A Registration-Aware Spatial-Spectral Diffusion Framework for Unregistered Hyperspectral-Multispectral Image Fusion

Peiran Liang[1,2], Yifei Zhao[1,2], Jiexiao Peng[1,2],
Qiman Li[1,2], Jiaxin Yu[1,2], and Jin Hu[1,2,3,4,*]
[1]College of Information Engineering, Northwest A&F University, Yangling, Shaanxi, China
[2]Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, Yangling, Shaanxi, China
[3]Shaanxi Engineering Research Center of Agricultural Information
Intelligent Perception and Analysis, Yangling, Shaanxi, China
[4]State Key Laboratory for Crop Stress Resistance and High-Efficiency Production, Yangling, Shaanxi, China
Email: {brockliang, ez4chen, jiexiaopeng, maztina, yvjiaxin, hujin007}@nwafu.edu.cn

***Abstract*—Hyperspectral and multispectral image fusion aims to combine a low-resolution hyperspectral image (Lr-HSI) with a high-resolution multispectral image (Hr-MSI) to generate a high-resolution hyperspectral image (Hr-HSI). Existing methods predominantly rely on fixed degradation distributions during training; real-world data acquired from disparate sensors often exhibit pose and photometric misalignments that severely degrade fusion quality. To address this, we propose SSRAF-Diff, a diffusion-based registration-aware spatial-spectral fusion framework. Adopting a dual-branch collaborative architecture with a Pyramid Attention-Guided Denoising (PAGD) module as the core generative unit, the framework simultaneously learns spectral fidelity and spatial detail recovery during the iterative denoising process. Specifically, the spectral branch models consistency to provide stable references, while the spatial branch enhances structural textures via a selective detail injection mechanism to suppress noise. Furthermore, we introduce a Spatial Deformation Registration Fusion (SDRF) module that estimates deformation fields from cross-modal similarity maps, enabling robust alignment and fusion despite geometric and photometric inconsistencies. Extensive experiments demonstrate that SSRAF-Diff outperforms various state-of-the-art (SOTA) methods.**

***Index Terms*—Hyperspectral image, spatial-spectral fusion, image registration, diffusion model, hyperspectral image super-resolution**

I. Introduction

Hyperspectral images (HSIs), characterized by abundant contiguous spectral bands, are invaluable in plant phenotyping [1], precision agriculture [2], and remote sensing [3]. However, due to hardware constraints, hyperspectral sensors cannot simultaneously achieve high spatial and spectral resolutions, which limits their practical application. Recently, computational imaging methods—such as spectral unmixing [4], spectral super-resolution [5], and spatial-spectral fusion [6]—have gained significant attention for reconstructing high-spatial-resolution HSIs (Hr-HSIs). To mitigate the low spatial resolution of HSIs, single-image super-resolution (SISR) has been explored but often suffers from hallucinated or lost details in complex textures due to its reliance on internal priors [7]. Conversely, spatial-spectral fusion, which combines a low-resolution HSI (Lr-HSI) with a high-resolution multispectral image (Hr-MSI), has proven particularly effective in maintaining both spatial and spectral integrity [8]. Therefore, this study focuses on the fusion-based paradigm.

Current HSI-MSI fusion methods are primarily divided into model-driven and deep learning-based approaches. Traditional methods rely on handcrafted priors to model mappings between Lr-HSI, MSI, and Hr-HSI, yet they are limited in handling high-dimensional, non-linear data and lack the capacity to capture complex features [9]. Recently, deep learning methods have made significant strides. While CNNs excel at local feature extraction and Transformers demonstrate superior global modeling [10], they often introduce distortions in high-frequency textures or subtle spectral features [11]. To further improve quality, Diffusion Models (DMs) have been introduced for their stable approximation of complex distributions through iterative denoising. Despite their advantages in detail recovery, most existing DM-based methods assume perfect spatial registration between Lr-HSI and Hr-MSI [12].

However, in real-world scenarios, the theoretical assumption of perfect alignment collapses. Disparate sensor configurations and dynamic acquisition environments inevitably introduce complex geometric distortions and photometric inconsistencies [13]. Current methodologies typically attempt to mitigate this via a sequential pipeline, yet this reveals a fundamental bottleneck: the decoupled nature of the prevailing "registration-then-fusion" paradigm. Such naive tandem approaches treat registration as an isolated pre-processing step, ignoring the intrinsic correlation between spatial alignment and spectral reconstruction. Consequently, distinct pose biases are often misinterpreted as textural differences, leading to irreversible error accumulation. Furthermore, without a feedback mechanism from the reconstruction stage, these methods fail to dynamically calibrate the deformation field, rendering them brittle when confronting non-rigid transformations and severe

*Corresponding author

spectral discrepancies [14].

To address these systemic limitations, we propose SSRAF-Diff, a registration-aware spatial-spectral fusion framework that shifts the paradigm from "static cascading" to "dynamic co-evolution." We conceptualize pose misalignment as a form of structured noise and synchronize the registration process within the diffusion denoising iterations. Specifically, to eliminate error accumulation caused by decoupled registration, we design a Spatial Deformation Registration Fusion (SDRF) module. Unlike static pre-processors, SDRF estimates dense deformation fields from cross-modal similarity maps in real-time, utilizing feature-level consistency constraints to dynamically calibrate geometric distortions and ensure that heterogeneous features converge toward a consistent manifold. Simultaneously, to resolve the conflict between noise suppression and detail preservation, we construct a dual-branch architecture centered on a Pyramid Attention-Guided Denoising (PAGD) module. The spectral branch models consistency to provide stable references, while the spatial branch employs a Selective Detail Injection Gating (SDIG) mechanism to differentiate between authentic structural details and misalignment artifacts, adaptively controlling the incorporation of high-frequency textures. In summary, our main contributions are as follows:

- We propose a co-evolutionary registration-fusion paradigm. By embedding the designed Spatial Deformation Registration Fusion (SDRF) module into the diffusion process, we achieve closed-loop optimization of geometric correction and spectral reconstruction. This effectively addresses the limitations of decoupled registration and fusion, eliminating irreversible error accumulation in multi-source observations.
- We construct a dual-branch progressive mechanism centered on Pyramid Attention-Guided Denoising (PAGD). Through the Selective Detail Injection Gating (SDIG), we adaptively regulate the intensity of high-frequency texture integration, thereby enhancing spatial resolution while strictly preserving spectral fidelity.
- Extensive experiments demonstrate that the proposed method outperforms existing state-of-the-art (SOTA) methods across three public datasets, exhibiting superior robustness, particularly in challenging scenarios involving non-rigid deformations and non-strict alignment.

