect filtering of critical features or retention of noise, which could affect the overall segmentation accuracy. For example, if non-target regions are incorrectly classified as part of the target, this could introduce artifacts in the final output. Conversely, under-segmentation may exclude important target features, reducing model performance. Future improvements could include adaptive thresholding techniques or integrating additional contextual information to mitigate such errors.

These advancements in the SUNet architecture and processing techniques effectively address specific challenges encountered in image segmentation tasks, resulting in improved accuracy and quality in the final output.

C. QUANTITATIVE REMOTE SENSING TRANSLATION (QRS-TRS)

1) QRS-TRS STRUCTURE

The QRS-Trs architecture (Figure 2) integrates MSUNet and Patch GAN to optimize local and global accuracy in image generation. This setup effectively addresses challenges in modeling and visual data synthesis, particularly for high-fidelity environmental feature representation.

2) GENERATOR

The MSUNet was employed as the generator, using a mask to filter out non-forest areas, thereby improving carbon stock estimation accuracy and model convergence speed. The discriminator utilized the Patch GAN, with specific loss functions to enhance training efficacy and reduce domain discrepancies in images taken at different times. By using SUNet as the generator and appropriate loss functions, we achieved efficient domain adaptation without needing a masking filter during this phase.

3) DISCRIMINATOR

Adversarial training stability is heavily influenced by the discriminator architecture. Our experiments showed that replacing traditional convolution with Swin Transformer Blocks (STB) and increasing model parameters did improve stability but also added significant overhead with marginal benefits. As a result, we continued using Patch GAN as the discriminator.

The original GAN discriminator produces a single binary value for the whole image, while Patch GAN evaluates each patch within the image as a fully convolutional network. This $n \times n$ matrix approach focuses on local patches, which effectively penalizes local inconsistencies. Our experiments showed that the patch discriminator successfully reduced blocking artifacts in generated images.

D. OBJECTIVE FUNCTION

1) \mathcal{L}_1 LOSS

Isola et al. [13] introduced the Pix2Pix Conditional Generative Adversarial Network (cGAN), utilizing the \mathcal{L}_1 -distance

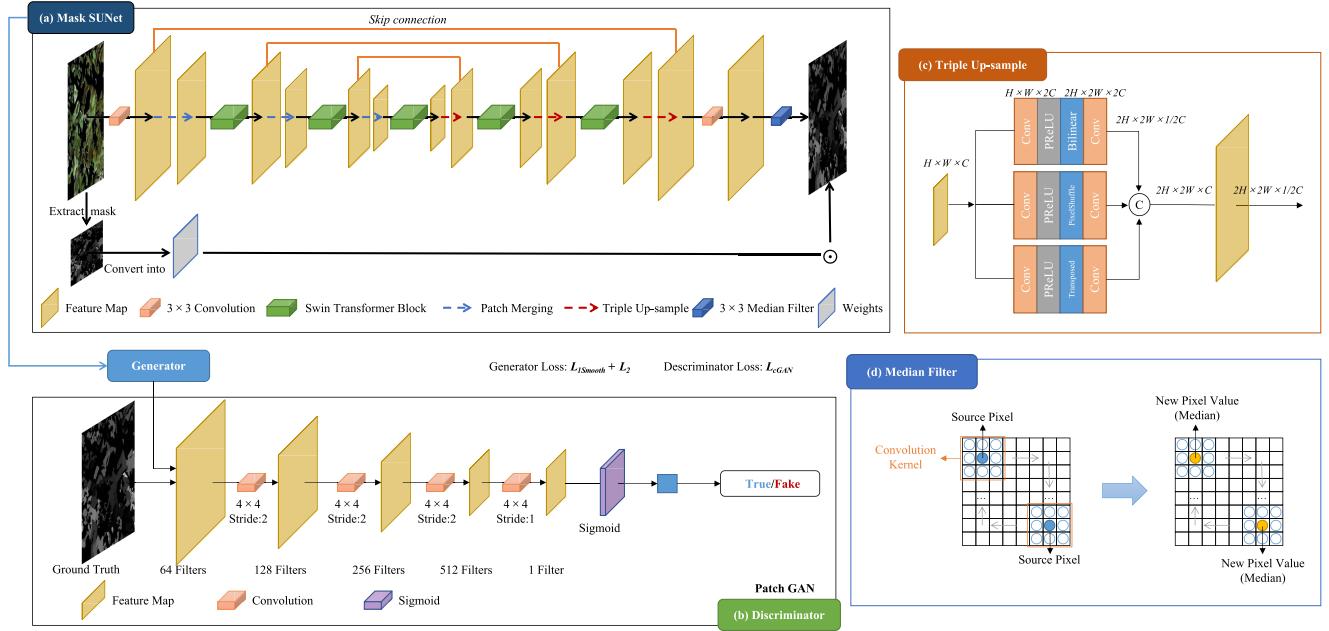


FIGURE 2. QRS-Trs architecture. (a) Generator: Mask SUNet module. (b) Discriminator: cGAN. The top panel represents Mask SUNet, while the bottom panel shows cGAN, the right-top panel shows triple up-sample, and the right-bottom panel depicts the median filter.

