Journal of Engineering Technology and Applied Physics

Enhancing Spectral Quality through Multisource Fusion of SAR and Optical Imagery Using Deep Learning

Cuong Ha Tuan
Southern Sub-Institute of Forest Inventory and Planning, Ho Chi Minh city, Vietnam.
Corresponding author: tuancuongdialyk38@gmail.com, *ORCiD*: 0009-0006-8886-7849

https://doi.org/10.33093/jetap.2026.8.1.4
Manuscript Received: 15 June 2025, Revised: 13 July 2025, Accepted: 9 Augst 2025, Published: 15 March 2026

Abstract **—This study evaluates the effectiveness of a multi-source satellite image fusion method using deep learning to enhance spectral feature quality in urban environments. The input data consist of synthetic aperture radar (SAR) images from Sentinel 1 and optical images from Landsat 9. Two deep learning models were implemented: patch-wise shallow convolutional neural networks (CNN) and convolutional autoencoders (CAE). Each of Red-Green-Blue optical band was fused separately with the VV/VH ratio image derived from radar data. The fusion results were assessed using RMSE, SSIM, UIQI indices, and the Pearson correlation coefficient. The CAE demonstrated better spectral reconstruction capability with lower RMSE and higher SSIM across all three bands. Conversely, the CNN model achieved higher UIQI on some bands and produced images with visually superior sharpness. Visual assessment indicated that CNN better preserves edge details and fine structures, while CAE generated smoother images but with some blurring of objects. These results suggest that deep learning-based fusion methods hold great potential for improving input image quality for urban analysis in areas affected by cloud cover.**

Keywords—Multisource image fusion, Sentinel 1, Landsat 9, Deep learning.

I. INTRODUCTION

The fusion of multi-source data, particularly SAR and optical imagery, is an emerging trend in remote sensing. These two data types differ in spectral properties, acquisition mechanisms, and applications. Optical sensors capture rich spectral information useful for surface classification but are limited by weather conditions. In contrast, SAR operates independently of weather and lighting, yet lacks spectral detail when used alone, limiting its effectiveness in land cover analysis.

To overcome and compensate for the limitations of optical and radar sensors [1, 2], missing information in one type of imagery can be supplemented by the other, thereby improving image quality and enhancing classification accuracy of surface objects [3, 4]. Currently, fusion methods for optical and SAR images can be broadly categorized into two main groups: transform domain methods and spatial domain methods [5].

- Transform domain methods are based on traditional multiscale transform theory. Common techniques in this group include wavelet transform [6], contourlet transform [7], and non-subsampled contourlet transform [8]. These methods often produce good fusion results in terms of preserving image details and texture, but they require careful design of fusion rules and involve high computational costs.

- Spatial domain methods process images directly based on pixel values or image regions. Common techniques include Intensity-Hue-Saturation (IHS), Principal Component Analysis (PCA), and Brovey transform [9]. Although these methods are simple and easy to implement, they strongly depend on manually designed fusion rules, leading to limited generalizability when applied to different data types.

Deep learning algorithms have opened a new research direction in recent years [10]. Deep neural networks possess the capability to model complex nonlinear relationships between remote sensing observations and geospatial parameters. This is a strength that traditional algorithms, especially physically interpretable models, cannot achieve [11].

Several commonly used architectures include convolutional neural networks (CNN), which are effective in extracting deeper spatial features and have

https://doi.org/10.33093/jetap.2026.8.1.4

Published by MMU PRESS. URL: https://journals.mmupress.com/index.php/jetap/index

shown high performance in noise reduction, resolution enhancement, object detection, and image classification tasks [12] and recurrent neural networks (RNN), which are well suited for handling time series data such as forecasting and classification tasks [13].

Recent advances in deep learning have enabled more effective approaches to multisource image fusion, particularly for combining SAR and optical data. Unlike traditional techniques such as PCA, IHS, or Wavelet, deep models like CNNs, GANs, and autoencoders offer better preservation of spatial and spectral details. For example, Zhang *et al.* [14] developed a dual-branch CNN that maintained structural fidelity, while Xiong *et al.* [15] used a conditional GAN to generate cloud-free optical images from SAR, achieving high SSIM (~0.995). Other architectures, including U-Net [16], attention-based autoencoders [17], and transformer variants [18, 19], have shown strong performance, underscoring the effectiveness of deep learning in SAR-optical fusion.

This study assessed the fusion of Sentinel 1 and Landsat 9 imagery acquired in 2014 over Ho Chi Minh City using deep learning methods. The integration of radar data shows strong potential for surface monitoring in urban areas characterized by complex structures. Additionally, Sentinel 1 helps mitigate cloud-related limitations in optical images, improving the overall quality of Landsat 9 data.

II. STUDY AREA

Ho Chi Minh city is the largest economic, cultural, and scientific center in Vietnam, located at approximately 10°45N and 106°40E. The city covers an area of around 2,095 square kilometers and consists of 22 districts, including Thu Duc city, with a population of over 9 million as of 2023 [20]. Figure 1 below shows the geographical location of Ho Chi Minh city in vietnam.