II. Related Work

A. Deep Learning-based Hyperspectral Image Fusion

Deep learning has become the mainstream approach for hyperspectral image (HSI) fusion due to its strong ability to model complex spectral–spatial correlations. Early methods commonly combined dimensionality reduction techniques, such as PCA [15], with 3D convolutional neural networks (CNNs) to jointly extract spectral and spatial features. To improve physical consistency and robustness, subsequent studies incorporated observation models into network design or synthesized physics-aware training data, forming model-based deep fusion frameworks [16]. Despite their effectiveness, most CNN-based methods are limited by local receptive fields, which restrict their capability to capture long-range dependencies and global contextual information [17].Recently, the Mamba architecture, derived from state space models (SSMs) [18], has attracted attention for its global modeling capability with linear complexity. However, its inherent one-dimensional causal scanning mechanism is not well suited to the non-causal and multi-dimensional nature of hyperspectral images [19], often leading to insufficient preservation of local details and limited exploitation of inter-band correlations.

In parallel, diffusion probabilistic models (DPMs) have demonstrated strong performance in image restoration and generation [20], benefiting from stable training and powerful generative priors. Recent HSI fusion works formulate spatial–spectral super-resolution as a diffusion process [21], significantly improving reconstruction robustness. These advances indicate that generative diffusion models provide a promising alternative for hyperspectral image fusion beyond conventional discriminative architectures.

B. Image Registration

Image registration aims to correct spatial misalignment among multi-source observations and is a crucial prerequisite for accurate image fusion. Traditional approaches rely on handcrafted features and interpolation-based deformation models, which are often computationally expensive and prone to suboptimal solutions under complex non-rigid transformations.Deep learning has substantially advanced this field. The Spatial Transformer Network (STN) introduced learnable parametric transformations within neural networks [22], while later unsupervised frameworks, such as VoxelMorph [23], enabled efficient end-to-end non-rigid registration by directly regressing dense deformation fields. To further handle large and complex deformations, cascaded coarse-to-fine strategies have been adopted, progressively refining alignment accuracy across multiple stages [24].

III. Proposed Method

A. Problem Formulation

Let $\mathbf{X} \in \mathbb{R}^{B\times h\times w}$ denote the input low-resolution hyperspectral image (Lr-HSI) and $\mathbf{Y} \in \mathbb{R}^{b\times H\times W}$ represent the high-resolution multispectral image (Hr-MSI). Here, $\{B, b\}$ signify the number of spectral bands, while $\{h, H\}$ and $\{w, W\}$ denote the spatial height and width, respectively. Typically, $b \ll B$ and $h \ll H, w \ll W$. The ultimate goal is to reconstruct the latent high-resolution hyperspectral image (Hr-HSI), denoted as $\mathbf{Z} \in \mathbb{R}^{B\times H\times W}$.

In ideal scenarios, the degradation models are expressed as:

$$\mathbf{X} = \mathbf{Z}\mathbf{K}\downarrow_s, \quad \mathbf{Y} = \mathbf{R}\mathbf{Z} \tag{1}$$

where $\mathbf{K}$ is the spatial blurring kernel, $\downarrow_s$ is the downsampling operator, and $\mathbf{R}$ is the spectral response matrix.

However, in real-world scenarios with disparate sensors, we define a composite geometric transformation operator $\mathcal{W}(\cdot)$ (encapsulating affine, perspective, and elastic deformations) and a photometric perturbation operator $\mathcal{P}(\cdot)$ (modeling color