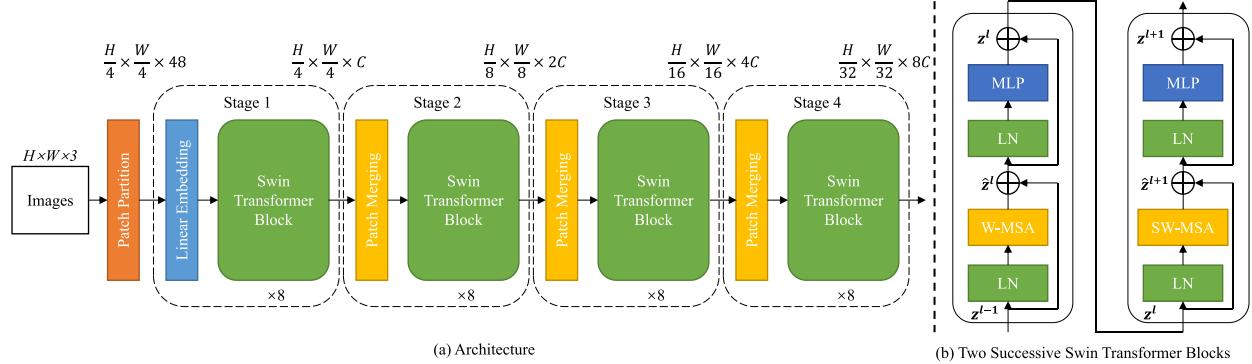


FIGURE 3. Swin Transformer (Swin-T) architecture. (a) is the overall architecture, and (b) is two successive Swin Transformer blocks.

as the loss function, as expressed in Eq.1. In this formulation, x represents the ground truth, while $G(u)$ denotes the synthetic image generated conditionally based on the input data u . The L_1 -distance, also known as the Mean Absolute Error (MAE) Loss, was selected due to its stable gradient across a range of input values, which mitigates the risk of gradient explosion and ensures a robust solution.

$$\mathcal{L}_1(G) = \mathbb{E}_{x \sim P_d(x)}[\|x - G(u)\|_1] \quad (1)$$

2) \mathcal{L}_2 LOSS

Pathak et al. [14] used the \mathcal{L}_2 -distance, as depicted in Eq.2. The \mathcal{L}_2 distance, also known as the Mean Square Error (MSE) Loss, is continuous and smooth at all points, easy to derive, and provides a more stable solution. However, it is sensitive to outliers; when the input value of the function

deviates significantly from the true value, the corresponding value of the loss becomes large, leading to a large gradient during gradient descent, which can result in gradient explosion. The \mathcal{L}_2 -distance is particularly suitable for regression tasks with small numerical characteristics and low dimensionality.

$$\mathcal{L}_2(G) = \mathbb{E}_{x \sim P_d(x)}[(x - G(u))^2] \quad (2)$$

3) $\mathcal{L}_{1SMOOTH}$ LOSS

When the difference between the predicted and true values was small (the absolute value of the difference being less than 1), \mathcal{L}_2 was applied; when the difference was large, a translation of \mathcal{L}_1 was employed. $\mathcal{L}_{1Smooth}$ is a combination of \mathcal{L}_1 and \mathcal{L}_2 , leveraging the advantages of both approaches. $\mathcal{L}_{1Smooth}$ addresses the problem of unsmoothed zeros and is more robust to outliers than \mathcal{L}_2 . With \mathcal{L}_2 , the gradient

is smaller, and the loss function is more rounded than \mathcal{L}_1 , which can result in faster convergence. Conversely, with \mathcal{L}_1 , the gradient is sufficiently small, more stable, and less prone to gradient explosion. In regression tasks, \mathcal{L}_1 is more suitable when there are larger values in the features.

$$\begin{aligned} \mathcal{L}_{1S}(G) \\ = & \begin{cases} E_{x \sim P_d(x)}[(x - G(u))^2 \times 0.5], & \text{if } \|x - G(u)\| < 1 \\ E_{x \sim P_d(x)}[\|x - G(u)\| - 0.5], & \text{otherwise} \end{cases} \end{aligned} \quad (3)$$

Huber loss, which blends the strengths of \mathcal{L}_1 and \mathcal{L}_2 , was not adopted in this study because $\mathcal{L}_{1Smooth}$ serves a similar purpose while being simpler to implement. Both losses provide robustness to outliers and smooth gradients for small errors, but $\mathcal{L}_{1Smooth}$ aligns more naturally with the QRS-Trs framework and its dual objectives of style transfer and carbon stock estimation.

Additionally, Huber loss requires fine-tuning the threshold parameter δ , which can introduce additional complexity without guaranteeing significant performance gains in this specific application. $\mathcal{L}_{1Smooth}$ effectively balances simplicity and performance, addressing both boundary artifacts and noise. Therefore, it was selected as the more practical choice for this study.

4) \mathcal{L}_{cGAN} LOSS

The original GAN comprised a generator (G) and a discriminator (D), where G and D participated in a minimax game. G can learn to map a random noise vector (z) and the observed image (u) to produce an output (y), as shown in Eq.4. The objective function of \mathcal{L}_{cGAN} is given in Eq.5.

$$\begin{aligned} G : \{z, u\} \rightarrow y \quad (4) \\ \mathcal{L}_{cGAN}(G, D) = E_{x \sim P_d(x)}[\log D(x, u)] \\ + E_{z \sim P_z(z)}[\log(1 - D(G(z, u)))] \quad (5) \end{aligned}$$

5) QRS-TRS LOSS

The loss functions of the Swin-Pix2Pix model used for style transfer in this paper were \mathcal{L}_{cGAN} and \mathcal{L}_1 , as shown in Eq.6. In contrast, the losses of the QRS-Trs model used for estimation included \mathcal{L}_{cGAN} , $\mathcal{L}_{1Smooth}$, and \mathcal{L}_2 , as shown in Eq.7, where $\lambda = 100$.

$$G_1^* = \arg \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_1(G) \quad (6)$$

$$G_2^* = \arg \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_2(G) + \lambda \mathcal{L}_{1S}(G) \quad (7)$$

E. EVALUATION METRICS

Mean Absolute Error (MAE) is the mean of the absolute error between the predicted and true values. Mean Squared Error (MSE) is the mean of the squared error between the predicted and true values. The Structure Similarity Index Measure (SSIM) is a full-reference image quality evaluation metric that measures image similarity in terms of luminance, contrast, and structure, respectively.

IV. RESULTS

A. IMAGE RECONSTRUCTION

Variations in sensors, lighting, and atmospheric conditions introduce domain shifts in multi-temporal satellite images, complicating carbon stock estimation. To mitigate these discrepancies, we employed a style transfer algorithm to harmonize the spatial distribution of images across different time periods. This process is crucial for ensuring the accuracy of carbon stock estimation.