The selection of Ho Chi Minh city as the study area aims to evaluate the relationship between urban development and environmental factors, particularly through the integration of optical and radar remote sensing data for monitoring spatial changes. Satellite data such as Sentinel 1 (SAR) and optical imagery provide a rich and reliable source of information for analyzing land use change, urban heat effects, and environmental degradation.

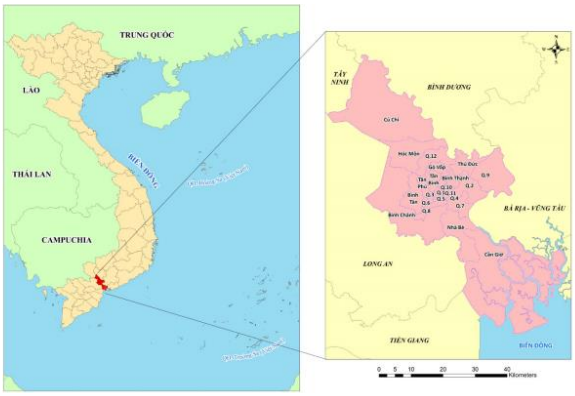


Fig. 1. Location of Ho Chi Minh city.

III. DATA AND METHODS

A. Satellite Data and Pre-processing

This study utilized two primary data sources: optical imagery from Landsat 9 and synthetic aperture radar (SAR) data from Sentinel 1, aiming to combine the strengths of both sensor types in a deep learning-based image fusion model. All input images were standardized to the WGS 84 UTM Zone 48N coordinate system, clipped to the study area boundary, and normalized to a pixel value range of 0 to 1 to enhance stability during the model training process. Detailed information about the two image sources used in this study is presented in Table I.

Table I. Detailed information on satellite images used.

Type	Date	Spectral bands used	Source
Level-2 Landsat 9	22/03/2024	Blue (Band 2) Green (Band 3) Red (Band4)	United States Geological Survey
Sentinel-1 C-band	31/03/2024	VV, VH polarization at 10m	Copernicus

The data used in this study include Level 2 Landsat 9 images acquired under low cloud conditions in 2024. Three spectral bands Red (Band 4), Green (Band 3), and Blue (Band 2) with a spatial resolution of 30 meters were selected to represent the visual and natural color information of the urban surface [21].

Sentinel 1 employs C band radar operating in the Interferometric Wide swath (IW) mode, which provides a ground resolution of 10 meters. The Sentinel 1 imagery was processed using the SNAP software (developed by ESA), following a standardized workflow. This includes the application of precise orbit files (apply orbit file), thermal noise removal, radiometric calibration, and speckle filtering using a 5×5 Lee filter. Finally, terrain correction was performed to ensure both geometric and geographic accuracy of the SAR images before fusion.

In this study, the VV/VH polarization ratio was calculated from radar images collected in 2024 to extract structural and textural information from the surface. One major advantage of radar data is its insensitivity to cloud cover or illumination conditions [22]. The VV/VH ratio has demonstrated improved capability in distinguishing urban features and vegetation when integrated with optical data [23].

B. Data Normalization

Input data normalization is an essential step in deep learning tasks to ensure that features share the same units and value distributions. This helps the model converge more efficiently and prevents issues such as “gradient vanishing” or “exploding gradients” [24, 25]. In this study, all optical image bands and the VV/VH ratio radar image were normalized to a standard range of 0 to 1 using the following formula:

$$\boldsymbol{x_{norm}} = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

, where x_{norm} is the normalized pixel value; x is the original pixel value. x_{max} and x_{min} represent the maximum and minimum pixel values in the corresponding image band.

For optical data such as Landsat 9 surface reflectance (SR) imagery, pixel values typically range from 0 to 10,000, since the product is provided at Level 2. However, for radar data VV/VH ratio, values are usually expressed in decibels (dB). Therefore, before normalization, they must be converted into linear scale using the following formula [26]:

$$\mathbf{dB} = 10 \text{ x } log_{10}\ (\text{DN}^2) \qquad (2)$$

, where 10 is a standard coefficient used in the conversion formula because the energy-based logarithmic function applies a base-10 logarithm.

DN refers to the pixel value of the uncalibrated radar image (amplitude).

To ensure spatial consistency between images from different satellite sources, geometric correction was applied after standardizing resolution and pixel values. Due to differences in orbit and sensor configuration, Landsat 9 and Sentinel 1 imagery may not align perfectly.

In this study, all data were reprojected to the WGS 84 UTM zone 48N system. Sentinel 1 images processed in SNAP and exported as GeoTIFF were used as spatial references. Landsat images were then aligned using image overlay techniques to reduce spatial mismatches. Accurate geometric correction is crucial for deep learning models to learn spatial relationships between radar and optical inputs at the pixel level, especially in image fusion and reconstruction tasks.

C. *Convolutional Autoencoder (CAE) Method*

The Convolutional Autoencoder (CAE) is a deep neural network architecture designed to learn a compressed representation of input images through two main phases: encoding and decoding [27]. The encoder utilizes convolutional layers (Conv2D) combined with non-linear LeakyReLU activation functions and max-pooling (MaxPooling2D) layers to extract critical spatial and spectral features. Conversely, the decoder reconstructs the image using upsampling layers (UpSampling2D) followed by convolution, with the bottleneck layer serving as a compressed latent representation of the input data [28].