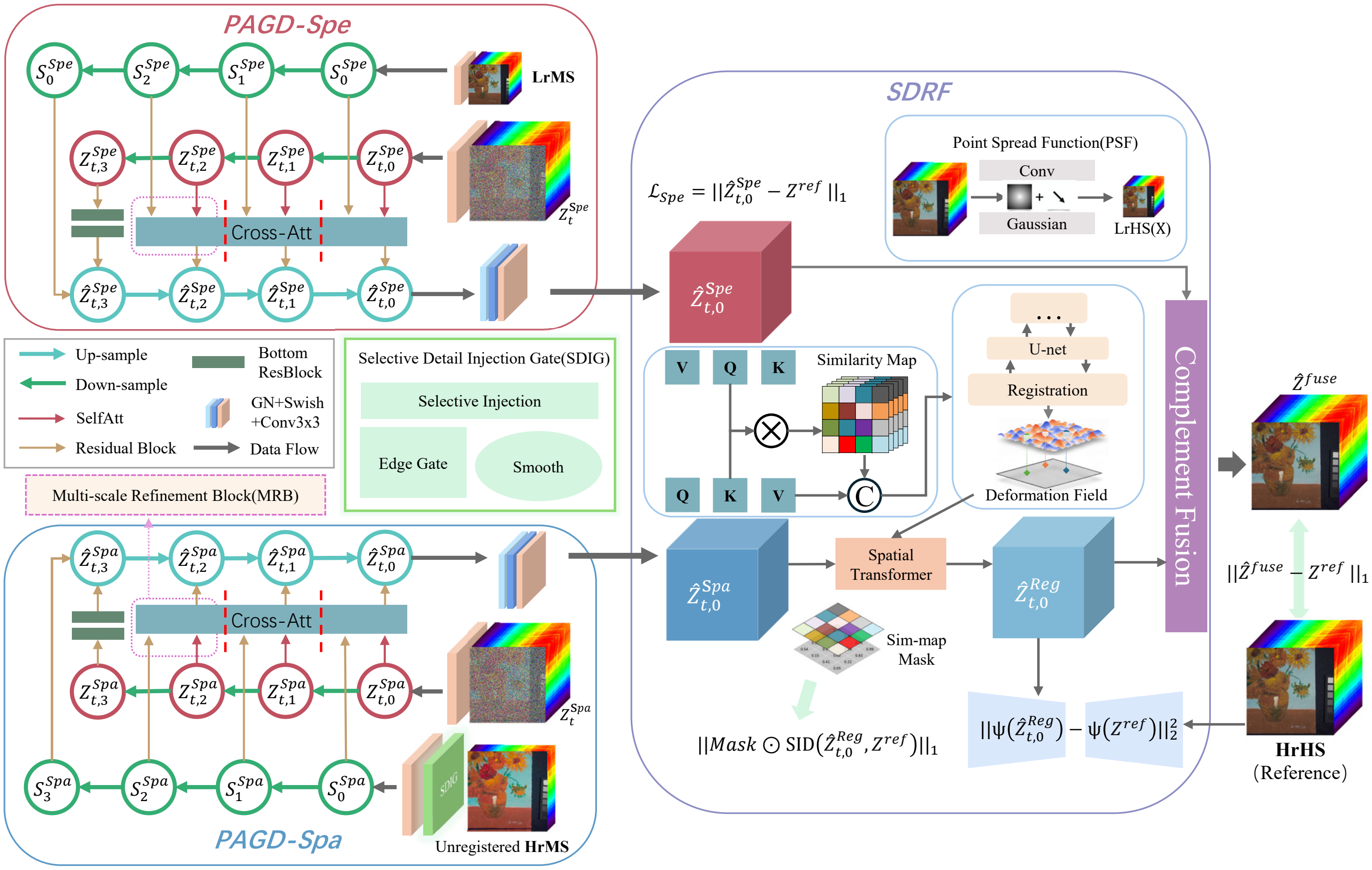

Fig. 1. The overall architecture of the proposed SSRAF-Diff framework.

temperature and saturation shifts). The realistic degradation for $\mathbf{Y}$ is reformulated as:

$$\mathbf{Y} = \mathbf{R} \cdot \mathcal{P}\left(\mathcal{W}\left(\mathbf{Z}\right)\right) \tag{2}$$

Our goal is to solve for an estimated HR-HSI, denoted as $\hat{\mathbf{Z}}$, which fully integrates the spatial details of $\mathbf{Y}$ and the spectral information of $\mathbf{X}$ despite the unregistration. We formulate the reconstruction objective as a superposition problem:

$$\hat{\mathbf{Z}} = \alpha\left[\mathcal{W}^{-1}\left(\mathcal{P}^{-1}(\mathbf{R}^{-1}\mathbf{Y})\right)\right] + \beta\left[\mathbf{K}^{-1}(\mathbf{X})\uparrow_s\right] \tag{3}$$

where $\mathbf{R}^{-1}$ denotes the pseudo-inverse of the spectral response, $\mathbf{K}^{-1}(\cdot)\uparrow_s$ represents the spatial upsampling operation corresponding to Eq. (1), and $\mathcal{W}^{-1}, \mathcal{P}^{-1}$ denote the inverse operations of the geometric and photometric distortions, respectively. The weighting parameters $\alpha, \beta \in [0,1]$ satisfy $\alpha + \beta = 1$.

Unlike ideal cases where a single solution exists, the unregistered nature implies that satisfying only the spatial constraint of $\mathbf{Y}$ or the spectral constraint of $\mathbf{X}$ yields incompatible results. Therefore, the optimal $\hat{\mathbf{Z}}$ is theoretically defined as a superposition of the two modal priors, conforming to the high-frequency spatial structure of $\mathbf{Y}$ while preserving the spectral manifold of $\mathbf{X}$ under the learned deformation field.

B. Overall Framework

Based on the realistic degradation scenarios established in the previous section, we propose SSRAF-Diff, a spectral-spatial collaborative registration-aware diffusion framework. The overall architecture is illustrated in Fig. 1. To address the discrepancy between spatial resolution and spectral information, we design a diffusion-based backbone to recover comprehensive Hr-HSIs from stochastic noise.

In the forward diffusion process, Gaussian noise is gradually added to the clean ground-truth Hr-HSI $\mathbf{Z}_0 \sim q(\mathbf{Z}_0)$ through a Markov chain. Using the reparameterization technique, the noisy state $\mathbf{Z}_t$ at any timestep $t \in [0, T]$ is obtained.

In the reverse denoising process, we propose a dual-branch collaborative mechanism centered on the Pyramid Attention-Guided Denoising (PAGD) module. Within this structure, multiple Multi-scale Refinement Blocks (MRB) are employed at each timestep, ensuring that fine-grained textures are selectively transferred to the high-fidelity spectral representation to obtain comprehensive spectral and spatial information.

Finally, the spatial deformation registration fusion (SDRF) module is introduced to perform feature-level registration and complementary fusion.