The study utilizes images from 2005, 2013, 2015, and 2020, where the 2005 image is a Landsat TM image (30 m resolution), and the later images are GF-1 WVF images (16 m resolution). The Swin-Pix2Pix model was applied for style transfer, aligning image features while preserving texture details. The results are displayed in Figure 4. In the figure, input x represents the images to be transformed, the 2020 image serves as the style reference, and the output $G(x)$ represents the style-transferred image that combines the content of input x with the style attributes of the 2020 ground truth image.

1) ENHANCING IMAGE STYLIZATION QUALITY

The Swin-Pix2Pix algorithm leverages the Swin Transformer's ability to capture global contextual features, ensuring uniform feature distribution and preventing distortions in de-clouded regions. This improvement is particularly beneficial in cloud-covered areas, where image restoration requires consistent texture generation. Unlike traditional style transfer methods, Swin-Pix2Pix effectively integrates stylized attributes such as spectral consistency and surface texture while maintaining the structural integrity of the original image.

2) OPTIMAL PARAMETER SELECTION FOR CLARITY AND STYLE ACCURACY

Experimental results indicate that using only \mathcal{L}_1 loss results in blurred images, whereas \mathcal{L}_2 loss alone produces sharper images but introduces spectral deviations. The optimal combination of \mathcal{L}_1 and \mathcal{L}_2 losses improves clarity while maintaining the original color spectrum, significantly enhancing the accuracy of environmental data interpretation.

B. CARBON STOCK ESTIMATION

1) COMPARISON OF ESTIMATION MODELS

Table 2 and Figure 5 present the comparative results of the QRS-Trs model against statistical models (OLS, GWR), machine learning models (RF, SVR), and deep learning models (CNN, Pix2Pix). We employed five-fold cross-validation for performance evaluation.

OLS, GWR, RF, and SVR models exhibited R^2 values below 0.5 and SSIM values below 0.4, indicating poor numerical and spatial accuracy. Their high RMSE and MAE values further suggest substantial estimation errors exceeding 100%. The deep learning models demonstrated superior performance, with CNN being the least effective

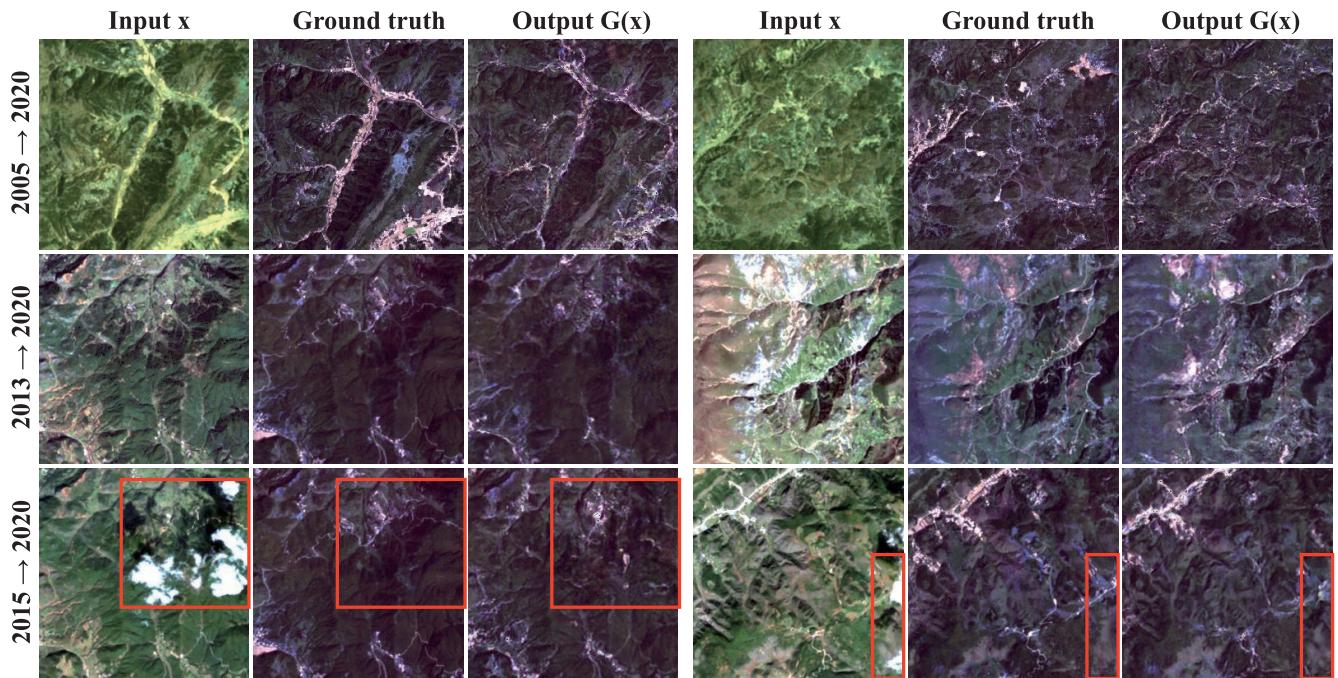


FIGURE 4. Style transfer results. Notes: The ground truth is the GF-1 WVF image in 2020, and the red box highlights the cloudy area.