In this study, a lightweight CAE model consisting of two encoding layers, a single bottleneck layer, and two decoding layers was implemented in Python 3.10 using Visual Studio Code. The development environment included TensorFlow 2.12, Keras, NumPy, Rasterio, and Matplotlib libraries.

The model was trained with input data of size 128×128 pixels, composed of two channels: (i) a multispectral optical patch (Red, Green, Blue), and (ii) a radar-derived VV/VH ratio patch. The output is an optical image reconstructed from the fused input pair. Mean Squared Error (MSE) was employed as the loss function, and the Adam optimizer (learning rate = 0.001) was used for training. The model contained approximately 11,817 trainable parameters, making it computationally suitable for medium-resolution satellite imagery such as Landsat.

Prior to training, all input images were normalized to the range (0, 1) and uniformly cropped into non-overlapping patches of 128×128 pixels. This patch-based approach was adopted to optimize memory usage and avoid out-of-memory (OOM) errors, which are common when training on high-resolution remote sensing data [29]. Figure 2 illustrates the CAE architecture, where each input is a dual-channel image combining optical and radar information. The network was trained for 15 epochs per spectral band, allowing it to learn effective optical reconstruction while leveraging structural information from radar data.

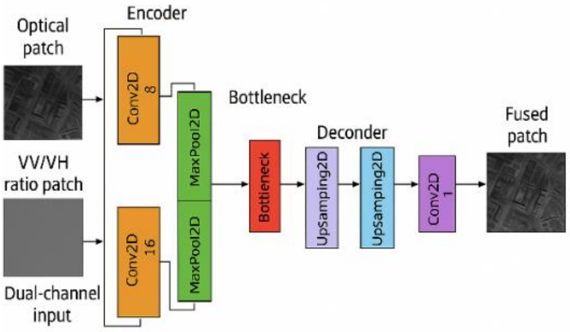
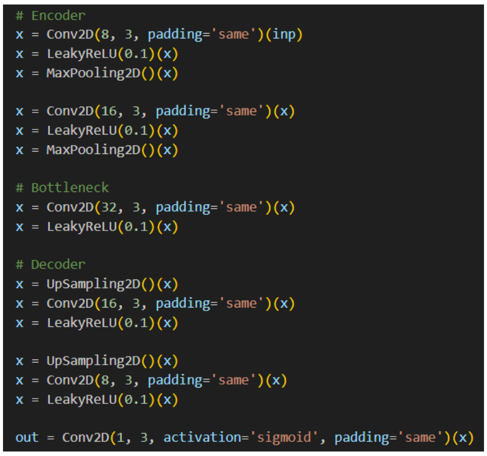


Fig. 2. Architecture of the convolutional autoencoder network for integrating SAR and optical imagery in this study.

The CAE model employed in this study follows a symmetric encoder–decoder architecture with a bottleneck layer in between (Figure 3). The model processes 128×128×2 patches, where the two channels correspond to co-registered optical and radar data. This lightweight architecture (≈11,800 trainable parameters) balances efficiency and performance, suitable for medium-resolution satellite data.



```
# Encoder
x = Conv2D(8, 3, padding='same')(inp)
x = LeakyReLU(0.1)(x)
x = MaxPooling2D()(x)

x = Conv2D(16, 3, padding='same')(x)
x = LeakyReLU(0.1)(x)
x = MaxPooling2D()(x)

# Bottleneck
x = Conv2D(32, 3, padding='same')(x)
x = LeakyReLU(0.1)(x)

# Decoder
x = UpSampling2D()(x)
x = Conv2D(16, 3, padding='same')(x)
x = LeakyReLU(0.1)(x)

x = UpSampling2D()(x)
x = Conv2D(8, 3, padding='same')(x)
x = LeakyReLU(0.1)(x)

out = Conv2D(1, 3, activation='sigmoid', padding='same')(x)
```

Fig. 3. Code used to implement CAE architecture in this study.

- *Encoder:* Two convolutional blocks extract spatial – spectral features. Each block consists of a 3×3 convolution (8 and 16 filters, respectively), LeakyReLU activation (slope = 0.1), and MaxPooling2D to downsample the input to 64×64 and then 32×32.

- *Bottleneck:* A 3×3 convolution with 32 filters and LeakyReLU captures abstract, high-level features while preserving spatial size.

- *Decoder:* The decoder upsamples the feature maps using UpSampling2D, followed by 3×3 convolutions (16 and 8 filters) and LeakyReLU activations. A final 3×3 convolution with sigmoid activation reconstructs the output image in the range (0, 1).

In this study, the convolutional autoencoder (CAE) model was trained for *15 epochs* per spectral band. This fixed epoch setting was chosen based on prior experimentation to balance model convergence and computational efficiency. Given the stable performance observed during training and the moderate model complexity (≈11,800 parameters), no dynamic early-stopping criterion was applied.

This approach helped the model learn how to reconstruct optical images from radar-optical input without causing overfitting.

D. Shallow Convolutional Neural Network (Shallow CNN) Method

In addition to the CAE model, this study also implemented a shallow convolutional neural network (CNN) to compare its performance in reconstructing optical images from radar data. This CNN has a simplified architecture, consisting of only a few consecutive convolutional layers and does not include the encoding or decoding components typically found in CAE. The detailed network structure is illustrated in Fig. 4.

