

Methods for hyperspectral and multispectral images fusion based on per pixel spectral transformations and cyclic consistency

© R.R. Khodyrev, S.A. Bazhenov, T.V. Kabanova, D.A. Vrazhnov[✉], Yu.V. Kistenev

Tomsk State University, Tomsk, Russia

[✉]e-mail: vda@iao.ru

Received December 05, 2025

Revised January 12, 2026

Accepted March 27, 2026

The paper presents two approaches for hyperspectral and multispectral images fusion. The first approach is based on the WMatrixGenerator module, which generates per pixel spectral transformation matrices with quadratic correction. The second approach uses the HSIMSiCycleGAN architecture with cyclical consistency and physically motivated losses. The experimental evaluation showed the competitiveness of the proposed methods for the PSNR, SSIM, SAM and ERGAS metrics on the Real HSI/MSI/PAN dataset. The methods work on paired observations of hyperspectral and multispectral images of the same scene, but do not require high-resolution reference data.

Keywords: image fusion, hyperspectral images, multispectral images, deep learning, non-negative least squares, CycleGAN, remote sensing, neural networks.

DOI: 10.61011/EOS.2026.05.63859.77-25

1. Introduction

Hyperspectral images (HSI) have a high spectral resolution providing detailed information about material composition, but usually have low spatial resolution [1]. Multispectral images (MSI), on the contrary, feature high spatial resolution with a limited number of spectral channels. Hyperspectral cameras on unmanned aerial vehicles (UAV) have significantly expanded the hyperspectral imaging capabilities for agroecological monitoring applications [2], for example, the spatial resolution has been increased from tens of meters (satellite images) to tens of centimeters [3]. However, taking into account that survey shall cover large areas with resolution of several centimeters, it is important to provide digital processing of hyperspectral images to increase image resolution. Methods for increasing the spatial resolution of hyperspectral data (HSI) or spectral resolution of multispectral data (MSI) by means of simultaneous processing [4] are currently being actively developed. Comparative properties of MSI and HSI are listed in Table 1.

Numerous approaches to solving the HSI-MSI fusion problem have been proposed over the recent years. They can be conventionally divided into several categories: methods based on component replacement, multiscale analysis, tensor decomposition, sparse representation and more recently deep learning methods. Table 2 shows the summary of main methods with brief description of their operating principles.

Methods for fusing hyperspectral and multispectral data evolve from linear factorization approaches to tensor ones and then to model-oriented and physically-based hybrid architectures. Basic linear methods, for example, nonnegative matrix factorization (NMF) and principal component analysis (PCA), provide low computational complexity and interpretability, however, are restricted by a global linearity premise and demonstrate noise susceptibility, leading to

moderate Peak Signal-to-Noise Ratios (PSNR) and insufficiently accurate maintaining of spectral form using Spectral Angle Mapper (SAM).

Next-level methods introduce structural and statistical limitations. Thus, hyperspectral super-resolution via subspace-based regularization (HySure) uses subspace-based regularization and joint estimation of the point spread function (PSF) and spectral response function (SRF), increasing stability with a correct degradation model. Nonnegative sparse representation (NSSR) relies on sparse representation with clustering and nonnegative dictionary, improving the PSNR estimate due to the growth of computational burden.

Tensor methods, for example, Coupled Sparse Tensor Factorization (CSTF), deep Tucker decomposition network (Deep Tucker), Unsupervised Deep Tensor Network (UDTN), use a multidimensional data structure. CSTF considers modality coupling, Deep Tucker adds neural network factor approximation, and UDTN achieves high metrics without paired reference data, requiring careful setting of hyperparameters. Model-based and coupled convolutional architectures such as Model Inspired Autoencoder (MIAE), Coupled Convolutional Neural Network with Adaptive Response Function Learning (HyCoNet), Coupled Unmixing Network with a cross attention mechanism (CuCANet), unsupervised Sparse Dirichlet-Net (uSDN), Modality-Decoupled Spatial-Spectral Fusion (MossFuse) integrate a priori assumptions into learning structures. They combine interpretability and nonlinearity, however, are sensitive to latent space dimensions and spectral response calibration accuracy. Physics-based and blind approaches: Physics-Based Generative Adversarial Network (Physics GAN), Blind Estimation Generative Adversarial Network (BliEstGAN), include an explicit or estimated spectral degradation model. The former achieves high

Table 1. Comparison of HSI and MSI

Description	Multispectral images	hyperspectral images
Number of channels	3–10	>100
Channel width	Wide (70–400 nm)	Narrow (5–10 nm)
Spectral resolution	Low	High
spatial resolution	To several centimeters	To tens of meters
Amount of data	Relatively small	Very large („curse of dimensionality“)
Key concept	Distinguish general object classes	Identify materials by their spectral signature

Table 2. Existing methods for fusing HSI and MSI images

Method	Category	Brief description
Nonnegative Matrix Factorization	Linear	Nonnegative Matrix Factorization for decomposition into basis spectra
Principal Component Analysis	Linear	Principal Component technique used to identify principal variance directions
Coupled Sparse Tensor Factorization	Tensor	Coupled Sparse Tensor Factorization taking into account coupling of modalities [5]
Hyperspectral Super-resolution via (Subspace-based) Regularization	Regularization	Subspace-based regularization with evaluation of point spread function and spectral response function [6]
Nonnegative Sparse Representation	Sparse	Nonnegative sparse representation with clustering and dictionary learning [7]
deep Tucker decomposition network	Tensor and deep learning	Tucker decomposition with neural network factor approximation [8]
Model Inspired Autoencoder	Model and deep learning	Autoencoder with integrated NMF structure [9]
unsupervised Sparse Dirichlet-Net	Deep learning	Unsupervised sparse Dirichlet-Net [10]
Coupled Convolutional Neural Network with Adaptive Response Function Learning	Deep learning	Coupled convolutional network with adaptive response function [11]
Coupled Unmixing Network with a cross-attention mechanism	Deep learning	Network with a cross-attention mechanism used for modality decoupling [12]
Unsupervised Deep Tensor Network	Tensor and deep learning	Unsupervised Deep Tensor Network [13]
Modality-Decoupled Spatial-Spectral Fusion	Deep learning	Decomposition into shared and complementary modalities [4]
Blind Estimation Generative Adversarial Network	Deep learning	Two-step approach with evaluation of spectral degradation [14]
Physics-Based Generative Adversarial Network	Deep learning	Integration of physical imaging model [15]
Registration And Fusion Nonconvex Low-Rank and Group-Sparse	regularization and deep and group learning	Registration with nonlocal regularization sparsity [16]
Efficient Hyperspectral Image Super-Resolution via Kolmogorov-Arnold Networks	Deep learning	Networks based on the Kolmogorov-Arnold theorem [14]

structural similarity index (SSIM) values with dependence on model correctness, the latter reduces the requirements for knowledge about the sensor, but complicates optimization stability. Methods with advanced regularization mechanisms (Registration And Fusion Nonconvex Low-Rank and Group-Sparse, RAF-NLRGS) and new approximating paradigms (Efficient Hyperspectral Image Super-Resolution via Kolmogorov–Arnold Networks, HSR-KAN) provide the lowest SAM values due to complex regularization and flexible nonlinear structures, being accompanied by substantial resource intensity and tool immaturity.

Regardless of significant success achieved in the field of HSI and MSI fusion, many existing methods face a set of constraints such as high computational complexity, insufficient noise and sensor variability robustness, and limited ability to generalize to new data. Modern approaches require vast labeled data or accurate knowledge of sensor specifications, which hinders their application. This study proposes solutions to these problems via development of two innovative approaches: pixel-by-pixel spectral transformation method based on neural network matrix generation and cycle consistency architecture using HSI/MSI paired observations without involvement of a high resolution hyperspectral reference image. Both approaches demonstrate a balance between recovery accuracy and practical applicability in real remote sensing conditions.