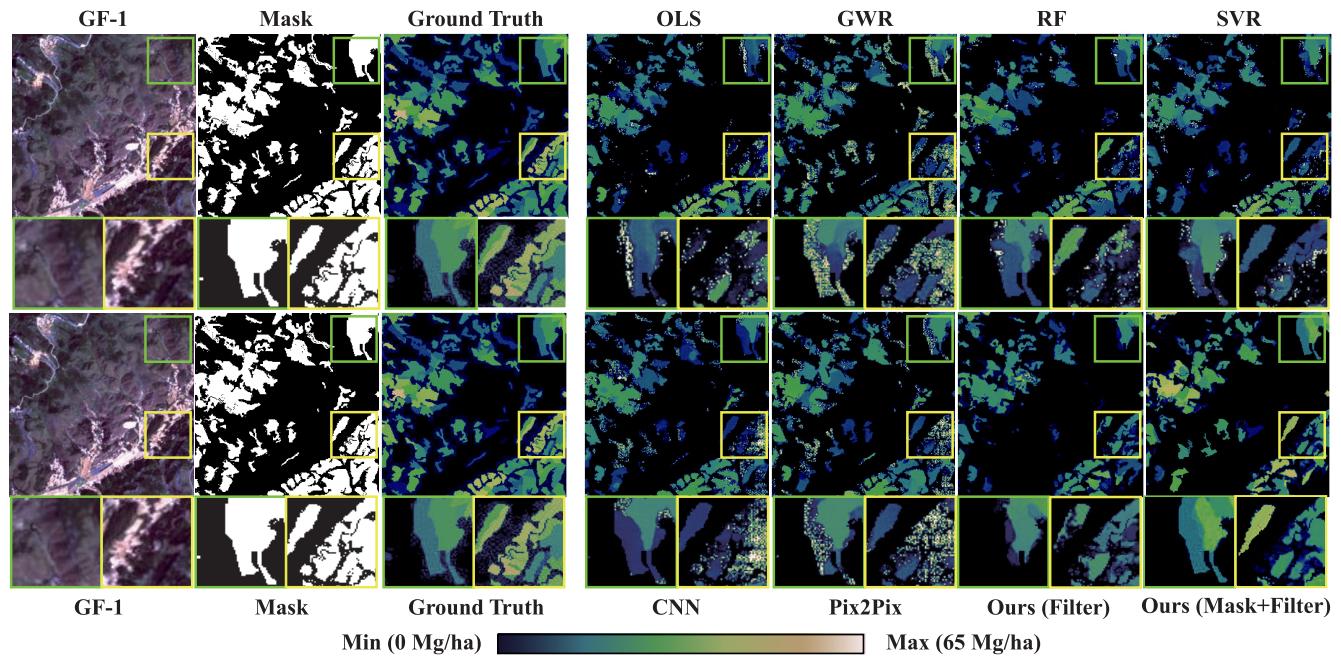


FIGURE 5. Carbon stock estimation results.

among them, followed by Pix2Pixel, while QRS-Trs outperformed all models. Except for CNN, all deep learning models achieved R^2 values above 0.5 and SSIM values above 0.65, highlighting improved consistency in both numerical accuracy and spatial representation. Notably, QRS-Trs achieved the highest accuracy (MAE = 16.29 Mg/ha, RMSE = 29.38 Mg/ha, R^2 = 0.71, SSIM = 0.75).

Among the traditional models, OLS and GWR were least effective in estimating carbon stock patterns, primarily due to their limited capability in modeling spatial dependencies. Machine learning models, RF and SVR, performed slightly better but lacked deep hierarchical feature extraction. CNN models were inconsistent, while the integration of Swin Transformer into Pix2Pixel significantly enhanced global

TABLE 2. Comparison of carbon stock estimation. Notes: OLS = Ordinary Least Squares; GWR = Geographically Weighted Regression; RF = Random Forest; SVR = Support Vector Regression; CNN = Convolutional Neural Network. **Bold** represents the best performance, and underline denotes the second best. MF = Median Filter. Mask and MF were not applied to OLS, GWR, RF, and SVR, as these models perform point-based fitting rather than spatial surface estimation.

Model	Mask	MF	MAE↓	RMSE↓	R ² ↑	SSIM↑
OLS	-	-	65.59	104.95	0.39	0.26
GWR	-	-	55.40	93.34	0.44	0.33
RF	-	-	52.98	89.93	0.43	0.37
SVR	-	-	50.22	85.85	0.43	0.36
Swin-Pix2Pix	×	×	18.98	34.52	0.60	0.70
	×	✓	18.39	33.65	0.63	0.72
	✓	×	<u>18.09</u>	<u>33.08</u>	0.64	0.70
QRS-Trs	✓	✓	16.29	29.38	0.71	0.75

feature extraction and model stability, resulting in the best overall performance.

2) ABLATION STUDY FOR DIFFERENT MODULES

To further investigate the contribution of individual components, an ablation study was conducted, as summarized in Table 2. The impact of the median filter and mask module was analyzed separately.

The median filter effectively reduced noise and improved numerical accuracy, albeit with minor blurring in block representations. The mask module provided a more significant improvement by enhancing boundary detail extraction and filtering non-vegetative areas, reducing noise in non-target regions. The combination of both modules in Swin-Pix2Pix led to the highest performance gains, attributed to enhanced boundary handling and effective global feature extraction via the transformer's attention mechanism. The median filter suppressed local anomalies, while the mask module removed irrelevant non-target noise, leading to significantly improved model accuracy. Traditional models, lacking global feature extraction capabilities, did not benefit substantially from these modules, underscoring the unique effectiveness of the proposed approach.

C. SPATIAL AND TEMPORAL VARIATION CHARACTERISTICS

The carbon stock dynamics in Huize County from 2005 to 2020 reveal significant spatial and temporal variations. Over this period, carbon stock increased in 44.04% of the area (2,593.76 km²), decreased in 10.22% (601.56 km²), and remained unchanged in 45.74%. Figure 6 and A1 illustrate the spatial distribution of these changes across four time periods.

The QRS-Trs model played a key role in producing high-resolution carbon stock estimations. By leveraging Swin-Pix2Pix for style transfer, the model ensured multi-temporal consistency in remote sensing images, reducing inter-domain discrepancies caused by sensor variations and atmospheric influences. Additionally, the mask module filtered out non-vegetation areas, while the median filter

suppressed anomalies, resulting in more stable and accurate carbon stock estimates.