C. Pyramid Attention-Guided Denoising (PAGD)

As the core component of the diffusion framework, PAGD utilizes a collaborative input strategy. First, a unified initial feature is constructed by concatenating the noisy state $\mathbf{Z}_t$, the LrHSI $\mathbf{X}$, and the HrMSI $\mathbf{Y}$ as the starting point for both branches:

$$\mathbf{Z}_{t,0}^{Spa} = \mathbf{Z}_{t,0}^{Spe} = \text{Conv}_{3\times3}([\mathbf{Z}_t, \mathbf{X}, \mathbf{Y}]) \tag{4}$$

The HrMSI and LrHSI serve as guidance factors respectively:

$$\mathbf{S}_0^{Spa} = \text{SDIG}(\mathbf{Y}), \quad \mathbf{S}_0^{Spe} = \mathbf{X} \tag{5}$$

Crucially, to resolve the conflict between authentic detail preservation and misalignment artifact suppression, we design the SDIG as a pre-injection gate for the spatial branch. By differentiating between structural details and pose biases, it adaptively regulates the high-frequency information flow to refine the guidance $\mathbf{Y}$ into $\mathbf{Y}'$ before multi-scale fusion:

$$\mathbf{Y}' = \mathbf{Y} + \alpha s_t(\tilde{\mathbf{Y}} - \mathbf{Y}) \odot \exp(-\beta|\nabla \mathbf{Y}|) \tag{6}$$

where $\tilde{\mathbf{Y}}$ is a smoothed version of $\mathbf{Y}$ and s_t is the timestep-dependent intensity. This design filters unreliable spatial details during the denoising stage. The dual-stream denoising process is then formulated as:

$$\hat{\mathbf{Z}}_{t,0}^{Spa} = \text{PAGD-Spa}\big(\mathbf{Z}_{t,0}^{Spa}, \mathbf{S}_0^{Spa}\big) \tag{7}$$
$$\hat{\mathbf{Z}}_{t,0}^{Spe} = \text{PAGD-Spe}\big(\mathbf{Z}_{t,0}^{Spe}, \mathbf{S}_0^{Spe}\big) \tag{8}$$

The dual branches consist of multiple multi-scale refinement blocks (MRB), indexed by l. As shown in Fig. 2, at each scale θ of the dual-branch pyramid U-Net, the l-th MRB integrates the decoding feature $\hat{\mathbf{Z}}_{t,\theta}^{Spa}$ with the fused attention feature $\mathbf{F}_{att}^l$ as follows:

$$\begin{cases} \hat{\mathbf{Z}}_{t,\theta}^{Spa,l} = \text{MRB}_t^l(\hat{\mathbf{Z}}_{t,\theta+1}^{Spa}, \mathbf{F}_{att}^l) \\ \mathbf{F}_{att}^l = \mathcal{R}\left(\left\{ \begin{aligned} &(\mathcal{R}(\mathbf{S}_\theta^l) + \text{Softmax}\left(\frac{\mathbf{Q}_S^l(\mathbf{K}_S^l)^T}{\sqrt{d}}\right)\mathbf{V}_S^l); \\ &(\mathcal{R}(\mathbf{Z}_{t,\theta}^l) + \text{Softmax}\left(\frac{\mathbf{Q}_S^l(\mathbf{K}_Z^l)^T}{\sqrt{d}}\right)\mathbf{V}_Z^l) \end{aligned} \right\}\right) \\ \mathbf{Q}_S^l, \mathbf{K}_S^l, \mathbf{V}_S^l = \phi(\mathbf{S}_\theta^l), \quad \mathbf{K}_Z^l, \mathbf{V}_Z^l = \rho(\mathbf{Z}_{t,\theta}^l) \end{cases} \tag{9}$$

where θ denotes the specific scale within the U-Net architecture and l represents the sequential index of the refinement block. In this formulation, the attention feature $\mathbf{F}_{att}^l$ is constructed by applying a residual-based fusion $\mathcal{R}(\cdot)$ on the concatenated results of self-attention (derived from guidance $\mathbf{S}_\theta^l$) and cross-attention (probing the noisy manifold $\mathbf{Z}_{t,\theta}^l$).

D. Spatial Deformation Registration Fusion (SDRF)

To eliminate the irreversible error accumulation caused by static pre-alignment, we design the SDRF module to perform dynamic geometric calibration within each noise level t. Unlike decoupled approaches, SDRF treats the spectral feature $\hat{\mathbf{Z}}_{t,0}^{Spe}$ as the reference and the spatial feature $\hat{\mathbf{Z}}_{t,0}^{Spa}$ as the target, continuously refining the alignment based on the evolving

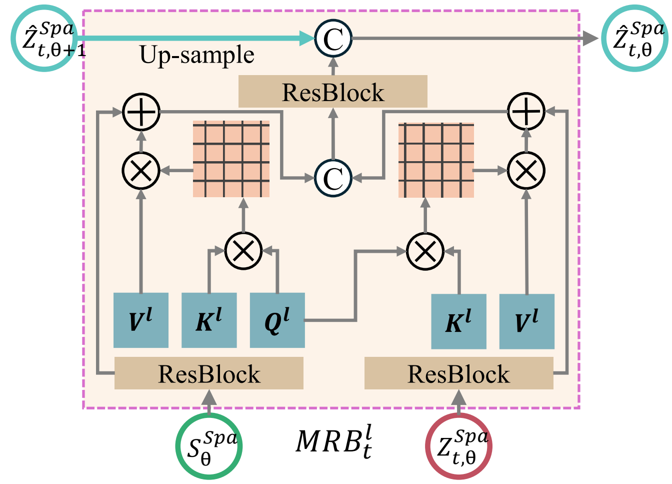

Fig. 2. Illustration of the l^{th} multi-scale refinement block (MRB) of spatial pyramid attention-guided denoising module (PAGD-Spa).

reconstruction state. For a specific scale i, the process is formulated as follows:

$$\phi^i = f_D\left(\text{concat}\left(\mathbf{Q}^i(\mathbf{K}^i)^T, \mathbf{V}^i\right)\right) \tag{10}$$
$$\hat{\mathbf{Z}}_t^{fuse} = f_{fuse}\left(\hat{\mathbf{Z}}_{t,0}^{Spa,i} \otimes \phi^i, \hat{\mathbf{Z}}_{t,0}^{Spe}\right) \tag{11}$$

As defined in Eq. (10), the dense deformation field ϕ^i is regressed from the cross-modal similarity matrix (computed via $\mathbf{Q}^i(\mathbf{K}^i)^T$) and value features. Subsequently, Eq. (11) performs spatial alignment using the warping operator $\otimes$ and fuses the registered features with the spectral reference. After T iterations, the final high-fidelity HrHSI is reconstructed.