The purpose of this study is to develop a HSI/MSI fusion method based on the cycle consistency architecture with physically-motivated loss. For this, the following tasks were solved.

- Comprehensive analysis of modern HSI and MSI fusion methods was performed, benefits and drawbacks of the methods were identified.
- A fusion method based on pixel-by-pixel spectral transformation using neural network matrix generation was developed and implemented with its performance estimated on real data.
- A fusion method based on cycle consistency architecture with physically-motivated loss was developed and implemented, the effect of various loss function components on quality metrics was investigated.
- Data preparation and learning procedures for the proposed methods were described to ensure repeatability of experiments.
- Experimental qualitative and quantitative estimation of the proposed approaches was performed on standard datasets, comparative analysis with the existing solutions was carried out using PSNR, SSIM, SAM and relative dimensionless global error in synthesis (ERGAS).

2. Materials and methods

2.1. Mathematical statement of problem

Fusion problem is formally stated as follows. Given that a low spatial resolution hyperspectral image $H \in \mathbb{R}^{h \times w \times L_h}$,

where $h \times w$ are spatial dimensions (height and width in pixels), L_h is the number of spectral channels (usually tens or hundreds). Also given that a high resolution multispectral image $M \in \mathbb{R}^{H \times W \times L_m}$, where $h_{\text{MSI}} \times w_{\text{MSI}}$ are spatial dimensions ($h_{\text{MSI}} \gg h$, $w_{\text{MSI}} \gg w$), L_m is the number of spectral channels ($L_m \ll L_h$, usually units). A high-quality hyperspectral image (HQHSI) $\hat{Y} \in \mathbb{R}^{h_{\text{MSI}} \times w_{\text{MSI}} \times L_h}$ having simultaneously a high spatial resolution (as that of MSI) and high spectral resolution (as that of HSI) shall be recovered.

Process of formation of observed data can be described by the following degradation models [17,18]:

$$H = D_s(Y) + n_h = E(Y) \downarrow_r + n_h,$$

$$M = D_\lambda(Y) + n_m = RY + n_m,$$

where $Y \in \mathbb{R}^{h_{\text{MSI}} \times w_{\text{MSI}} \times L_h}$ is the sought-for high spatial and spectral resolution HQHSI; $D_s(Y)$ is the spatial degradation operator, including blurring of $E(Y)$ and downsampling $\downarrow_r$ with $r = h_{\text{MSI}}/h = w_{\text{MSI}}/w$; D_λ is the spectral degradation operator approximated by the linear transformation through the sensor's spectral response matrix $R \in \mathbb{R}^{L_m \times L_h}$; n_h and n_m are additive noise.

When developing deep learning methods for the HSI–MSI fusion problem, an important concept is a receptive field — an input data area that affects computation of a particular output element. Increase in the receptive field allows the model to consider a wider spatial and spectral context for predicting, which is critical for accurate recovery of spectral characteristics and maintaining spatial structures.

Constraints of survey instruments result in a compromise between the spatial and spectral resolutions. Obtaining high spatial resolution HSI is directly associated with significant technical and cost-based difficulties. Fusion methods are designed to overcome this barrier via mathematical processing of available low resolution HSI (LR–HSI) and high resolution MSI (HR–MSI) of the same scene (Figure 1).

2.2. Data description and preprocessing

This study used synthetic data generated on the basis of real hyperspectral images from the CAVE set (<http://www.cs.columbia.edu/CAVE/databases/multispectral/>). Model learning was performed on real scenes: Washington DC Mall, Pavia Center and Botswana, and the Real HSI/MSI/PAN dataset (<https://www.kaggle.com/datasets/shuangliangli123/real-hsimsipan-image-dataset>), while testing was performed on CAVE-based synthetic data. In the given configuration, dimensions of low-resolved hyperspectral cubes $64 \times 64 \times 76$ (spatial dimensions $h \times w$ and the number of spectral channels L_h), and dimensions of multispectral high resolution images are $192 \times 192 \times 8$ (spatial dimensions and the number of channels L_m), which corresponds to a space scaling coefficient of 3 and is used for controlled tests with known high resolution reference image.

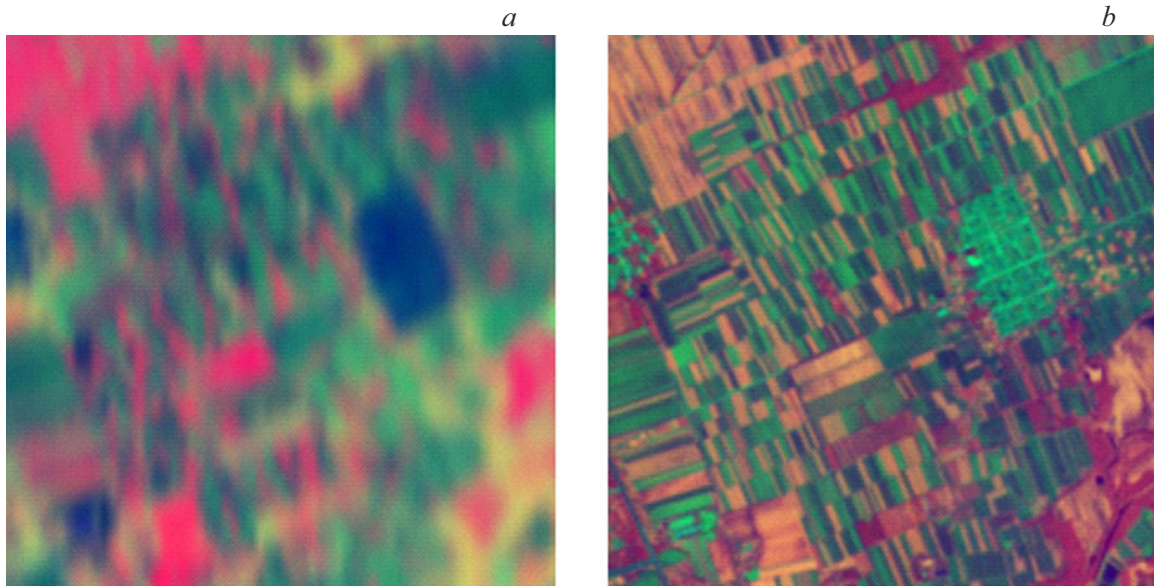


Figure 1. Example of low spatial resolution HSI (a) and high spatial resolution MSI (b) of the same scene (from <https://www.kaggle.com/datasets/shuangliangli123/real-hsimsipan-image-dataset>). For visualization of multichannel images, PCA is used: the first three main components are used that contain the principal portion of the effective signal, are normalized and displayed into corresponding channels of the RGB color model with projection to the first three main components and reflection into the appropriate color space.

Generation of synthetic test data from the CAVE dataset included the following degradation stages to mimic realistic observation conditions.

1. Spatial blurring: the space-wise Gaussian filter with $\sigma_{\text{blur}} = 0.8$ was applied to reference HQHSI to simulate optical system constraints and atmospheric effects.

2. Noise addition: additive Gaussian noise with zero mean and standard deviation $\sigma_{\text{noise}} = 0.8$ was added to blurred images to mimic sensor's instrumental noise. Final values were cut to fit the $[0,1]$ range.

3. Downsampling: downsampling with $r = 3$ was performed using the area interpolation method to obtain $64 \times 64 \times 76$ LR-HSI.