V. DISCUSSION

A. REDUCING THE IMPACT OF LOW-QUALITY DATA

Satellite imagery is essential for long-term, large-scale environmental assessments, but it often suffers from distortions due to sensor variations and illumination effects. Traditional correction methods, such as matched histogram and relative radiation correction, have notable limitations [15], [16]. Matched histogram techniques enhance contrast but are sensitive to noise and struggle with nonlinear transformations, making them highly dependent on precise target histograms. Relative radiation correction improves brightness and details under varying lighting conditions but does not fully mitigate spectral inconsistencies across different image sources.

Cloud cover has been a persistent challenge in highland regions, restricting the availability of high-quality data over time [17]. To address this, we employed a transfer learning-based image style transfer model, integrating the Pix2Pix framework with the Swin Transformer to enhance global feature extraction. This approach minimizes domain offsets and optimizes content and style fusion, leading to improved image consistency and enhanced spectral feature representation compared to traditional Pix2Pix. The Swin Transformer further improves texture preservation and adapts better to varying environmental conditions, ultimately enhancing the quality of reconstructed images.

B. ENHANCING CARBON STOCK ESTIMATION ACCURACY

Huize County's diverse forest resources provide a strong basis for carbon stock estimation, but its varied altitude and terrain introduce significant challenges. Current methods for estimating carbon stocks rely on satellite-based LiDAR and optical imaging. While LiDAR is highly accurate, it faces scalability limitations due to cost and coverage constraints. In contrast, GF-1 optical imagery offers high spatio-temporal resolution, making it more suitable for large-scale applications. However, optical images are susceptible to cloud cover and sensor-induced discrepancies, reducing estimation reliability.

To address these limitations, we implemented an advanced image transfer method (Figure 4), which mitigates cloud interference while maintaining high accuracy. Nevertheless, achieving consistently accurate predictions remains challenging due to the complexity of nonlinear relationships in carbon stock estimation. Traditional machine learning models such as OLS, RF, and SVR [18], [19], [20] struggle with these complexities, particularly in spatially heterogeneous landscapes.

Our QRS-Trs model overcomes these challenges by integrating a Transformer-based backbone with deep learning and attention mechanisms for global feature extraction. Additionally, the incorporation of a median filter refines local features and effectively isolates non-vegetated areas. These enhancements lead to substantial improvements in estimation

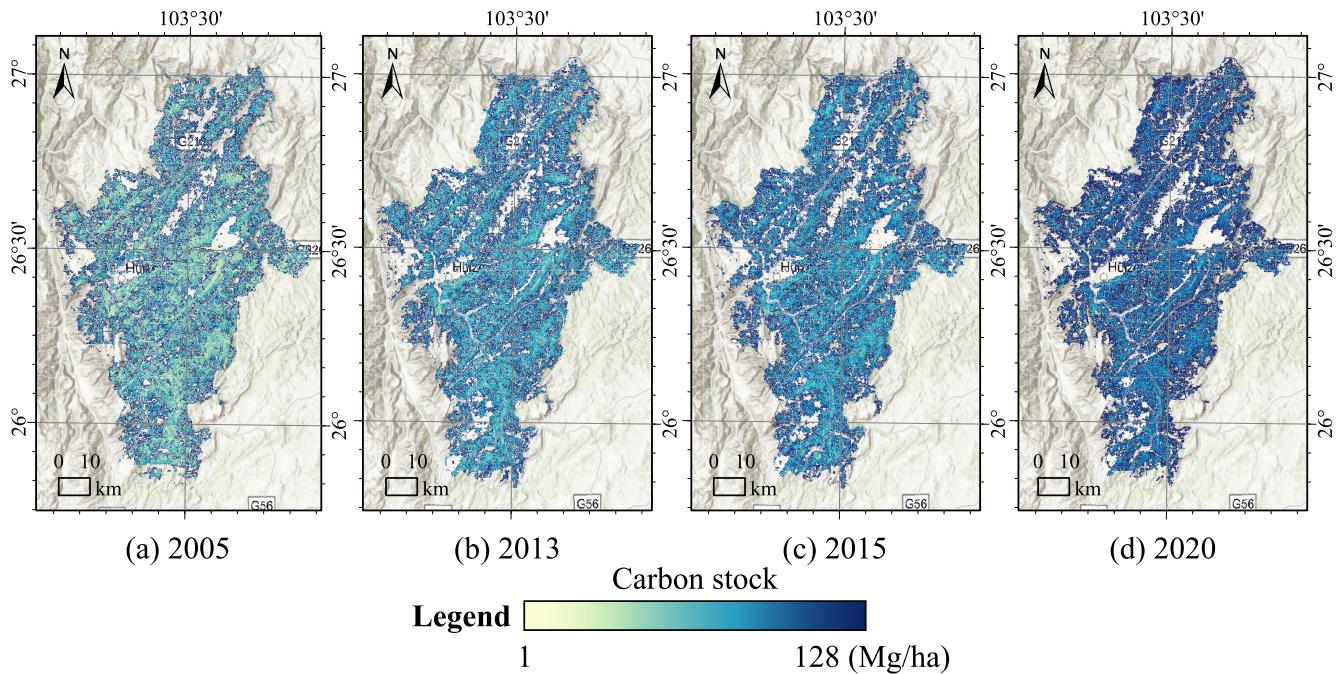


FIGURE 6. Carbon stock temporal and spatial variation characteristics. We selected the results of four months for visualization.

accuracy, achieving an RMSE of 29.38 Mg/ha, which approaches the performance of LiDAR-based estimations (25.64 Mg/ha) [21], while outperforming coarse-resolution multi-temporal imagery (approximately 30 Mg/ha) [22]. Although previous studies focused primarily on biomass estimation [21], [22], the strong correlation between biomass and carbon stocks ensures the relevance of these comparisons.

The use of high-resolution, multi-temporal images presents a significant opportunity to further improve carbon stock estimation. These findings suggest that integrating deep learning-based domain adaptation techniques with traditional remote sensing approaches can advance both methodological frameworks and practical applications, enabling more accurate, large-scale environmental monitoring.