E. Training Objectives

To train the proposed framework, we define a multi-task loss function to constrain both the final fusion result and the intermediate branch outputs. The total loss $\mathcal{L}_{total}$ consists of the following components:

1) Global Reconstruction Loss: We employ the ℓ_1 norm to measure the final fused result:

$$\mathcal{L}_{fuse} = \|\hat{\mathbf{Z}}_t^{fuse} - \mathbf{Z}_{ref}\|_1 \tag{12}$$

2) Branch-Specific Loss: To ensure the spectral branch maintains spectral fidelity during the denoising process, we introduce a fidelity loss $\mathcal{L}_{spe}$:

$$\mathcal{L}_{spe} = \|\hat{\mathbf{Z}}_{t,0}^{Spe} - \mathbf{Z}_{ref}\|_1 \tag{13}$$

3) Registration-Aware Similarity and Alignment Loss: To supervise the aligned spatial features, ensure cross-modal consistency, and focus training constraints on misaligned regions, we design the $\mathcal{L}_{\text{RASAL}}$:

$$\mathcal{L}_{\text{RASAL}} = \mathcal{L}_{\text{SID}}^{\text{mask}} + \mathcal{L}_{\text{feat}} + \mathcal{L}_{\text{smooth}} \tag{14}$$

where $\mathcal{L}_{\text{SID}}^{\text{mask}} = \|\mathbf{M} \odot \text{SID}(\hat{\mathbf{Z}}_{t,0}^{reg}, \mathbf{Z}_{ref})\|_1$ is the Spectral Information Divergence (SID) loss modulated by the registration confidence mask $\mathbf{M}$ to suppress gradients in mismatched areas.

TABLE I
QUANTITATIVE COMPARISON ON UNREGISTERED DATASETS (4× SR). BEST RESULTS ARE IN **BOLD**, SECOND-BEST ARE UNDERLINED.

Method	CAVE				Harvard				ICVL			
	PSNR ↑	*SAM* ↓	*ERGAS* ↓	*SSIM* ↑	*PSNR* ↑	*SAM* ↓	*ERGAS* ↓	*SSIM* ↑	*PSNR* ↑	*SAM* ↓	*ERGAS* ↓	*SSIM* ↑
CNMF	28.0289	8.3010	6.3987	0.8592	28.3953	9.6361	6.6015	0.8604	29.1892	6.7080	3.4592	0.9335
PSRT	44.5194	3.4598	2.0989	0.9838	44.5682	4.1463	2.4556	0.9731	46.6836	0.8390	0.3296	0.9928
FusionMamba	42.8205	4.1608	2.1826	0.9740	43.0596	5.6169	2.7614	0.9628	46.7106	0.8778	0.3123	0.9906
S2CycleDiff	41.4115	5.9241	2.3693	0.9695	42.0969	6.4592	3.0833	0.9661	46.4868	0.8654	0.2857	0.9856
M2DTN	45.3785	3.0837	1.8218	0.9839	45.3205	4.7292	2.1466	0.9826	46.0357	1.0962	0.2997	0.9884
DSPNet	45.7413	2.9079	1.5171	0.9896	46.3849	4.4224	2.0272	0.9806	47.9888	**0.8075**	0.2948	0.9916
LRTN	47.9689	2.8523	1.4192	0.9874	46.7233	4.0760	1.9922	0.9825	48.7751	0.8533	0.2879	0.9925
SRLF	46.7131	2.7394	1.4138	0.9906	45.7294	4.0631	2.4701	0.9718	49.5021	0.8476	0.2703	0.9955
SSRAF (Ours)	**49.0515**	**2.7012**	**1.3152**	**0.9942**	**47.7886**	**4.0527**	**1.9509**	**0.9866**	**52.0595**	0.8226	**0.2674**	**0.9962**

The term $\mathcal{L}_{\text{feat}} = \|\psi(\hat{\mathbf{Z}}_{t,0}^{reg}) - \psi(\mathbf{Z}_{ref})\|_2^2$ ensures feature-level consistency, where $\psi(\cdot)$ denotes a pre-trained feature extractor. Additionally, $\mathcal{L}_{\text{smooth}}$ constrains the spatial smoothness of the deformation field.

The final training objective is:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{fuse} + \lambda_2 \mathcal{L}_{spe} + \lambda_3 \mathcal{L}_{\text{RASAL}} \tag{15}$$

IV. EXPERIMENTS

A. Experimental Settings

Datasets: We evaluate the effectiveness of the proposed SSRAF on three publicly available datasets: CAVE [25], Harvard [26], and ICVL [27]. All three datasets consist of hyperspectral images with 31 spectral bands, with spatial resolutions of 512×512, 2704×3376, and 1300×1392, respectively. For the Harvard and ICVL datasets, each image is center-cropped into two or four 512×512 patches, respectively. The CAVE dataset contains 32 images, from which we select 20 for training and 12 for testing. The Harvard dataset includes 50×2 patches, with 80 used for training and 20 for testing. The ICVL dataset comprises 200×4 patches, with 640 allocated for training and 160 for testing.