4. Spectral degradation: HR-MSI were formed using linear spectral transformation through the matrix $R \in \mathbb{R}^{8 \times 76}$ simulating the spectral response of the multispectral sensor.

Data preprocessing included the following stages, implemented in the provided software.

1. Download and normalization: HSI and MSI input data was downloaded and normalized via the min-max-normalization to $[0,1]$ to ensure comparability of values and stability of subsequent calculations:

$$x_{\text{norm}} = \frac{(x - x_{\min})}{x_{\max} - x_{\min}}.$$

2. Image cropping: to ensure size compatibility and cutting into patches (small fixed-size image fragments used for neural network learning), original HSI and MSI were cut so that their spatial dimensions (height and width in pixels) were multiples of a patch size. For MSI, a spatial scaling coefficient relative to HSI (64×64) was considered.

3. Cutting into patches: sets of 64×64 learning patches for HSI and 192×192 learning patches for MSI without overlapping were formed from prepared images to facilitate the increase in the learning sample size and model stability.

4. Change of HSI size: to compare with MSI, a low resolution HSI was increased to MSI dimensions using bicubic interpolation to ensure correct comparison of spatial structures.

These stages ensure high quality preparation of input data for delivery to fusion models and provide higher reliability of results.

2.3. Quality assessment metrics

Quantitative quality assessment of recovered images uses a set of labels covering various aspects of similarity to the reference image.

PSNR is a photometric accuracy metric that evaluates the RMS error of pixel amplitudes [19]:

$$\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}_I^2}{1/(L_H N) \sum_b \sum_i (H_{b,i} - \hat{H}_{b,i})^2} \right),$$

where L_H is the number of spectral channels, N is the total number of pixels, MAX_I is the maximum possible pixel value, $H_{b,i}$ is the reference pixel value i in channel b , $\hat{H}_{b,i}$ is the recovered pixel value.

SSIM is the structural similarity metric that evaluates similarity of local patches in brightness, contrast and structure [20]:

$$\text{SSIM}(x, y) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)},$$

where x, y are comparable local fragments (patches); μ_x, μ_y are mean brightnesses of fragments x, y ; σ_x, σ_y are dispersions (contrast) of fragments x, y ; σ_{xy} is covariance of fragments x, y (structural similarity measure); C_1, C_2 are small constants for division stabilization.

SAM a spectral form metric that measures an angle between spectral pixel vectors (in degrees) [21]:

$$\text{SAM} = \frac{180}{\pi} \frac{1}{N} \sum_i \arccos \left(\frac{\langle h_i, \hat{h}_i \rangle}{\|h_i\|_2 \|\hat{h}_i\|_2 + \varepsilon} \right),$$

where N is the total number of pixels per image, h_i is the spectrum of reference pixel i , $\hat{h}_i$ is the spectrum of recovered pixel i , $\langle h_i, \hat{h}_i \rangle$ is the scalar product, $\|h_i\|_2$ is the Euclidian norm h_i , ε is the small constant for numerical stability (to prevent division by zero).

ERGAS is the metric of relative integral error by channels considering the signal scale [22]:

$$\text{ERGAS} = 100 \frac{1}{d} \sqrt{\frac{1}{L_h} \sum_b \left(\frac{\sqrt{\frac{1}{N} \sum_i (H_{b,i} - \hat{H}_{b,i})^2}}{\mu_b} \right)^2},$$

where d is the downsampling coefficient (scaling coefficient); μ_b mean reference image value in channel b .

Combined use of (PSNR, SSIM, SAM, ERGAS) covers photometric, structural and spectral quality of recovery.

2.4. Description of proposed approaches

Analysis of existing solutions and tests was used as the basis for development and study of two main approaches to HSI and MSI fusion.

Approach 1: NMF-based spectral transformation

NMF fundamentals. NMF is a method for decomposing the nonnegative matrix $X \in \mathbb{R}_+^{m \times n}$ into a product of two smaller nonnegative matrices: $X \approx WH$, where $W \in \mathbb{R}_+^{m \times r}$ and $H \in \mathbb{R}_+^{r \times n}$, while the factorization rank $r \ll \min(m, n)$. In terms of image fusion, NMF is interpreted as follows: the matrix W constitutes basis spectral profiles (end-members), and the matrix H constitutes mixing coefficients of these profiles for each pixel. NMF is formulated as an optimization problem [23,24]:

$$\min_{W, H \geq 0} \|X - WH\|_F^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Nonnegativity of factors provides physical interpretability corresponding to the additive nature of spectral measurements. For HSI–MSI fusion, classical NMF finds a global spectral transformation matrix W connecting the MSI and HSI spaces.

Proposed expansion. This approach is focused on finding nonlinear transformation to map MSI into the HSI space. The basic linear part implies that each HSI spectral channel can be approximated by the linear combination

of MSI channels: $H_R \approx WM$, where $H_R \in \mathbb{R}^{L_h \times N}$ is the hyperspectral image increased to the MSI dimensions via interpolation and expanded into the matrix (expansion of the tensor $\mathbb{R}^{h_{\text{MSI}} \times w_{\text{MSI}} \times L_h}$ into the matrix $\mathbb{R}^{L_h \times N}$, where $N = h_{\text{MSI}} \times w_{\text{MSI}}$ is the total number of pixels for this 76×36864 configuration), $M \in \mathbb{R}^{L_m \times N}$ is the multispectral image (expansion of the tensor $\mathbb{R}^{h_{\text{MSI}} \times w_{\text{MSI}} \times L_h}$ into the matrix $\mathbb{R}^{L_m \times N}$ for this 8×36864 configuration) and $W \in \mathbb{R}^{L_h \times L_m}$ is the sought-for spectral transformation matrix (in this implementation 76×8). However, to consider nonlinear effects in sensors' spectral response and atmospheric distortions, square-law correction is added to the linear transformation.

The approach is implemented using **WMatrixGenerator**, a module that generates pixel-by-pixel (tensor) spectral transformation matrices W' and square-law correction coefficients, which constitutes nonlinear generalization of global NMF to local adaptive transformation.

WMatrixGenerator architecture. This module implements generation of the spatial variant (pixel-by-pixel) matrix of spectral transformation W' and square-law polynomial coefficients for MSI mapping into the HSI space (Figure 2). Concatenation of the interpolated low resolution HSI H_R and normalized MSI serves as the input, M is the number of channels.

Module specification includes the following elements.

• **Input:** $H_R \in \mathbb{R}^{\text{Batch} \times L_h \times h_{\text{MSI}} \times w_{\text{MSI}}}$ and $M \in \mathbb{R}^{\text{Batch} \times L_m \times h_{\text{MSI}} \times w_{\text{MSI}}}$ (tensors reduced to a spatial resolution), where *Batch* is the number of batches.

• **Output:** spectral weight matrices $W' \in \mathbb{R}^{\text{Batch} \times L_m \times h_{\text{MSI}} \times w_{\text{MSI}}}$ and square-law polynomial coefficients $A, B, C \in \mathbb{R}^{\text{Batch} \times L_m \times h_{\text{MSI}} \times w_{\text{MSI}}}$.

• **Application:** linear transformation $y_{\text{lin}} = W'M$ followed by square-law correction $\hat{y} = Ay_{\text{lin}}^2 + By_{\text{lin}} + C$.

Architecture (designations: *Chan* — basic channels, d_{emb} — positional encoding dimension) implements the following sequence.

• **Positional encoding:** normalized coordinate grids are formed and transformed into positional representations $E \in \mathbb{R}^{\text{Batch} \times d_{\text{emb}} \times h_{\text{MSI}} \times w_{\text{MSI}}}$ through a small MLP, adding the information about pixel location to the model.