VI. CONCLUSION

This study introduced the QRS-Trs model for carbon stock estimation, integrating Swin-Pix2Pix for enhanced **domain alignment** and **de-clouding**. By leveraging Swin Transformer for global feature extraction and incorporating a mask and median filter, QRS-Trs improved the accuracy and consistency of multi-temporal remote sensing data.

A. KEY CONTRIBUTIONS

(1) The Swin-Pix2Pix model effectively reduced discrepancies in multi-temporal images, improving long-term carbon stock estimation. (2) The mask and median filter enhanced spatial precision by eliminating non-target regions and reducing anomalies. (3) The QRS-Trs model achieved high accuracy ($MAE = 16.29$ Mg/ha, $RMSE = 29.38$ Mg/ha, $R^2 = 0.71$, $SSIM = 0.75$), outperforming traditional models in spatial consistency.

B. FINDINGS AND IMPLICATIONS

Between 2005 and 2020, carbon stock increased in 44.04% of the study area, highlighting a positive ecological trend. QRS-Trs demonstrates strong potential for large-scale forest monitoring, offering a scalable and robust approach to carbon stock estimation.

VII. SUPPLEMENTARY

A. PREPROCESSING

1) FEATURE EXTRACTION

The data analysis involved extracting a comprehensive set of features, categorized into topographic, spectral, texture, and vegetation index features, utilizing a total of 50 bands, as detailed in Table 3.

2) TOPOGRAPHIC FEATURES

Derived from the Digital Elevation Model (DEM), these features encompass 11 bands including slope and aspect, providing essential terrain context for the study area.

3) SPECTRAL FEATURES

These consist of four bands from the GF-1 WVF images, capturing the primary spectral information required for the analysis.

4) TEXTURE FEATURES

Utilizing the Gray-Level Co-occurrence Matrix (GLCM), texture features were extracted, including mean and variance calculations for each of the four GF-1 bands. This resulted in a total of 32 feature bands, enriching the analysis with detailed textural insights of the vegetation cover.

TABLE 3. Features information. Notes: B = Band.

Features	Spectral	Topographical			Vegetation Index	Texture	
Indicators	B1	Slope	Plan Convexity	Maximum Curvature	NDVI	Mean	Dissimilarity
	B2	Aspect	Longitudinal Convexity	RMS	DVI	Variance	Entropy
	B3	Shaded Relief	Cross Sectional Convexity	Slope Percent	RVI	Homogeneity	Second Moment
	B4	Profile Convexity	Minimum Curvature	-	-	Contrast	Correlation
Total	4 bands	11 bands			3 bands	32 bands	

5) VEGETATION INDEX FEATURES

Vegetation indices such as the Normalized Difference Vegetation Index (NDVI), Difference Vegetation Index (DVI), and Ratio Vegetation Index (RVI) were computed using Eq.A1~A3. These indices are calculated from the Near-Infrared (NIR) and Red bands, providing critical measures of vegetation health and density.

The features extracted from these diverse data sources integrate to form a robust dataset for subsequent analysis, enhancing the understanding of the forest resources in Huize County.

$$NDVI = \frac{\rho_{NIR} - \rho_{RED}}{\rho_{NIR} + \rho_{RED}} \quad (A1)$$

$$RVI = \frac{\rho_{NIR}}{\rho_{RED}} \quad (A2)$$

$$DVI = \rho_{NIR} - \rho_{RED} \quad (A3)$$

6) FEATURE SCREENING

To overcome the limitations of relying on a single correlation metric, our study utilized a comprehensive approach by selecting bands that exhibited the highest correlation with carbon stock. We incorporated a variety of correlation measures, including Pearson, Spearman, and Kendall coefficients [23], Cosine Similarity [24], and distance measures such as Euclidean, Manhattan, and Chebyshev [25]. These metrics helped determine the correlation between each band and the measured carbon stock values. From the cumulative ranking of each band, a comprehensive score was derived, allowing for an objective selection of the most relevant bands. The bands that showed the highest correlation, based on this integrated scoring method, were the 19th (GLCM-Mean_Band1), 35th (GLCM-Mean_Band3), and 27th (GLCM-Mean_Band2). Detailed correlation data for each band is presented in Table A3. This multi-index approach ensures a robust selection process, mitigating the risk of bias inherent in single-index evaluations.

7) MASK CALCULATION

In this study, the Normalized Difference Vegetation Index (NDVI) was employed to delineate vegetation from non-vegetation areas effectively. NDVI was chosen for its ability to capture the spectral characteristics of forest land, which were used to create a mask for the study area. The calculation

of the threshold for this vegetation index, which determines the distinction between vegetated and non-vegetated regions, is detailed in Eq.A4. This method enhances the accuracy of the forest land analysis by providing a clear separation based on spectral data, thereby facilitating more precise environmental assessments.

$$M = \bar{M} - 2\sigma \quad (A4)$$

where, M is the threshold, \bar{M} is the average, and σ is the standard deviation of the vegetation index. With this formula, the threshold value for 2020 was 0.3951 (~ 0.40), where \bar{M} was 0.69 and σ was 0.15, and the NDVI was binarized, and the area above the threshold value was regarded as vegetation (set as 1) and the area below the threshold value was regarded as non-vegetation (set as 0). This extraction method was used to apply the mask calculation for the remaining years.

B. STUDY AREA

Huize County was situated in the northeastern part of Yunnan Province and the northwestern part of Qujing City. It was located at the junction of Yunnan, Sichuan, and Guizhou Provinces. The county occupied a land area of 5,889 km² as shown in Fig. 1. The county was characterized by mountainous terrain that gradually decreased from west to east. The county's highest peak was 4,017 m above sea level, which was the highest peak in Qujing City. In contrast, its lowest point was only 695 m above sea level, making it the lowest point in Qujing City. According to data from the Third National Land Survey with December 31, 2019, forest land covered 3,080.53 km² ($\sim 4,620,800$ mu), with arboreal forest land accounting for 82.39% of the total area. The complexity of the county's physical geography presented a significant challenge, even though the abundance of forest resources provided ample samples for the study.