Competing Methods and Evaluation Metrics: We compare SSRAF against several state-of-the-art (SOTA) methods designed for unregistered image fusion, including SRLF [28], M2DTN [29], LRTN [30], DSPNet [31], and FusionMamba [32]. Additionally, for comprehensive evaluation, we include fusion methods originally designed for registered images, such as S2CycleDiff [21] and PSRT [33], as well as the traditional baseline method CNMF [34]. To ensure consistency, all deep learning-based methods are trained using the same dataset configurations. We select four quantitative metrics to evaluate the reconstructed high-resolution hyperspectral images (HrHSI): Peak Signal-to-Noise Ratio (PSNR), Spectral Angle Mapper (SAM), Structural Similarity Index (SSIM), and Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS).

Implementation Details: The LrHSI ($128 \times 128 \times 31$) and HrMSI ($512 \times 512 \times 3$) for the three datasets are generated via Wald's protocol [35]. Furthermore, we apply affine transformations, perspective perturbations, elastic deformations, and color temperature/white balance perturbations to the HrMSI to simulate the spatial and photometric discrepancies defined in Eq. (3). The experiments are implemented using the PyTorch framework and trained on four NVIDIA GeForce A40 GPUs. For all datasets, the batch size is set to 8, and the models are trained for 120,000 iterations. We use the Adam optimizer with a learning rate of 0.0001. The total diffusion timesteps are set to $T = 2000$. For the total loss function, the weighting coefficients are set to $\lambda_1 = 1$, $\lambda_2 = 0.3$, and $\lambda_3 = 0.1$.

TABLE II
ABLATION STUDY OF THE PAGD MODULE.

Variants	PSNR ↑	SAM ↓	ERGAS ↓	SSIM ↑
CAVE Dataset				
w/o Guidance	34.7334	4.5366	4.5626	0.9567
w/o Cross-Attn	48.9917	2.7607	1.4022	0.9935
w/o SDIG	48.8635	2.7993	1.4118	0.9931
SSRAF	**49.0515**	**2.7012**	**1.3152**	**0.9942**
Harvard Dataset				
w/o Guidance	32.8067	5.6125	3.6422	0.9356
w/o Cross-Attn	47.4599	4.1328	1.9571	0.9852
w/o SDIG	47.3862	4.5307	1.9656	0.9830
SSRAF	**47.7886**	**4.0527**	**1.9509**	**0.9866**

B. Quantitative and Qualitative Results

Quantitative Results: Table I reports the objective evaluation results for the fusion task. Our SSRAF consistently achieves the best performance across all metrics, significantly outperforming the comparison methods. These results demonstrate the superior ability of SSRAF to mitigate spatial misalignments while preserving high-fidelity fusion quality.

Qualitative Results: To visually evaluate the reconstruction quality, we provide pseudo-color images and Mean Absolute Error (MAE) residual maps of samples randomly selected from the three datasets in Fig. 3. As illustrated in the magnified regions, our method recovers sharper texture details and exhibits superior spectral fidelity. Compared to other methods, the residual maps generated by SSRAF exhibit the lowest intensity, confirming its superior capability in reconstructing fine spatial structures and maintaining spectral consistency.

C. Ablation studies

To verify the effectiveness of the proposed modules, we conducted extensive ablation experiments on two datasets. The quantitative results are summarized in Table II.

1) Pyramid Attention Guided Denoising (PAGD) Module: To validate the impact of the guidance conditions, we configured a dual-branch setup with unconditional inputs, where the