• **Input convolution:** $[H_R, M]$ concatenation undergoes Conv2d with the 3×3 kernel and ReLU, forming the basic feature tensor F_0 .

• **Downsampling and receptive field expansion:** positional representations are added to features, then downsampling and residual processing blocks are accomplished (stride2 for $\text{Chan} \rightarrow 2\text{Chan}$ and $2\text{Chan} \rightarrow 4\text{Chan}$ transitions) for context accumulation [25].

• **Upsampling:** two transported convolutional layers ($k = 4, s = 2, p = 1$) return the spatial resolution and reduce the number of channels ($4\text{Chan} \rightarrow 2\text{Chan} \rightarrow \text{Chan}$) without appearance of checkerboard artifacts.

• **Final processing:** residual block and 3×3 convolution form the smoothed representation F_{ref} , then one-dimensional convolution is used to find W' and (A, B, C) .

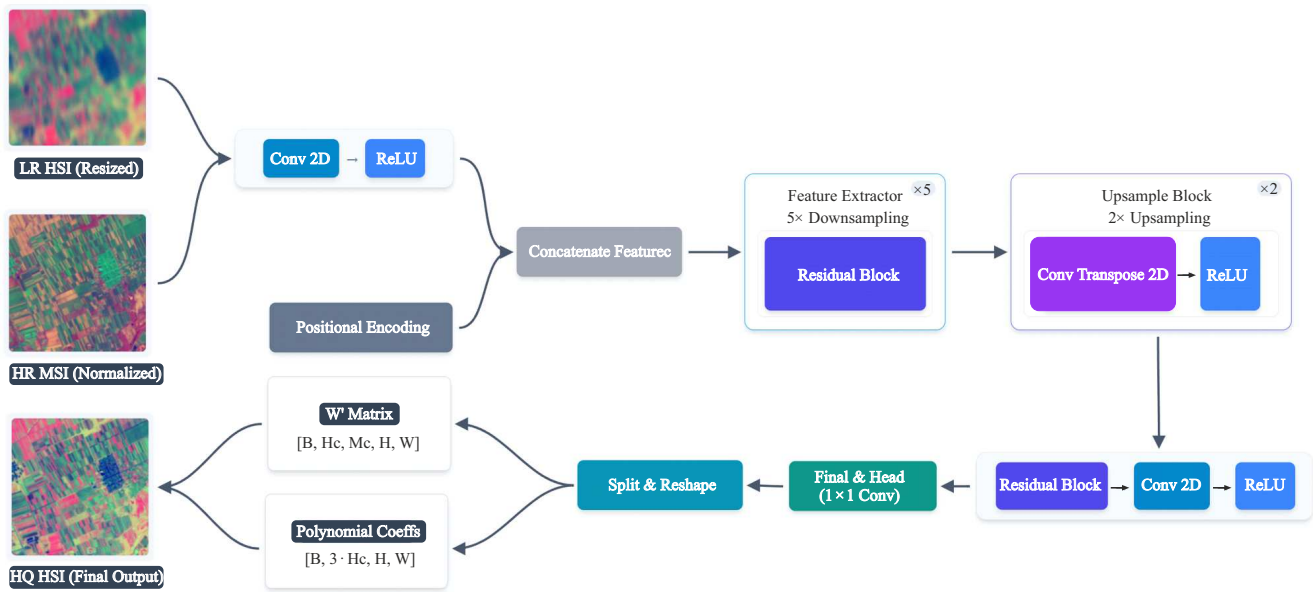


Figure 2. WMatrixGenerator flowchart visualization. Here, Conv2D is a filter that extracts spatial and spectral features from images through convolution, ResidualBlock is a block with bypass connection designed to train the deep network without loss of information about the original signal, Positional Encoding is a mechanism for adding pixel position to data for local adaptation of the spectral transformation, Concatenate Features is an operation of feature concatenation from HSI and MSI for simultaneous processing, Feature Extractor is a data compression module for context analysis and extraction of main scene features, Upsample Block (ConvTranspose2D) is a block for high spatial resolution recovery from compressed features, Final Head (1×1 Conv) is a layer that converts the accumulated features into particular matrix parameter values, Split & Reshape is a technical operation for data separation and change of data dimensions for correct multiplication, W' Matrix/Polynomial Coeffs are network output parameters (matrix and coefficients) for mathematical transformation of the MSI spectrum into HSI.

Nonlinear transformation: linear convolution of the MSI spectrum is performed for each pixel:

$$y_{\text{lin}}[\text{batch}, l_h, h, w] = \sum_j W'[\text{batch}, l_h, j, h, w] \times M[\text{batch}, l_h, h, w], \quad (1)$$

where batch is a batch number, l_h is the spectral channel, h and w are the height and width in pixels, respectively. Then, a limited square-law polynomial is applied to y_{lin} :

$$\hat{y}[\text{batch}, l_h, h, w] = A[\text{batch}, l_h, h, w] y_{\text{lin}}^2[\text{batch}, l_h, h, w] + B[\text{batch}, l_h, h, w] y_{\text{lin}}[\text{batch}, l_h, h, w] + C[\text{batch}, l_h, h, w], \quad (2)$$

where A, B undergo scaling by the hyperbolic tangent function for numerical stability:

$$A \leftarrow 0.1 \tanh(\cdot), \\ B \leftarrow 1 + 0.5 \tanh(\cdot).$$

Final high resolution image is formed with residual connection to improve the learning stability:

$$Y_{\text{fake}} = y_{\text{lin}} + \hat{y}. \quad (3)$$

Properties and connection with NMF: Pixel-by-pixel matrices W' provide local adaptivity, square-law correction

adds moderate nonlinearity, and downsampling expands the receptive field. Linear part y_{lin} generalizes NMF to local pixel transformations, where the global matrix W is replaced with adaptive $W'_{h,w}$. The module can learn both in forced learning mode (with reference targets W_{NMF}) and in end-to-end learning mode together with cycle loss.

PseudoInverseGenerator W^+ . Symmetric module builds pixel-by-pixel pseudoinverse matrices for transition from the HSI space back to MSI and learns together with W' using the cycle flowchart below. Pseudoinversibility term in the PseudoInverseGenerator module description refers to recovery of the linear spectral transformation component.

- **Input convolution:** $\text{Conv2d}(L_H \rightarrow C, 3 \times 3)$ with ReLU.

- **Extraction and upsampling:** residual block cascade with downsampling to $4C$ followed by ConvTranspose2d upsampling to the initial resolution.

- **Final block:** residual block and 3×3 convolution form smoothed representation.

- **Output:** $\text{Conv2d}(C \rightarrow L_M L_H, 1 \times 1)$ with scaling to the tensor $W^+ \in \mathbb{R}^{\text{Batch} \times L_M \times L_H \times h_{\text{MSI}} \times w_{\text{MSI}}}$.

Co-learning of W' and W^+ in the cycle architecture decrease error accumulation with the MSI–HSI–MSI series transition, provides mutual regularization and connects the

method with the classical NMF paradigm through local adaptive transformation matrices.

In the proposed model, full transformation is a composition of direct-matrix linear mapping and additive nonlinear polynomial correction. Inverse transformation is symmetrically built.

Pseudoinversibility is implemented at two levels. At the linear level, the orthogonal loss function ensures pseudoinversibility of the inverse transformation matrix relative to the direct transformation matrix. This in turn ensures correct recovery of the basic linear signal structure. At the nonlinear level, full pseudoinversibility of mapping and compensation of quadratic terms are achieved by means of inverse polynomial parameter learning under the action of the cycle consistency loss function.