C. DATA SOURCES

The study leveraged the data from the Third National Land Survey (Third Survey), conducted by the Forestry and Grassland Bureau of Huize County. This comprehensive survey, spanning from October 8, 2017, to December 31, 2019, and completed in 2020, captured more than 70 attributes including accumulation, dominant tree species, small group area, and tree species structure. These attributes provide

TABLE 4. Data information.

Name	Abbreviation	Format	Date	Resolution	Product ID/Source
GF-1 WVF	GF-1	GeoTIFF	2020/8/27	16 m	GF1_WFV3_E103.7_N25.6_20200827_L1A0005020371
GF-1 WVF	GF-1	GeoTIFF	2020/8/27	16 m	GF1_WFV3_E104.2_N27.3_20200827_L1A0005020368
ALOS PALSAR DEM	DEM	GeoTIFF	2006~2011	12.5 m	AP_24748_FBD_F0520_RT1.dem.tif
ALOS PALSAR DEM	DEM	GeoTIFF	2006~2011	12.5 m	AP_24748_FBD_F0510_RT1.dem.tif
ALOS PALSAR DEM	DEM	GeoTIFF	2006~2011	12.5 m	AP_24748_FBD_F0500_RT1.dem.tif
ALOS PALSAR DEM	DEM	GeoTIFF	2006~2011	12.5 m	AP_19132_FBD_F0510_RT1.dem.tif
Global Canopy Height 2020	Canopy Height	GeoTIFF	2020	10 m	-
Forest Resource Management Inventory Data	-	Shapefile	2020	-	Forestry and Grassland Bureau of Huize County

a reliable representation of the forest resources in the study area, which is essential for accurate forest resource estimation. The detailed and precise data refined from this survey enhance our understanding of the changes in land use and natural resource changes across China.

For the remote sensing component, the study utilized GF-1 WVF image data from the first satellite of China's high-resolution Earth observation system, launched in April 2013. These images, with a spatial resolution of 16 meters, ensure temporal consistency by covering two scenes recorded on August 27, 2020, thus providing comprehensive coverage of the study area. The specifics of these images are detailed in Table 4.

Prior to 2013, Landsat TM image data, with a spatial resolution of 30 meters, were primarily used to monitor the area over the long term. To maintain data comparability, only the Red, Green, Blue, and NIR bands were selected from the GF-1 WVF and Landsat TM images. Landsat images within the study area, identified by Path 129 and Rows 41 and 42, were seamlessly stitched using the ENVI mosaic tool, and color correction was applied by histogram matching, as detailed in Table 1.

Additionally, an ALOS PALSAR DEM with a spatial resolution of 12.5 meters served as a morphological reference. The Advanced Land Observing Satellite (ALOS), operational from 2006 to 2011, aimed to enhance mapping accuracy, regional land cover observations, disaster monitoring, and resource surveys. The DEM data, integral to our study, provided detailed measurements throughout all seasons and times of the day, as outlined in Table 4.

REFERENCES

- [1] R. Zhang, X. Zhou, Z. Ouyang, V. Avitabile, J. Qi, J. Chen, and V. Giannico, "Estimating aboveground biomass in subtropical forests of China by integrating multisource remote sensing and ground data," *Remote Sens. Environ.*, vol. 232, Oct. 2019, Art. no. 111341.
- [2] Z. Liu, F. Zhao, X. Liu, Q. Yu, Y. Wang, X. Peng, H. Cai, and X. Lu, "Direct estimation of photosynthetic CO₂ assimilation from solar-induced chlorophyll fluorescence (SIF)," *Remote Sens. Environ.*, vol. 271, Mar. 2022, Art. no. 112893.
- [3] I. E. Teubner, M. Forkel, G. Camps-Valls, M. Jung, D. G. Miralles, G. Tramontana, R. van der Schalie, M. Vreugdenhil, L. Mössinger, and W. A. Dorigo, "A carbon sink-driven approach to estimate gross primary production from microwave satellite observations," *Remote Sens. Environ.*, vol. 229, pp. 100–113, Aug. 2019.
- [4] A. Hamedianfar, C. Mohamedou, A. Kangas, and J. Vauhkonen, "Deep learning for forest inventory and planning: A critical review on the remote sensing approaches so far and prospects for further applications," *Forestry, Int. J. Forest Res.*, vol. 95, no. 4, pp. 451–465, Aug. 2022.
- [5] M. Santoro, O. Cartus, and J. E. S. Fransson, "Dynamics of the Swedish forest carbon pool between 2010 and 2015 estimated from satellite L-band SAR observations," *Remote Sens. Environ.*, vol. 270, Mar. 2022, Art. no. 112846.
- [6] M. Chopping, Z. Wang, C. Schaaf, M. A. Bull, and R. R. Duchesne, "Forest aboveground biomass in the southwestern United States from a MISR multi-angle index, 2000–2015," *Remote Sens. Environ.*, vol. 275, Jun. 2022, Art. no. 112964.
- [7] Z. Yu, J. Wang, X. Yang, and J. Ma, "Superpixel-based style transfer method for single-temporal remote sensing image identification in forest type groups," *Remote Sens.*, vol. 15, no. 15, p. 3875, Aug. 2023.
- [8] N. Lang, N. Kalischek, J. Armston, K. Schindler, R. Dubayah, and J. D. Wegner, "Global canopy height regression and uncertainty estimation from GEDI LiDAR waveforms with deep ensembles," *Remote Sens. Environ.*, vol. 268, Jan. 2022, Art. no. 112760.
- [9] B. Fan, Y. Dai, and M. He, "SUNet: Symmetric undistortion network for rolling shutter correction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4521–4530.
- [10] C.-M. Fan, T.-J. Liu, and K.-H. Liu, "SUNet: Swin transformer UNet for image denoising," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2022, pp. 2333–2337.
- [11] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
- [12] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2021, pp. 205–218.
- [13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.
- [14] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.
- [15] L. Zhang, G. Su, J. Yin, Y. Li, Q. Lin, X. Zhang, and L. Shao, "Bioinspired scene classification by deep active learning with remote sensing applications," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 5682–5694, Jul. 2022.
- [16] D. P. Roy, J. Ju, P. Lewis, C. Schaaf, F. Gao, M. Hansen, and E. Lindquist, "Multi-temporal MODIS–Landsat data fusion for relative radiometric normalization, gap filling, and prediction of Landsat data," *Remote Sens. Environ.*, vol. 112, no. 6, pp. 3112–3130, Jun. 2008.
- [17] L. E. Christovam, M. H. Shimabukuro, M. D. L. B. T. Galo, and E. Honkavaara, "Pix2pix conditional generative adversarial network with MLP loss function for cloud removal in a cropland time series," *Remote Sens.*, vol. 14, no. 1, p. 144, Dec. 2021.
- [18] S. Eckert, "Improved forest biomass and carbon estimations using texture measures from WorldView-2 satellite data," *Remote Sens.*, vol. 4, no. 4, pp. 810–829, Mar. 2012.
- [19] B. K. V. Yadav and S. Nandy, "Mapping aboveground woody biomass using forest inventory, remote sensing and geostatistical techniques," *Environ. Monitor. Assessment*, vol. 187, no. 5, pp. 1–12, May 2015.
- [20] C. Zhang, S. Denka, H. Cooper, and D. R. Mishra, "Quantification of sawgrass Marsh aboveground biomass in the coastal Everglades using object-based ensemble analysis and Landsat data," *Remote Sens. Environ.*, vol. 204, pp. 366–379, Jan. 2018.