The shallow CNN model in this study was designed using a patch-wise processing approach, combining optical data with the VV/VH polarization ratio from Sentinel-1 radar imagery to reconstruct individual spectral channels. By incorporating structural information from radar data, the model enhances detail and reduces noise, particularly under challenging optical conditions such as cloud cover or atmospheric disturbances [30].

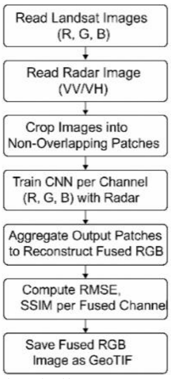

Read Landsat Images (R, G, B)
Read Radar Image (VV/VH)
Crop Images into Non-Overlapping Patches
Train CNN per Channel (R, G, B) with Radar
Aggregate Output Patches to Reconstruct Fused RGB
Compute RMSE, SSIM per Fused Channel
Save Fused RGB Image as GeoTIF

Fig. 4. Architecture of the shallow CNN model used in this study.

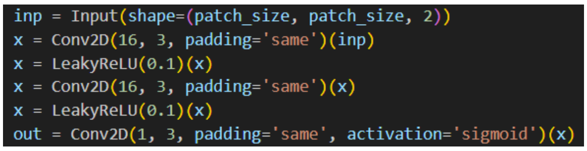
The network architecture consists of three consecutive convolutional layers. The first two layers use 16 filters of size 3×3 with LeakyReLU activation, while the final layer applies a single filter with a sigmoid function to normalize the output to the (0, 1) range. Input patches are sized 64×64×2, corresponding to two channels: one from optical imagery and one from radar. The model structure is as follows:

- *First Conv2D layer:* 16 filters (3×3), 2 input channels → 304 trainable parameters

- *Second Conv2D layer:* 16 filters (3×3), 16 input channels → 2320 parameters

- *Final Conv2D layer:* 1 filter (3×3) → 145 parameters

Figure 5 illustrates a shallow CNN segment consisting of three consecutive convolutional layers.



```
inp = Input(shape=(patch_size, patch_size, 2))
x = Conv2D(16, 3, padding='same')(inp)
x = LeakyReLU(0.1)(x)
x = Conv2D(16, 3, padding='same')(x)
x = LeakyReLU(0.1)(x)
out = Conv2D(1, 3, padding='same', activation='sigmoid')(x)
```

Fig. 5. Code used to implement shallow CNN architecture.

For the shallow CNN model, a patch-wise training approach was used, and each spectral channel was trained independently for 5 epochs. The choice of a lightweight network and low epoch count was guided by computational constraints and the need to avoid memory overflow during training on high-resolution satellite data. While no best epoch was explicitly selected, the 5-epoch training allowed the model to converge adequately, as indicated by evaluation metrics such as RMSE and SSIM. Future implementations could benefit from incorporating validation-based early stopping to fine-tune the number of training iterations.

Due to memory constraints and the high resolution of input images, the model was trained in a patch-wise manner. Each 64×64 patch was treated as an independent training sample. The loss function used was Mean Squared Error (MSE), computed between the original and reconstructed images. After training, the model predicted each patch individually, and the full image was reconstructed by averaging overlapping pixels [28]. Patch-wise training not only reduces memory usage but also improves the ability to learn local features effectively [29].

E. Image Quality and Structural Evaluation Indices

In this study, three widely used quantitative metrics were employed to evaluate the quality of the fused images: Root Mean Square Error (RMSE), Structural Similarity Index (SSIM), and Universal Image Quality Index (UIQI). Each metric captures a different aspect of similarity or discrepancy between the fused image and the original reference image.

- *Root Mean Square Error (RMSE)*

Root Mean Square Error (RMSE) is a commonly used metric for calculating the average deviation between corresponding pixel values of the fused image and the reference image. It is sensitive to large errors and reflects the overall level of deviation [31, 32]. The formula is expressed as follows:

$$\textbf{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - Q_i)^2} \qquad (3)$$

, where P_i denotes the pixel value in the fused image, Q_i denotes the corresponding pixel value in the reference image and N is the total number of pixels.

RMSE ranges from 0 to ∞ and reflects the absolute error between the fused image and the reference image. A lower RMSE value indicates smaller error and higher similarity to the original image.

- *Structural Similarity Index Measure (SSIM)*

Structural Similarity Index (SSIM) measures the similarity between two images in terms of structure, luminance, and contrast [31]. The formula is defined:

$$\textbf{SSIM (x, y)} = \frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \qquad (4)$$

, where

μ_x, μ_y denote the mean values of x and y;

σ_x^2, σ_y^2, are the variances of x, y;

σ_{xy} is the covariance between the two images;

C_1, C_2 are small constants to stabilize the division.

This index ranges from –1 to 1, where a value of 1 indicates perfect similarity. In contrast, values close to 0 or negative suggest a significant difference between the fused image and the reference image.

- *Universal Image Quality Index (UIQI)*

The Universal Image Quality Index (UIQI) is a widely used metric proposed by Wang and Bovik [33], designed to quantify the overall similarity between two images. UIQI models image quality distortion based on three main components: correlation, standard deviation, and mean luminance. The calculation formula is as follows:

$$\textbf{UIQI} = \frac{4\sigma_{xy}\mu_x\mu_y}{(\sigma_x^2+\sigma_y^2)(\mu_x^2+\mu_y^2)} \qquad (5)$$

, where

μ_x, μ_y are the mean values of images x and y;

σ_x^2, σ_y^2, are their variances;

σ_{xy} is the covariance between the two images.