Thus, the generated inverse matrix ensures linear basis pseudoinversibility, while nonlinear residues are inverted by the learning polynomial coefficients of the pseudoinverse generator within end-to-end network optimization.

Approach 2: HSI-MSI CycleGAN architecture (cycle model with pixel-by-pixel spectral kernels)

The second approach is implemented as the HSI-MSI CycleGAN cycle architecture (Figure 3) that expands the classical cycle consistency idea [26] through using (i) pixel-by-pixel direct and inverse transformation spectral matrices, (ii) parametrized quadratic nonlinearities, (iii) positional representations for increasing local adaptivity. Model consists of two reciprocal mappings: $G_M \rightarrow H$, $G_H \rightarrow M$.

Direct mapping. Interpolated HSI H_R and multispectral image M are input. WMatrixGenerator forms pixel-by-pixel matrices W' and coefficients A, B, C , then linear spectral mixing and square-law correction are performed using equations (1)–(3).

Inverse mapping. PseudoInverseGenerator module receives Y_{fake} and generates pseudoinverse matrices W^+ together with nonlinear correction coefficients, which makes it possible to recover MSI and close the cycle:

$$\begin{aligned} x_{\text{lin}}[\text{batch}, l_h, h, w] &= \sum_{i=1}^{L_H} W^+[\text{batch}, l_h, i, h, w] \\ &\quad \times Y_{\text{fake}}[\text{batch}, i, h, w], \\ \dot{x}[\text{batch}, l_h, h, w] &= A'[\text{batch}, l_h, h, w] \\ &\quad \times (x_{\text{lin}}[\text{batch}, l_h, h, w])^2 + B'[\text{batch}, l_h, h, w] \\ &\quad \times x_{\text{lin}}[\text{batch}, l_h, h, w] + C'[\text{batch}, l_h, h, w], \\ X_{\text{rec}} &= x_{\text{lin}} + \dot{x}. \end{aligned}$$

Both generators use positional encoding concatenated with features before blocks with downsampling and upsampling to increase the effective receptive field while retaining the local adaptivity.

Direct and inverse transformation orthogonality is ensured by additional regularization that minimizes the difference

between the product of W^+W' and a unit matrix for each pixel:

$$L_{\text{ortho}} = \|W^+W' - I\|_1.$$

Loss function includes the cycle consistency, spectral and spatial degradation components, gradient alignment, SAM loss, SSIM loss, reconstruction component and matrix norm regularization, which is reflected in Table 3.

Learning scheme. Learning is performed in three stages, each with 75 epochs. WMatrixGenerator is pre-initialized from the linear approximation and then all blocks learn together.

Stage 1 (epochs 1–75). Weights $\lambda_{\text{spectral}} = \lambda_{\text{spatial}} = \lambda_{\text{grad}} = \lambda_{\text{SAM}} = 1$, $\lambda_{\text{SSIM}} = 2$ are used without the reconstruction component ($\lambda_{\text{recon}} = 0$). Learning rate is equal to $1e-4$ and ensures linear mapping stabilization.

Stage 2 (epochs 76–150). Reconstruction loss with respect to HSI ($\lambda_{\text{recon}} = 5$) is added and spectral-angular penalty ($\lambda_{\text{SAM}} = 5$) is augmented. Other weights are maintained, learning step decreases to $1e-5$.

Stage 3 (epochs 151–225). Weight coefficients coincide with the second stage, however, the optimization step increases to $3e-5$ for fine adjustment and maintaining the SAM.

All stages use data splitting into learning, validation and test samples without intersections between scenes. Early stopping is controlled by validation SAM and SSIM, and small weights of spectral smoothness and distribution loss maintain the learning stability.

3. Results and discussion

3.1. Comparative performance analysis of the existing methods

For performance evaluation of the proposed methods, quantitative comparison with modern approaches to HSI–MSI fusion was performed. Table 4 shows PSNR, SSIM, SAM and ERGAS values for various methods, including approaches proposed in this work.

For numerical metrics, successive increase in PSNR and SSIM and decrease in ERGAS are observed during transition from linear models to hybrid and GAN-/attention architectures. Mean values by groups of methods can be generalized as follows.

- Linear methods (NMF, PCA): mean PSNR ≈ 18.2 dB, SSIM ≈ 0.40 , SAM $\approx 12.6^\circ$, ERGAS ≈ 3.34 — basic quality level with high interpretability.
- Regularization and sparse methods (HySure, NCSR): mean PSNR ≈ 29.1 dB, SSIM ≈ 0.87 , SAM $\approx 15.4^\circ$, ERGAS ≈ 2.77 — substantial improvement of structural consistency with growing computational complexity.
- Tensor models (CSTF, Deep Tucker, UDTN): mean PSNR ≈ 34.5 dB, SSIM ≈ 0.94 , SAM $\approx 7.8^\circ$, ERGAS ≈ 1.30 — significant growth through multidimensional composition.

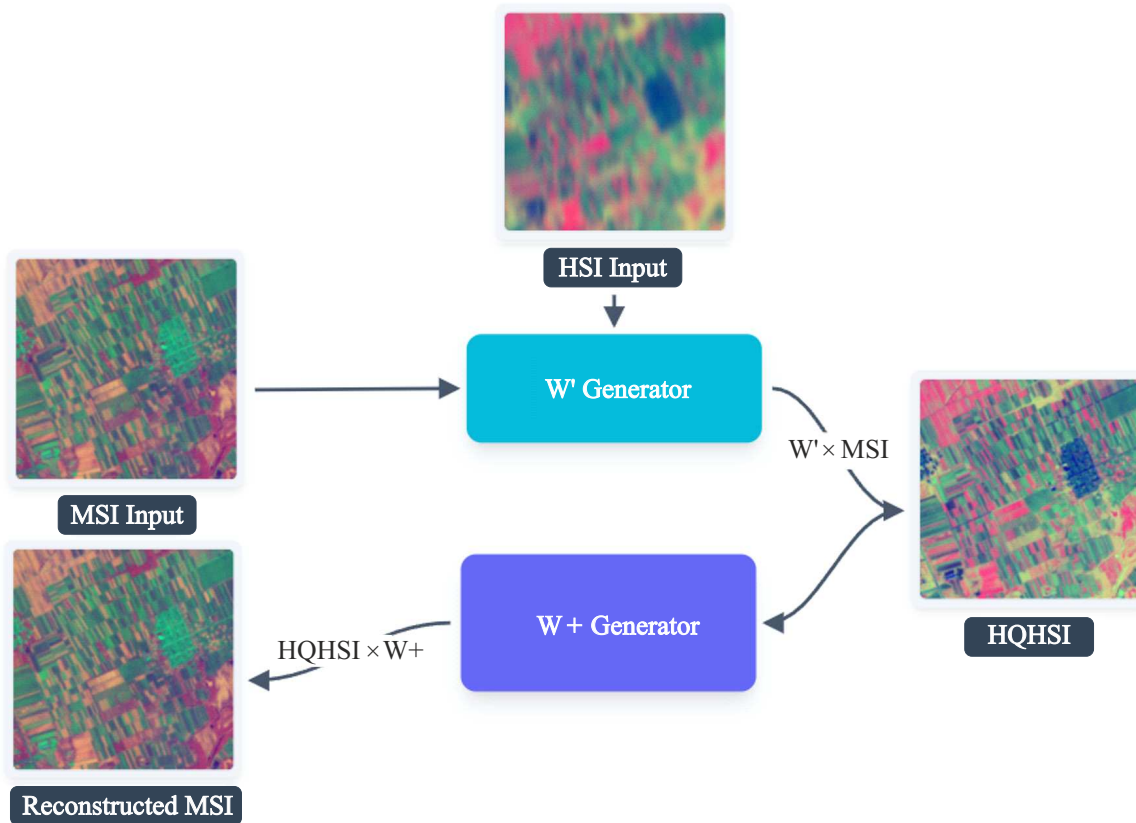


Figure 3. HSIMSICycleGAN flowchart visualization. Output MSI and HIS are supplied to the W' Generator block for generation of the W' matrix, then W' and MSI are mixed in HQHSI. At the next step, the W^+ Generator block generates the W^+ matrix and the latter is mixed with HQHSI for generation of the final reconstructed MSI.