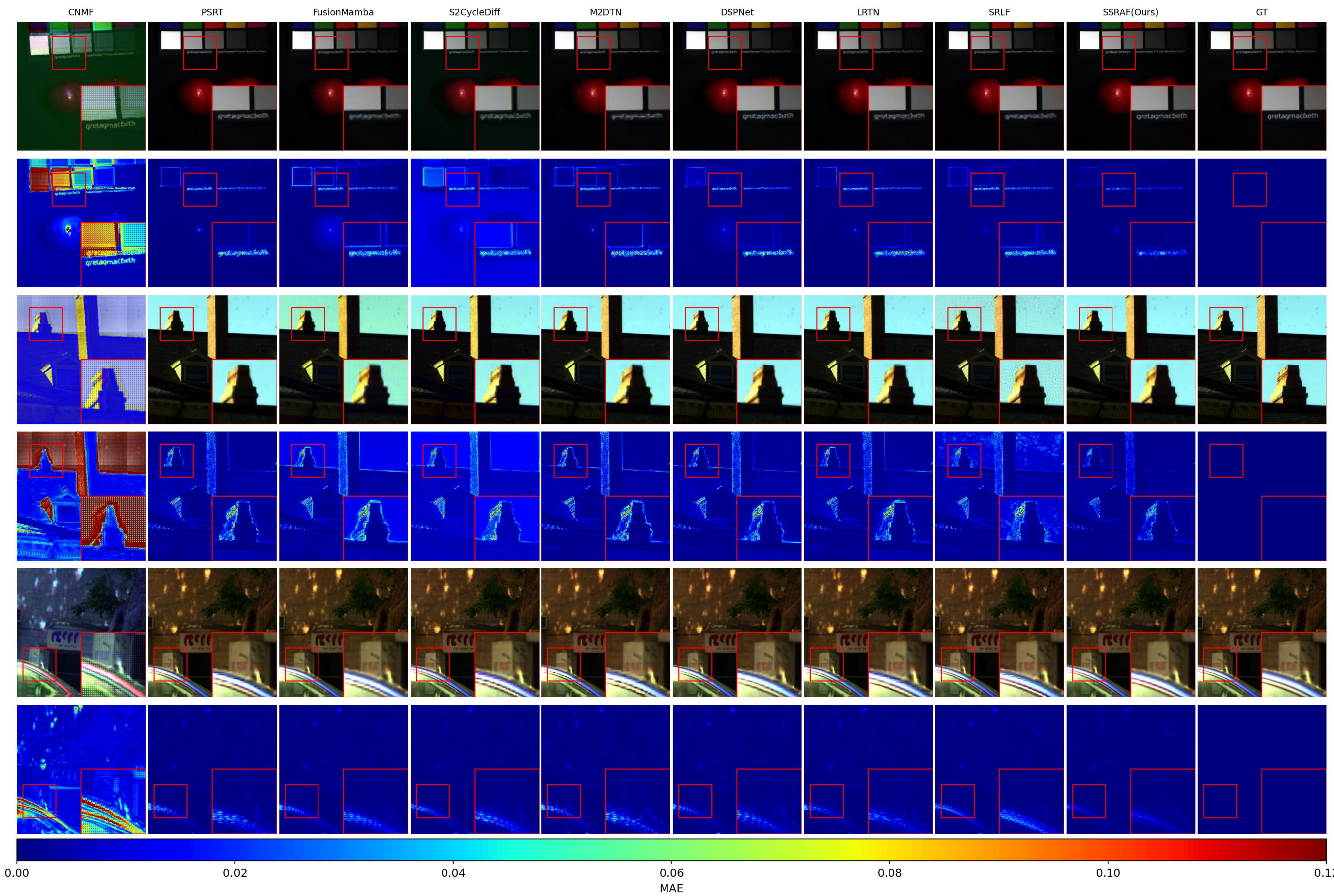

Fig. 3. Visual results and reconstruction error maps of different methods on three unregistered datasets. Darker error maps indicate better reconstruction performance.

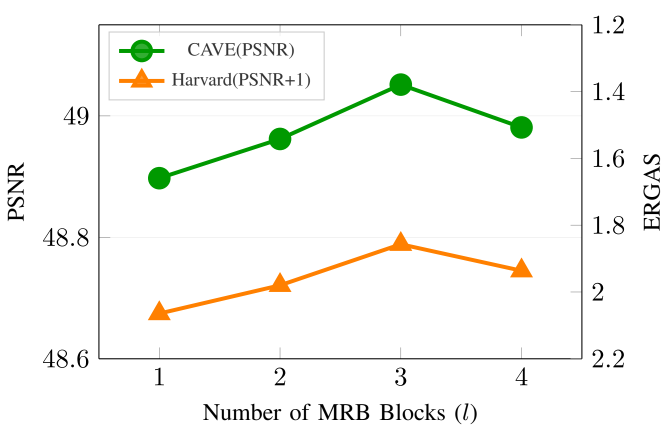

Fig. 4. Ablation of MRB block count (l). Both datasets reach the peak at $l = 3$.

cross-attention in the Multi-scale Refinement Block (MRB) was simplified to self-attention on input features while maintaining other structures. As shown in Table II, the multi-scale conditional input significantly enhances spatial details and spectral fidelity, demonstrating the superiority of the conditional diffusion framework in generating high-resolution hyperspectral images (HrHSI).

2) Multi-scale Refinement Block (MRB): To further evaluate the capability of the attention mechanism in feature extraction, we implemented a baseline without the attention mechanism (where guidance conditions are directly concatenated with intermediate features). Table II indicates that our attention-based MRB extracts finer spectral and spatial information. Additionally, we ablated the number of MRBs. As illustrated in Fig. 4, the optimal performance is achieved when the number of blocks is set to 3.

3) Selective Detail Injection Gating (SDIG): To verify the ability to adaptively control the intensity of high-frequency texture injection, we replaced the SDIG with a standard convolutional layer. The results in Table II show that SDIG effectively suppresses interference from unreliable details during reconstruction, thereby improving the overall spatial quality.

V. Conclusion

This paper presents SSRAF-Diff, a novel diffusion-based registration-aware framework designed to simultaneously address geometric and photometric misalignments in unregistered hyperspectral and multispectral image fusion. By employing a dual-branch collaborative architecture centered on the proposed Pyramid Attention-Guided Denoising (PAGD) module, the framework effectively preserves global spectral consistency while recovering fine-grained spatial details throughout the iterative reverse diffusion process. Furthermore, the integrated Selective Detail Injection Gating (SDIG) mechanism adaptively regulates high-frequency texture infusion to suppress noise amplification, while the Spatial Deformation Registration Fusion (SDRF) module significantly enhances structural robustness against complex spatial discrepancies. Extensive quantitative and qualitative experiments on three public datasets demonstrate that SSRAF-Diff consistently outperforms existing state-of-the-art methods, establishing a robust and highly generalizable solution for challenging fusion tasks under complex misalignment scenarios.

Acknowledgment

This work was supported by the Science and Technology Project of the Ministry of Agriculture and Rural Affairs, Open Project of the Arid Zone Agriculture Shaanxi Laboratory Construction (Grant No. 2024ZY-JCYJ-02-16), and Artificial Intelligence Platform of the College of Information Engineering, Northwest A&F University.

References

[1] M. D. Raya-Sereno, M. Alonso-Ayuso, J. L. Pancorbo, J. L. Gabriel, C. Camino, P. J. Zarco-Tejada, and M. Quemada, "Residual effect and n fertilizer rate detection by high-resolution vnir-swir hyperspectral imagery and solar-induced chlorophyll fluorescence in wheat," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[2] J. Jia, J. Chen, X. Zheng, Y. Wang, S. Guo, H. Sun, C. Jiang, M. Karjalainen, K. Karila, Z. Duan, T. Wang, C. Xu, J. Hyyppä, and Y. Chen, "Tradeoffs in the spatial and spectral resolution of airborne hyperspectral imaging systems: A crop identification case study," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.