The UIQI value ranges from –1 to 1. A value of 1 indicates that the fused image is structurally, contrast-wise, and luminance-wise identical to the reference image, meaning the image quality is fully preserved. Conversely, values close to 0 or negative indicate significant differences between the two images, reflecting low fusion quality.

F. Spatial Quality Metrics

To objectively evaluate the spatial fidelity of the fused images, four commonly used spatial quality metrics were employed: *Shannon Entropy (EN), Sobel Edge Strength, Spatial Frequency (SF), and Gradient Magnitude (GM).* These metrics quantify various aspects of spatial detail, edge sharpness, and pixel intensity variation.

- *Shannon Entropy (EN)*

Shannon entropy (EN) measures the information content or complexity of an image. It is defined as [34]:

$$\textbf{EN} = -\sum_{i=0}^{L-1} p(i) \times log_{2p(i)} \qquad (6)$$

, *where*

p(i) is the probability of gray level iii in the image;

L is the number of gray levels.

Higher value indicates a more complex image with greater variability in pixel intensities, which is typically associated with higher spatial richness.

- *Sobel edge strength*

This metric measures the average strength of edges detected using the Sobel operator, capturing structural sharpness [35]:

$$\textbf{E} = \frac{1}{MN} - \sum_{x=1}^{M}\sum_{y=1}^{N}\sqrt{G_x^2(x,y) + G_y^2(x,y)} \qquad (7)$$

, where

G_x, G_y are horizontal and vertical Sobel gradients;

M, N are the number of image rows and columns.

Higher Sobel values reflect sharper edges and better preservation of fine spatial structures.

- *Spatial Frequency (SF)*

Spatial frequency reflects how rapidly image intensity changes across space, indicating textural richness [36]. A higher SF implies more textural detail and better spatial resolution in the image:

$$\textbf{SF} = \sqrt{RF^2 + CF^2} \qquad (8)$$

with *RF (Row Frequency):* pixel intensity variation across rows:

$$\textbf{RF} = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=2}^{M}[I\ (i,j) - I\ (i,j-1)]^2} \qquad (9)$$

CF (Column Frequency): variation across columns:

$$\textbf{CF} = \sqrt{\frac{1}{MN}\sum_{i=2}^{M}\sum_{j=1}^{M}[I\ (i,j) - I\ (i,j-1)]^2} \qquad (10)$$

- *Gradient magnitude (GM)*

Gradient Magnitude measures the average brightness transition strength between neighboring pixels [37]:

$$\textbf{GM} = \frac{1}{MN} - \sum_{x=1}^{M}\sum_{y=1}^{N}\sqrt{\left(\frac{\partial I}{\partial x}\right)^2\left(\left(\frac{\partial I}{\partial y}\right)^2\right)^2} \qquad (11)$$

IV. RESULTS AND DISCUSSION

A. Visual Evaluation of Image Quality

When comparing the results of the two models at 10-meter resolution, the differences between the fused RGB bands are quite clear.

Red band: The shallow CNN gives a much sharper image, with finer details in urban features like roads and buildings. Edges appear cleaner, and high-reflectance areas are more distinct, which helps in identifying different land uses. On the other hand, the CAE output looks more blurred. It smooths out noise well, but this also causes a loss of detail, especially along boundaries.

Green band: In this band, the shallow CNN result shows visible grid lines that break the image continuity. These artifacts likely come from the model not fully learning the connection between neighboring patches. In contrast, CAE produces a more consistent image without obvious blocky patterns. However, it also makes the image look softer, and some small features can be hard to recognize.

Blue band: The CNN again performs better in keeping surface patterns clear. Features like rooftops and industrial zones stand out more. That makes it useful for identifying man-made structures. Meanwhile, the CAE image looks more uniform and avoids breaking up low-detail areas, but at the cost of some sharpness.

Figure 6 below shows the results of combining the two image sources, SAR and optical, using the shallow CNN and the CAE networks.

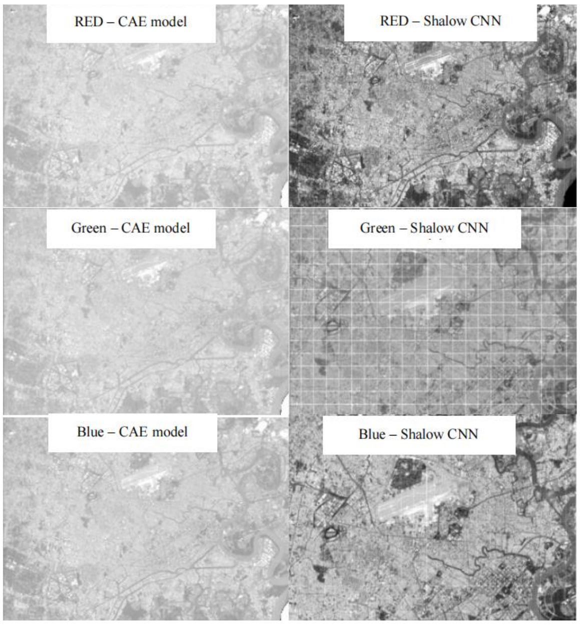


Fig. 6. Visual differences between the three RGB bands reconstructed by the CAE and shallow CNN models at 10m in the central area of Ho Chi Minh city.