Table 3. Loss function components and weights at three learning stages

Component	Short formula	Purpose	Stage 1	Stage 2	Stage 3
Cycle	$\ M - X_{\text{rec}}\ _1$	MSI→HSI→MSI consistency	1.0	1.0	1.0
Spectral (HSI→MSI)	$\ Conv_{1 \times 1}(Y_{\text{fake}}) - M\ _1$	SRF approximation	1.0	1.0	1.0
Spatial (HR→LR)	$\ down_{area}(blur(Y_{\text{fake}})) - H_R\ _1$	Degradation alignment	1.0	1.0	1.0
Gradient	$\sum_d \ \partial_d(Y_{\text{fake}}) - \partial_d(M)\ _1$	Boundary maintenance	1.0	1.0	1.0
SAM	$SAM(Y_{\text{fake}}, up_{\text{bicubic}}(H_R))$	Spectral form	1.0	5.0	5.0
SSIM	$1 - SSIM(Conv_{1 \times 1}(Y_{\text{fake}}), M)$	Local structural affinity	2.0	2.0	2.0
Recon (HR→HSI)	$\ Y_{\text{fake}} - up(H_R)\ _1$	Photometric reconstruction	0.0	5.0	5.0
Spectral smoothness	$\ Y_{\text{fake}}[:, 1, :] - Y_{\text{fake}}[:, :, -1]\ _1$	Spectral smoothing	0.0	0.0	0.0
Distribution	$W_1(hist(Y_{\text{fake}}), hist(up(H_R)))$	Channel histogram alignment	1.0	1.0	
Othogonality	$\ W^+ W' - I\ _1$	Local matrix pseudoinversibility	0.1	0.1	0.1
Regularization	$\ W'\ _2^2 + \ W^+\ _2^2$	Weight norm control	0.01	0.01	0.01

• Modern deep learning approaches (MIAE, HSR–KAN): mean PSNR ≈ 40.3 dB, SSIM ≈ 0.98 , SAM $\approx 5.5^\circ$, ERGAS ≈ 1.20 — maximum indicators with high resource intensity and data requirements.

Critically small SAM values are achieved for models with explicit structural regularization (RAF–NLRGS: 1.063°) and hybrid architectures with tensor components (Deep Tucker: 4.23° , HSR–KAN: 3.17°). Approaches

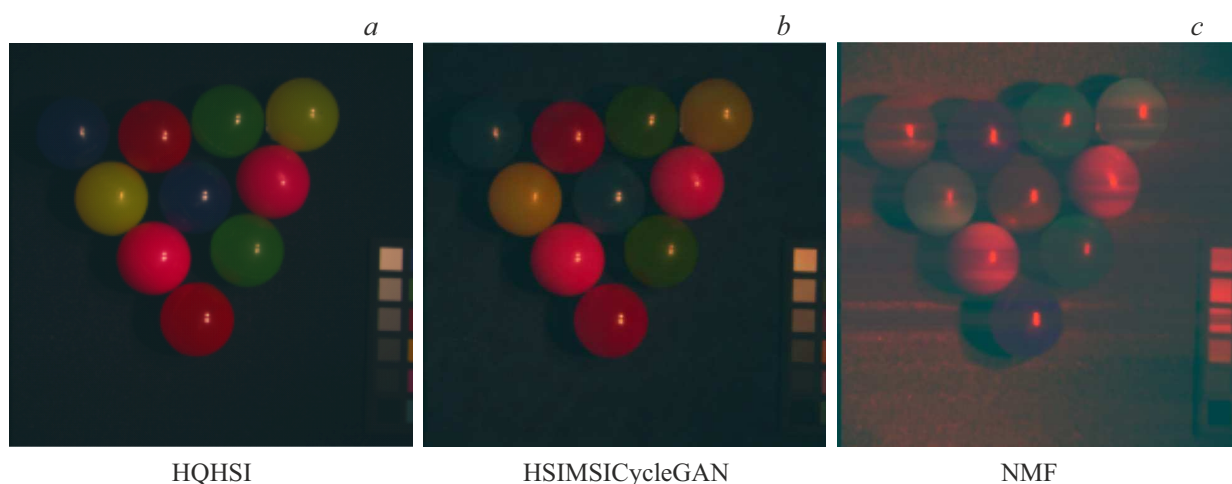


Figure 4. Example of a scene consisting of a synthetic CAVE-based dataset: (a) HQHSI (high quality hyperspectral image), (b) recovered high resolution HSI built by HSIMSiCycleGAN, (c) NMF-based method result. Visualization of hyperspectral images (a,b,c) uses the principal component analysis (PCA) with projection on the first three principal components followed by mapping into RGB.

Table 4. Comparative performance metrics for various HSI–MSI upsampling methods ($\uparrow$ means that a higher value is better, $\downarrow$ means that a lower value is better)

Method	PSNR $\uparrow$	SSIM $\uparrow$	SAM $\downarrow$	ERGAS $\downarrow$
NMF	17.17	0.3865	13.32	3.56
PCA	19.23	0.4217	11.87	3.12
CSTF	28.55	0.904	12.26	1.94
HySure	27.63	0.841	16.63	4.01
NSSR	30.59	0.899	14.17	1.53
Deep Tucker	35.18	0.938	4.23	1.40
MIAE	35.58	0.927	10.38	0.98
uSDN	37.74	0.974	9.28	0.70
HyCoNet	38.72	0.966	6.90	1.94
Physics-Based GAN	38.97	0.990	5.53	1.20
UDTN	39.8	0.982	6.78	0.55
MossFuse	42.05	0.990	6.57	0.42
BliEstGAN	43.9	0.990	6.89	–
RAF-NLRGS	43.57	0.991	1.063	2.075
HSR–KAN	44.89	0.991	3.17	2.04
Proposed methods				
NMF	11.72	0.2185	31.76	–
HSIMSiCycleGAN	25.30	0.8133	13.51	–

proposed in the work supplement the existing solution range. HSIMSiCycleGAN demonstrates intermediate re-

sults (PSNR = 25.30 dB, SSIM = 0.8133, SAM = 13.51°), surpassing basic methods and being inferior to advanced hybrid architectures. Basic WMatrixGenerator is still limited (PSNR = 11.72 dB, SSIM = 0.2185, SAM = 31.76°), indicating that further regularization and architecture improvement are required. Modern deep learning methods (MossFuse, HSR–KAN, RAF–NLRGS) provide the best quantitative indicators and confirm the performance of locally adaptive nonlinear transformations. Proposed HSIMSiCycleGAN provides a fair trade-off between the quality and data availability, relying on paired HSI/MSI observations without a reference HQHSI (Figure 4, 5).

There is generally a steady quality improvement trend during transition from global linear models to locally adaptive nonlinear architectures that combine physical knowledge and learning components.