[3] X. Guo, X. Huang, L. Zhang, L. Zhang, A. Plaza, and J. A. Benediktsson, "Support tensor machines for classification of hyperspectral remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 6, pp. 3248–3264, 2016.

[4] Y. Yu, E. Pan, Y. Ma, X. Mei, Q. Chen, and J. Ma, "Unmixdiff: Unmixing-based diffusion model for hyperspectral image synthesis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–18, 2024.

[5] Y. Cai, J. Lin, Z. Lin, H. Wang, Y. Zhang, H. Pfister, R. Timofte, and L. Van Gool, "Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2022, pp. 745–755.

[6] A. Guo, R. Dian, and S. Li, "Unsupervised blur kernel learning for pansharpening," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020, pp. 633–636.

[7] ——, "Unsupervised blur kernel learning for pansharpening," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020, pp. 633–636.

[8] S. Mei, G. Zhang, N. Wang, B. Wu, M. Ma, Y. Zhang, and Y. Feng, "Lightweight multiresolution feature fusion network for spectral super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[9] Y. Zhang, C. Yan, F. Dai, and Y. Ma, "Efficient parallel framework for h.264/avc deblocking filter on many-core platform," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 510–524, 2012.

[10] H. Hassan, M. N. Khan, S. O. Gilani, M. Jamil, H. Maqbool, A. W. Malik, and I. Ahmad, "H.264 encoder parameter optimization for encoded wireless multimedia transmissions," *IEEE Access*, vol. 6, pp. 22 046–22 053, 2018.

[11] M. Mancini, M. F. Naeem, Y. Xian, and Z. Akata, "Learning graph embeddings for open world compositional zero-shot learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 3, pp. 1545–1560, 2024.

[12] C. Wu, D. Wang, Y. Bai, H. Mao, Y. Li, and Q. Shen, "Hsr-diff: Hyperspectral image super-resolution via conditional diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 7083–7093.

[13] J. Nie, L. Zhang, W. Wei, C. Ding, and Y. Zhang, "Unsupervised deep hyperspectral super-resolution with unregistered images," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6.

[14] J. Fang, J. Yang, A. Khader, and L. Xiao, "Deep unfolding network enhanced by transformer priors for unregistered hyperspectral and multispectral image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.

[15] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 639–643, 2017.

[16] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 11, pp. 5345–5355, 2018.

[17] J.-F. Hu, T.-Z. Huang, L.-J. Deng, H.-X. Dou, D. Hong, and G. Vivone, "Fusformer: A transformer-based fusion network for hyperspectral image super-resolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[18] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2024. [Online]. Available: https://openreview.net/forum?id=AL1fq05o7H

[19] S. Chen, L. Zhang, and L. Zhang, "Hsrmamba: contextual spatial-spectral state space model for single hyperspectral image super-resolution," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, ser. IJCAI '25, 2025. [Online]. Available: https://doi.org/10.24963/ijcai.2025/91

[20] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

[21] J. Qu, J. He, W. Dong, and J. Zhao, "S2cyclediff: Spatial-spectral-bilateral cycle-diffusion framework for hyperspectral image super-resolution," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, pp. 4623–4631, Mar. 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/28262

[22] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf

[23] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: A learning framework for deformable medical image registration," *arXiv preprint arXiv:1809.05231*, 2018. [Online]. Available: https://arxiv.org/abs/1809.05231

[24] S. Zhao, Y. Dong, E. I.-C. Chang, and Y. Xu, "Recursive cascaded networks for unsupervised medical image registration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[25] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum," *IEEE Transactions on Image Processing*, vol. 19, no. 9, pp. 2241–2253, 2010.

[26] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *CVPR 2011*, 2011, pp. 193–200.

[27] B. Arad and O. Ben-Shahar, "Sparse recovery of hyperspectral signal from natural rgb images," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 19–34.

[28] Y. Liu, J. Liu, R. Dian, and S. Li, "A selective re-learning mechanism for hyperspectral fusion imaging," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 7437–7446.

[29] W. Dong, Y. Xu, J. Qu, and S. Hou, "Learning Multi-Modal Cross-Scale Deformable Transformer Network for Unregistered Hyperspectral Image Super-resolution," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, pp. 1573–1581, Mar. 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/27923

[30] Y. Liu, R. Dian, and S. Li, "Low-Rank Transformer for High-Resolution Hyperspectral Computational Imaging," *International Journal of Computer Vision*, vol. 133, no. 2, pp. 809–824, Feb. 2025. [Online]. Available: https://link.springer.com/10.1007/s11263-024-02203-7

[31] Y. Sun, H. Xu, Y. Ma, M. Wu, X. Mei, J. Huang, and J. Ma, "Dual spatial–spectral pyramid network with transformer for hyperspectral image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[32] S. Peng, X. Zhu, H. Deng, L.-J. Deng, and Z. Lei, "FusionMamba: Efficient Remote Sensing Image Fusion with State Space Model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024, arXiv:2404.07932 [cs]. [Online]. Available: http://arxiv.org/abs/2404.07932

[33] S.-Q. Deng, L.-J. Deng, X. Wu, R. Ran, D. Hong, and G. Vivone, "Psrt: Pyramid shuffle-and-reshuffle transformer for multispectral and hyperspectral image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[34] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled Nonnegative Matrix Factorization Unmixing for Hyperspectral and Multispectral Data Fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 2, pp. 528–537, Feb. 2012. [Online]. Available: http://ieeexplore.ieee.org/document/5982386/

[35] L. Wald and T. Ranchin, "Fusion of images and raster-maps of different spatial resolutions by encrustation: An improved approach," *Computers, Environment and Urban Systems*, vol. 19, no. 2, pp. 77–87, 1995, 1GIS/LIS '95 Central Europe. [Online]. Available: https://www.sciencedirect.com/science/article/pii/019897159500014Y