When examining the output images, it becomes clear that the shallow CNN offers better visual clarity. The images appear sharper, with urban features such as roads and buildings standing out more clearly. This is especially useful for identifying man-made structures or monitoring detailed surface changes. However, this model also introduces some blocky artifacts, likely due to the patch-based training not fully preserving continuity across neighboring regions.

On the other hand, the CAE model tends to smooth out noise effectively, producing more visually stable results. But this smoothing also reduces texture and sharp edges, particularly around areas with high contrast or complex structures. As a result, while CAE preserves spectral consistency, it may miss certain spatial details important for object recognition or boundary detection.

Improving either model could involve adding skip connections, applying attention layers, or experimenting with hybrid architectures to better balance spatial and spectral fidelity.

B. Quantitative Comparison of CNN Shallow and CAE Models in Image Fusion

The performance of the two fusion models was evaluated using four commonly used image quality indices: Root Mean Square Error (RMSE), Structural Similarity Index (SSIM), Universal Image Quality Index (UIQI), and the Pearson correlation coefficient (r). These metrics assess both spectral fidelity and spatial consistency between the fused and original images. The results are summarized in Table II.

Table II. Comparison of image quality evaluation indices.

Band	Model	RMSE	SSIM	UIQI	Pearson's r
Red	Shallow CNN	0.0442	0.9753	0.0407	0.99
	CAE	0.0067	0.9911	0.047	0.94
Green	Shallow CNN	0.0496	0.9677	0.0611	0.83
	CAE	0.0077	0.9828	0.0585	0.94
Blue	Shallow CNN	0.0379	0.9818	0.0674	0.99
	CAE	0.0073	0.9894	0.0485	0.94

In all three bands, CAE model consistently achieved lower RMSE values, indicating better spectral reconstruction with less absolute error. SSIM scores were high for both models, but CAE slightly outperformed CNN, suggesting superior spatial consistency. Although CNN produced higher UIQI values in the green and blue bands, typically linked to high visual contrast, the CAE achieved the best performance in the red band, which commonly emphasizes sharp man made structures. Pearson's r also revealed interesting differences: CNN performed very well in red and blue bands (r ≈ 0.99) but dropped sharply in green (r ≈ 0.83), whereas CAE maintained more balanced correlation values across all bands (r ≈ 0.94).

C. Spatial Quality Metrics Across Original and Reconstructed Images (CNN Shallow vs CAE Model)

To assess the spatial fidelity of the reconstructed images, four commonly used spatial quality metrics, including: *Entropy, Sobel Edge, Spatial Frequency, and Gradient Magnitude* were computed for the red, green, and blue bands. The evaluation was conducted using the normalized original Landsat 9 imagery and the corresponding fused outputs generated by the Shallow CNN and CAE models. Table III presents the calculated Shannon Entropy (EN) values, representing the information content of the fused images obtained by the two methods.

Table III. Comparison of image structural complexity (entropy) across original image, shallow CNN, and CAE.

Band	Original	Shallow CNN	CAE
Red	4.75	8.02	2.46
Green	4.65	8.37	2.49
Blue	4.59	8.32	2.60

The shallow CNN model produced significantly higher entropy values across all 3 bands (Red ≈ 8.02, Green ≈ 8.37, Blue ≈ 8.32), reflecting greater spectral complexity and information content. In contrast, the CAE model achieved entropy values of only around 2.46 – 2.60, indicating that its reconstructed images contained less gray-level variation, resulting in lower spatial information.

Table IV. Comparison of edge intensity (sobel edge) among original, shallow CNN, and CAE images.

Band	Original	Shallow CNN	CAE
Red	0.0143	0.0099	2.46
Green	0.0106	0.013	2.91
Blue	0.0109	0.013	3.25

Table IV presents the calculated Sobel edge metric (E) results, reflecting the edge sharpness. The CAE model significantly outperformed in the Sobel edge metric, with values ranging from 2.46 to 3.25, compared to the much lower scores of shallow CNN (≈ 0.0098 – 0.0130). This suggests that CAE model was more effective at preserving edges and structural details, resulting in higher image sharpness.

Table V. Spatial frequency analysis across original and fused RGB bands from CNN shallow and CAE.

Band	Original	Shallow CNN	CAE
Red	0.0111	0.0104	1.46
Green	0.0099	0.0089	1.69
Blue	0.0104	0.0109	1.66

Table V presents the Spatial frequency (SF) index results, reflecting the textural richness of the images. The CAE model also exhibited significantly higher spatial frequency values (≈ 1.46 – 1.68), indicating greater pixel intensity variation in the fused images, which contributes to a sharper visual impression compared to CNN-generated images (with the SF value ≈ 0.0089 –0.0108).

Table VI. Comparison of overall sharpness using gradient magnitude across original images, CNN, and CAE.