3.2. Discussion of results

Quantitative evaluation of the proposed methods has shown differences in their performance. HSIMSiCycleGAN achieves moderate reconstruction quality (PSNR = 25.30 dB, SSIM = 0.7546) with relatively adequate retention of spectral response curve (SAM = 14°), which holds the method promising for tasks where simultaneous HSI/MSI observations are available, but a reference HQHSI is absent. WMatrixGenerator in basic configuration demonstrates limited performance (PSNR = 11.72 dB, SSIM = 0.2185, SAM = 31.76°), indicating that integration of more complex regularization mechanisms is required.

Employed architecture combines locally adaptive matrices, moderate nonlinearity and cycle consistency, providing the interpretability of W' and W^+ transformations and improving the balance between photometric, structural and spectral metrics.

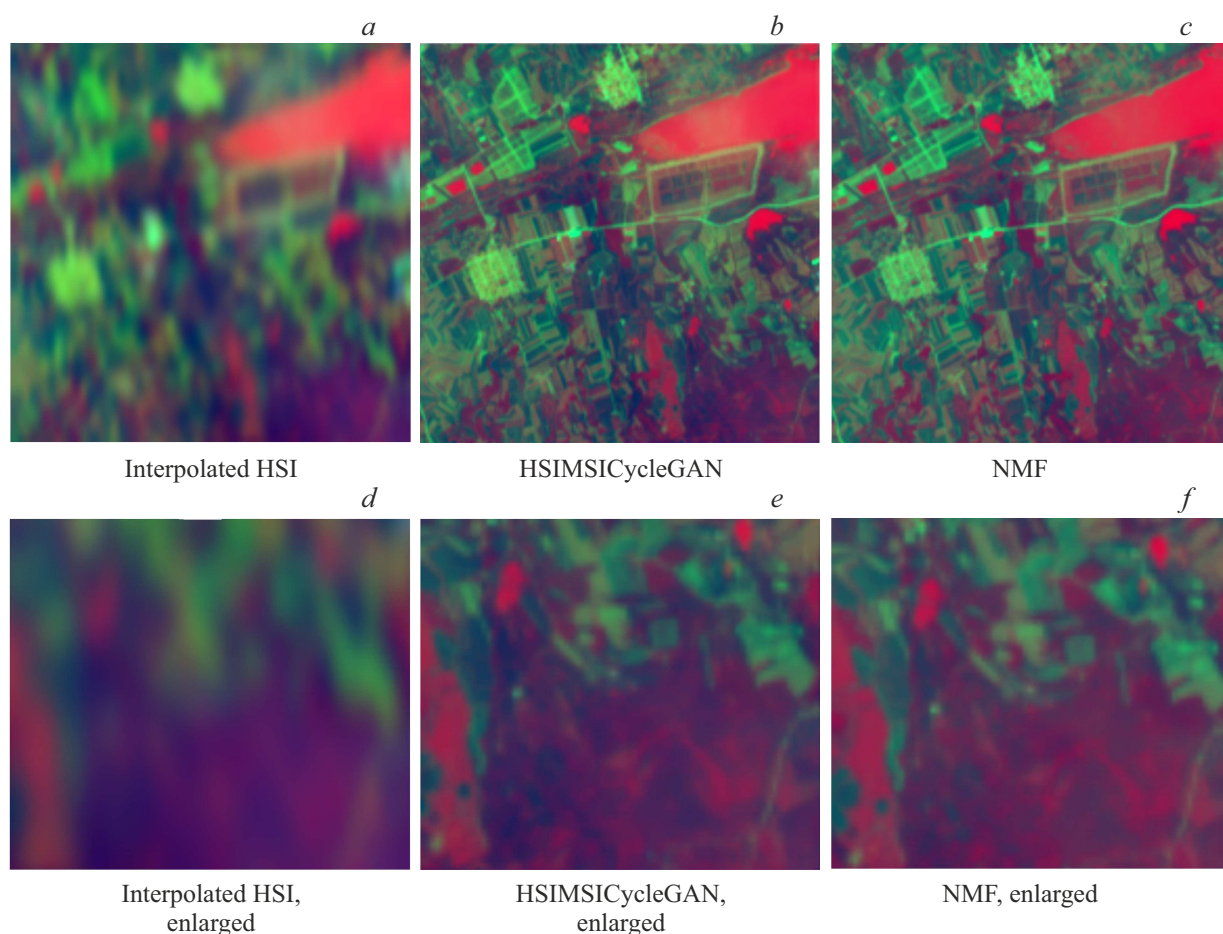


Figure 5. Visual comparison of fusion results (*a,b,c*) and corresponding enlarged fragments (*d,f,g*). Left to right: interpolated HSI (192×192 , from Real HSI/MSI/PAN), HSIMSiCycleGAN result (192×192) and NMF-based method result (192×192).

In image analysis tasks, interpretability serves as a key factor used to verify physical soundness of recovered data. It ensures that the result follows from mathematically justified measurement transformation, but is not an artifact of nontransparent computational procedures [27].

In the NMF-based approach, this interpretability is provided by strict compliance of the linear mixing model with nonnegativity constraints ($W, H \geq 0$), where components carry direct physical sense of spectral signatures and their concentrations.

Within the proposed HSIMSiCycleGAN architecture, strict nonnegativity constraint for generated W' matrices becomes weaker. This allows simulation of complex dependences and spectral correction effects beyond the simplest additive model. Benefit of this approach is in implementing a spectral subtraction mechanism: negative weights can be used to suppress a spurious signal from overlapping broad MSI channels with recovery of narrow HSI bands. Such mechanism provides higher photometric accuracy of spectral reconstruction which is potentially unattainable with strict nonnegativity constraints.

W' matrix in this case is interpreted as a dynamic spectral reconstruction operator adapting to the local scene

context, rather than as a set of static basis spectral profiles. Maintaining the transformation matrix structure makes it possible to directly trace the contribution of each MSI channel to formation of the final spectrum, providing transparency of the analysis model as opposed to fully hidden deep neural network layers.

Strengths of the study

- Modern HSI–MSI fusion methods have been reviewed.
- Two new approaches have been proposed, implemented and preliminary studied: NFM-based pixel-by-pixel spectral transformation and HSIMSiCycleGAN architecture.
- Practical data preparation and learning procedures have been described to facilitate the repeatability of results.

Shortcomings and limitations

- The obtained estimates are preliminary in nature and require extended comparison with many advanced methods on various datasets.
- Neural network model learning, in particular HSIMSiCycleGAN, is accompanied with a high computational cost.

- Thorough selection of hyperparameters and architectural components is required for stable convergence.
- It should be noted that patch-based learning can result in artifacts during final fusion, however, this issue was not addressed in this work.

Development prospects

It is planned to perform large-scale experimental validation on various datasets with various spatial and spectral characteristics, and to study alternative attention mechanisms and regularization strategies. Integration of physically justified imaging models, development of light architectures for embedded systems and employment of transfer learning for quick adaptation to new sensor platforms are of interest.

Scope of application of this technique includes both enhancing the resolution of hyperspectral and multispectral images obtained from satellites and UAV and wide application in biophotonics [28]. Thus, in [29], neural network methods for super-resolution reconstruction are used to build predictive models of diagnosis of plant diseases using hyperspectral images. Moreover, hyperspectral visualization of human skin is a potential neoplasm diagnosis tool [30–32]. Enhancement of resolution of such images using RGB images [33] will also substantially improve the HSI resolution.

4. Conclusion

The work reviews modern methods for fusing hyperspectral and multispectral images and two new approaches to solving this problem. Experimental evaluation has shown that the HSIMSiCycleGAN architecture demonstrated moderate results (PSNR = 25.30 dB, SSIM = 0.7546, SAM = 14°), surpassing classical methods, but being inferior to modern advanced solutions. A key benefit is the capability of learning on paired observations of HSI/MSI without using high spatial resolution reference data, which is important for remote sensing applications.