- [21] L. Cao, N. C. Coops, J. L. Innes, S. R. J. Sheppard, L. Fu, H. Ruan, and G. She, "Estimation of forest biomass dynamics in subtropical forests using multi-temporal airborne LiDAR data," *Remote Sens. Environ.*, vol. 178, pp. 158–171, Jun. 2016.
- [22] Y. Chen, X. Feng, B. Fu, H. Ma, C. M. Zohner, T. W. Crowther, Y. Huang, X. Wu, and F. Wei, "Maps with 1 km resolution reveal increases in above- and belowground forest biomass carbon pools in China over the past 20 years," *Earth Syst. Sci. Data*, vol. 15, no. 2, pp. 897–910, Feb. 2023.
- [23] J. Byakatonda, B. P. Parida, D. B. Moalafhi, and P. K. Kenabatho, "Analysis of long term drought severity characteristics and trends across semiarid Botswana using two drought indices," *Atmos. Res.*, vol. 213, pp. 492–508, Nov. 2018.
- [24] J. Yin and S. Sun, "Incomplete multi-view clustering with cosine similarity," *Pattern Recognit.*, vol. 123, Mar. 2022, Art. no. 108371.
- [25] D. A. Zaitsev, "A generalized neighborhood for cellular automata," *Theor. Comput. Sci.*, vol. 666, pp. 21–35, Mar. 2017.



ZHENYU YU (Member, IEEE) received the Ph.D. degree in geographic information systems (GIS), in June 2022. She was a Postdoctoral Researcher with the School of Geographical Sciences and Remote Sensing, Guangzhou University, and the Remote Sensing Big Data Intelligent Application Innovation Center. She is currently a Researcher with Universiti Malaya. She has published more than 30 papers in top international conferences and journals, such as AAAI, WR, WRR, and JOH. Her research interests include AI for science, computer vision, and remote sensing. She serves as a Program Committee Member for IJCAI, contributing to the review and selection process of high-quality research in artificial intelligence.



JINLIAN WANG received the B.Sc. degree in physical geography from Peking University, in 1987, the M.Sc. degree from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, in 1997, and the Ph.D. degree in computer science from Aix-Marseille University, France, in 2015. He is an Expert in remote sensing and spatial information technology and an Academician with the International Academy of Astronautics (IAA). He is currently a Professor with the School of Geographical Sciences and Remote Sensing, Guangzhou University, and the Director of the Remote Sensing Big Data Intelligent Application Innovation Center. He previously was the Deputy Director, a Researcher, and a Ph.D. Supervisor with the Institute of Remote Sensing and Digital Earth, CAS, and the Executive Deputy Director of the National Engineering Laboratory for Remote Sensing Satellite Applications. His research interests include remote sensing big data, spatial information processing, and their applications.



HANQING CHEN received the Ph.D. degree in geodesy and geomatics engineering from Hohai University, in 2021. He was a Postdoctoral Researcher with the School of Geographical Sciences and Remote Sensing, Guangzhou University. He is currently an Associate Professor with Guangdong Ocean University. He has led projects funded by China Postdoctoral Science Foundation and a key laboratory and has contributed to national research programs. He has published 21 articles in leading international and domestic journals, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, HESS, and JHM, with ten as the first author, including four CAS TOP articles. His research interests include hydrological remote sensing, global change, and the development and application of high-precision global precipitation products. He independently developed the MSMGPE global land precipitation dataset (0.25°/1-hour). He is also a member of the Space Earth Science Professional Committee, Chinese Society of Space Science. He has received multiple honors, including the Second Prize of the Jiangxi Remote Sensing Science and Technology Award, in 2018, and the three National Graduate Scholarships.



MOHD YAMANI IDNA IDRIS is currently a Professor with Universiti Malaya. He has been a Faculty Member, since 2000, contributing more than two decades to education, research, and innovation. He has published extensively in reputed journals and successfully supervised 25 Ph.D. candidates. He actively participates in grant evaluation panels and Ph.D. examinations worldwide. He has held key leadership roles, including the Head of Department and the Deputy Dean, driving research and academic excellence. His research interests include the Internet of Things (IoT) and image processing, with significant advancements in computer vision and pattern recognition. His contributions have earned him multiple excellence service awards and recognition among Stanford's Top 2% Scientists, in 2024. Beyond research, he serves as a Reviewer and an Editorial Board Member for journals, such as *Pattern Recognition* and *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*.