Band	Original	Shallow CNN	CAE
Red	0.0018	0.0013	0.315
Green	0.0014	0.0016	0.37
Blue	0.0014	0.0017	0.416

Table VI presents the Gradient Magnitude (GM) index results, reflecting the brightness transition intensity between neighboring pixels. The gradient magnitude values of the CAE model ranged from 0.315 to 0.416, hundreds of times higher than those of CNN (≈ 0.0012 – 0.0016), reinforcing the observation that CAE model produces images with sharper edges and stronger transitions between bright and dark regions.

The results indicate that the CAE network produces images with sharp edges and clear spatial details (high Sobel, SF, and GM values), whereas the shallow CNN maintains higher spectral complexity (high Entropy), preserving spectral information better but with lower sharpness.

D. Comparison of The Performance between Other Methods in Image Fusion

- *Traditional transformation methods:*

Many studies have employed traditional pixel-level image fusion algorithms such as Principal Component Analysis (PCA), Intensity-Hue-Saturation (IHS) transformation, wavelet transform (WT), and non-subsampled contourlet transform (NSCT) to integrate radar (SAR) and optical image. These approaches are typically evaluated using quantitative indices such as RMSE, SSIM, UIQI, and the Pearson correlation coefficient between the original images and fused images.

For example, Ding *et al.* [18] compared four widely used techniques: PCA, IHS, WT, and a hybrid HIS+NSCT model on SAR and Landsat imagery ((the results are shown in Table VII)). The results indicated that PCA produced the highest error (RMSE ≈ 12.75) and the lowest spectral fidelity (UIQI ≈ 0.451), whereas the HIS+NSCT model achieved the lowest RMSE (≈ 8.24) and the highest UIQI (~0.856), suggesting a superior ability to preserve spectral characteristics. The remaining methods (IHS and Wavelet) showed intermediate performance metrics. For example, the IHS method resulted in RMSE ≈ 11.77, SSIM ≈ 0.843, and UIQI ≈ 0.715, while the WT produced RMSE ≈ 10.36, SSIM ≈ 0.795, and UIQI ≈ 0.662.

Table VII. Quantitative results (RMSE, SSIM, UIQI) of typical SAR – Optical image fusion methods (IHS, PCA, Wavelet, HIS–NSCT) based on experimental data (Ding et al. 2025).

Method	RMSE	SSIM	UIQI
IHS	11.77	0.843	0.715
PCA	12.75	0.862	0.451
Wavelet	10.36	0.795	0.662
HIS + NSCT	8.235	0.658	0.856

In addition, the study by MatecConf (2018) demonstrated that the combination of WT and IHS reported UIQI values ranging from 0.80 to 0.93 across bands 1, 2, and 3 of Landsat 5, which were significantly higher than those obtained using the Gram-Schmidt method (approximately 0.44 to 0.49).

Overall, these traditional approaches tend to produce greater spectral distortion (as indicated by lower UIQI values) compared to more advanced multi-resolution integration techniques such as NSCT.

- *Application deep learning method in intergrating SAR and optical image:*

Recently, deep learning models have been proposed for fusing radar and optical imagery. For instance, Luo *et al.* [38] developed DAFCNN, a dual-branch CNN integrated with attention mechanisms to fuse SAR and multispectral images. On a dataset of 656 samples, DAFCNN achieved a Pearson correlation coefficient (CC) of approximately 0.980 and an SSIM of about 0.939, significantly outperforming traditional methods. In the same experiment, the IHS method yielded only CC ≈ 0.291 and SSIM ≈ 0.036; NSCT achieved CC ≈ 0.261 and SSIM ≈ 0.185; while a multi-scale CNN (MSDCNN) reached CC ≈ 0.891 and SSIM ≈ 0.847. The summary of the results from Luo's research group is presented in Table VIII.

Table VIII. Quantitative results (RMSE, SSIM, UIQI) of typical SAR – Optical image fusion methods (IHS, PCA, Wavelet, HIS–NSCT) based on experimental data [39].

Method	CC (Pearson)	SSIM
IHS	0.2909	0.0362
NSCT	0.2607	0.1845
Wavelet	0.7017	0.5699
MSDCNN	***0.8913***	***0.8466***
DAFCNN	***0.9801***	***0.9394***

Recent studies have shown that generative adversarial networks (GANs) can produce highly effective results in both SAR to optical image translation and image fusion. For example, Xiong *et al.* [15] proposed a conditional GAN to reconstruct optical images from SAR data in cloudy conditions. The model achieved strong performance with SSIM around 0.995, PSNR near 54 dB, RMSE approximately 0.0067 (on normalized data), and a Pearson correlation coefficient close to 0.96. These results suggest that GAN-based models can accurately recover spatial and spectral details even when optical information is missing or degraded.

- *Benchmark comparison:*

Recent studies have demonstrated that deep learning such as CNNs, GANs, and ResNet-based models consistently outperform traditional fusion techniques in terms of quantitative metrics. Such as, the DAFCNN model achieved a correlation coefficient (CC) of approximately 0.98 and SSIM of 0.94, whereas IHS method resulted in CC around 0.29 and SSIM as low as 0.036. Similarly, a conditional GAN reported SSIM up to 0.995 and CC near 0.96, significantly higher than those observed in classical methods like PCA or IHS, which showed SSIM values between 0.7 - 0.8, and relatively low correlation.