Results confirm the trend of deep learning method predominance over classical approaches with importance of balance between accuracy and computational efficiency. The study provides a basis for further research in the field of uncontrolled fusion of multimodal satellite data.

Method selection shall be based on specific features of the task and available resources. For quick applications with moderate quality requirements, classical methods (NMF, PCA) are recommended. In the case when there are paired observations of low resolution HSI and high resolution MSI, but a reference HQHSI is not available, it is reasonable to use HSIMSiCycleGAN type cycle architectures. To achieve the maximum quality with vast labeled data, modern deep methods with physical constraints and hybrid architectures are preferable.

Funding

The study was funded by the Ministry of Science and Higher Education of the Russian Federation (grant № 075-15-2024-557 as of 25.04.2024).

Conflict of interest

The authors declare no conflict of interest.

References

- [1] *Hyperspectral Data Exploitation: Theory and Applications*, ed. by C.I. Chang (John Wiley & Sons, 2007). DOI: 10.1002/9780470124628.ch3
- [2] T. Adão, J. Hruška, L. Pádua, J. Bessa, E. Peres, R. Morais, J.J. Sousa. *Remote Sensing*, **9** (11), 1110 (2017). DOI: 10.3390/rs9111110
- [3] G. Avola, A. Matese, E. Riggi. *Remote Sensing*, **15** (7), 1917 (2023). DOI: 10.3390/rs15071917
- [4] S. Du, Y. Zou, Z. Wang, X. Li, Y. Li, C. Shang, Q. Shen. *arXiv:2412.04802*, (2024). DOI: 10.48550/arXiv.2412.04802
- [5] J. Yang, L. Xiao, Y.-Q. Zhao, J.C.-W. Chan. *IEEE Transactions on Neural Networks and Learning Systems*, **35** (9), 13017–13031 (2024). DOI: 10.1109/TNNLS.2023.3266038
- [6] J. Liu, Z. Wu, L. Xiao, X.-J. Wu. *IEEE Transactions on Geoscience and Remote Sensing*, **60**, 1–12 (2021). DOI: 10.1109/TGRS.2022.3143156
- [7] K. Zheng, L. Gao, W. Liao, D. Hong, B. Zhang, X. Cui. *IEEE Transactions on Neural Networks and Learning Systems*, **59** (3), 2487–2502 (2021). DOI: 10.1109/TGRS.2020.3006534
- [8] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, Z. Xu. *European Conference on Computer Vision* (Cham: Springer International Publishing, 2020), p. 208–224. DOI: 10.1007/978-3-030-58526-6_13
- [9] Y. Qu, H. Qi, C. Kwan. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2511–2520 (2018). DOI: 10.48550/arXiv.1904.12175
- [10] S. Li, R. Dian, L. Fang, J.M. Bioucas-Dias. *IEEE Transactions on Image Processing*, 2018. DOI: 10.1109/TIP.2018.2836307
- [11] M. Simões, J. Bioucas-Dias, L.B. Almeida, J. Chanussot. *IEEE Transactions on Geoscience and Remote Sensing*, 2015. DOI: 10.1109/TGRS.2014.2375320
- [12] W. Dong, F. Fu, G. Shi, X. Cao, J. Wu, G. Li, X. Li. *IEEE Transactions on Image Processing*, 2016. DOI: 10.1109/TIP.2016.2542360
- [13] J. Yang, L. Xiao, Y. Zhao, J.C.-W. Chan. *IEEE Transactions on Neural Networks and Learning Systems*. **35** (9), 13017–13031 (2023). DOI: 10.1109/TNNLS.2023.3266038
- [14] B. Li, X. Wang, H. Xu. *arXiv:2409.06705*, 2024. DOI: 10.48550/arXiv.2409.06705
- [15] H. Wang, Y. Xu, Z. Wu, Z. Wei. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. DOI: 10.1109/TNNLS.2024.3457781
- [16] K. Yang, M. Bai, T. Lu. *arXiv preprint arXiv:2407.05279*, 2024. DOI: 10.48550/arXiv.2407.05279
- [17] L. Wald. *Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions* (Presses des MINES, 2002).

- [18] R.A. Schowengerdt. *Remote Sensing: Models and Methods for Image Processing*, 3rd ed. (Academic Press, London, 2006).
- [19] A. Hore, D. Ziou. 20th International Conference on Pattern Recognition. IEEE, 2366–2369 (2010). DOI: 10.1109/ICPR.2010.579
- [20] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli. IEEE Transactions on Image Processing, **13** (4), 600–612 (2004). DOI: 10.1109/TIP.2003.819861
- [21] R.H. Yuhas, A.F.H. Goetz, J.W. Boardman. JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop, vol. 1: AVIRIS Workshop (1992).
- [22] K. Kolikov, G. Krastev, Y. Epitropov, D. Hristozov. Chemo-metrics and Intelligent Laboratory Systems, **102** (1), 15–19 (2010). DOI: 10.1016/j.chemolab.2010.03.001
- [23] G.H. Golub, C.F. Van Loan. *Matrix Computations*, 4th ed. (Baltimore, MD, 2013). DOI: 10.56021/9781421407944
- [24] G. Strang. *Introduction to Linear Algebra* (Wellesley-Cambridge Press, 2022). DOI: 10.1137/1.9781733146678
- [25] K. He, X. Zhang, S. Ren, J. Sun. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), p. 770–778 (2016). DOI: 10.1109/CVPR.2016.90
- [26] J.-Y. Zhu, T. Park, P. Isola, A. Efros. Proceedings of the IEEE International Conference on Computer Vision, 2223–2232 (2017). DOI: 10.1109/ICCV.2017.244
- [27] Z. Salahuddin, H.C. Woodruff, A. Chatterjee, P. Lambin. Computers in Biology and Medicine, **140**, 105111 (2022). DOI: 10.1016/j.compbiomed.2021.105111
- [28] M. Yako. Optical Review, 1–13 (2025). DOI: 10.1007/s10043-025-01001-x
- [29] M.M. Antony, C.S.S. Sandeep, H.-T. Lim, M.V. Matham. J. Biomedical Photonics & Engineering, **9** (3), 030304 (2023). DOI: 10.18287/JBPE23.09.030304
- [30] I.A. Bratchenko, V.P. Sherendak, O.O. Myakinin, D.N. Artemyev, A.A. Moryatov, E. Borisova, L. Avramov, L.A. Zherdeva, A.E. Orlov, S.V. Kozlov, V.P. Zakharov. J. Biomedical Photonics & Engineering, **4** (1), 5–12 (2018). DOI: 10.18287/JBPE17.04.010301
- [31] B.V. Grechkin, V.O. Vinokurov, Y.A. Khristoforova, I.A. Matveeva. J. Biomedical Photonics & Engineering, **9** (4), 50–55 (2023). DOI: 10.18287/JBPE23.09.040304
- [32] I.A. Bratchenko, O.O. Myakinin, V.P. Sherendak, P.N. Volkhin, Y.A. Khristoforova, L.A. Bratchenko, D.N. Artemyev, A.A. Moryatov, O.V. Polschikova, A.S. Machikhin, V.E. Pozhar, S.V. Kozlov, V.P. Zakharov. J. Biomedical Photonics & Engineering, **4** (4), 040301 (2018). DOI: 10.18287/JBPE18.04.040301
- [33] Z. Lai, Y. Fu, J. Zhang. IEEE Transactions on Neural Networks and Learning Systems, 2024. DOI: 10.1109/TNNLS.2023.3340561

Translated by E.Ilyinskaya