These benchmark results provide a valuable reference for evaluating the performance of the proposed models in this study. Specifically, the CAE model presented here achieved SSIM values between 0.9828 and 0.9911 across R-G-B bands, with Pearson correlation coefficients around 0.94. In contrast, the shallow CNN model yielded higher visual sharpness and correlation in selected bands, with CC reaching 0.99 in the red and blue channels. These findings confirm that the deep learning-based fusion models developed in this study are comparable to or even surpass existing benchmarks in both spectral fidelity and spatial detail preservation.

E. Limitation and Future Directions

One key limitation of this study is the absence of land cover classification or image interpretation, which prevents the use of standard evaluation metrics such as precision, recall, or F1 score. The current assessment is based primarily on spatial quality indicators, which, while useful for measuring sharpness and structural preservation, do not fully reflect the semantic fidelity or practical utility of the reconstructed images.

Moreover, without ground truth labels, it is not possible to evaluate the reconstructed outputs in terms of land cover accuracy or their reliability for tasks such as land use mapping, forest monitoring, or change detection. This limits the generalizability and real-world applicability of the proposed models.

Future research should consider incorporating land cover classification or expert-based interpretation to provide a more comprehensive evaluation. This would enable the use of classification-based metrics such as overall accuracy, F1 score, or intersection over union, thereby clarifying the operational potential of deep learning models for environmental monitoring and resource management.

V. CONCLUSION

This study looked at how well two neural network models could combine radar and optical satellite images for Ho Chi Minh city. The first model used was a convolutional autoencoder. It was trained on small image patches that included both optical data from Landsat 9 and the VV/VH ratio from Sentinel-1. This model gave good results in terms of spectral accuracy. Specifically, it produced lower RMSE and higher SSIM scores in all three color bands, which suggests that it did a good job capturing spectral patterns. That said, the images it created often looked overly smooth and lacked sharp details. This might be an issue for applications that need more precise spatial features.

The second model was a shallow convolutional neural network. Its outputs were visually clearer, especially in places with dense or complex urban layouts. It also scored higher on the Pearson correlation in the red and blue bands. That indicates it was more successful at keeping the original spatial patterns intact.

Overall, the autoencoder seems more useful when keeping the spectral information consistent is important. In contrast, the shallow CNN might work better for tasks that need sharper, more detailed visuals. Both models show how deep learning can help combine radar and optical data, especially in areas where cloud cover often limits the use of regular satellite images.

In future research, more advanced deep learning architectures (e.g., ResNet, GAN, Transformer) may be incorporated to further improve spectral detail and spatial consistency. In addition, the proposed method could be applied to different geographic regions to evaluate its robustness and generalizability.

ACKNOWLEDGEMENT

The author would like to thank the editorial board of Journal of Engineering Technology and Applied Physics for their support and guidance in preparing this manuscript.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Ha Tuan Cuong: Conceptualization, Methodology, Investigation, Formal Analysis, Validation, Visualization, Writing – Original Draft, Review & Editing

CONFLICT OF INTERESTS

No conflict of interests was disclosed.

ETHICS STATEMENTS

Our publication ethics follow The Committee of Publication Ethics (COPE) guideline. https://publicationethics.org/

REFERENCES

[1] M. Schmitt, L. H. Hughes and X. X. Zhu, "The SEN1-2 Dataset for Deep Learning in SAR-Optical Data Fusion," *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. IV-1, pp. 141–146, 2018.

[2] Y. Ban, O. Yousif and H. Hu, "Fusion of SAR and Optical Data for Urban Land Cover Mapping and Change Detection," *Global Urban Monitoring and Assessment through Earth Observation,* CRC Press, pp. 353-386, 2014.

[3] C. Pohl and J. L. Van Genderen, "Multisensor Image Fusion in Remote Sensing: Concepts, Methods and Applications," *Int. J. Remote Sens.*, vol. 19, no. 5, pp. 823–854, 1998.

[4] Y. Jiao and R. S. Blum, "A Fusion Method of SAR and Optical Images for Urban Object Extraction," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XXXVII, part B7, Beijing, pp. 1257-1260, 2008.

[5] C. Hua, Z. Tang, K. Zhang, W. Jiang and H. Wang, "Study on Infrared Camouflage of Landing Craft and Camouflage Effect Evaluation," *Infrared Technol.*, vol. 30, pp. 379–383, 2008.

[6] M. N. Do and M. Vetterli, "The Contourlet Transform: An Efficient Directional Multiresolution Image Representation," *IEEE Trans. Image Process.*, vol. 14, pp. 2091–2106, 2005.

[7] Q. Xiao-Bo, Y. Jing-Wen, X. Hong-Zhi and Z. Zi-Qian, "Image Fusion Algorithm Based on Spatial Frequency-Motivated Pulse Coupled Neural Networks in Nonsubsampled Contourlet Transform Domain," *Acta Autom. Sin.*, vol. 34, pp. 1508–1514, 2008

[8] A. L. Da Cunha, J. P. Zhou and M. N. Do, "The Nonsubsampled Contourlet Transform: Theory, Design, and Applications," *IEEE Trans. Image Process.*, vol. 15, pp. 3089–3101, 2006.

[9] D. Limpitlaw and R. Gens, "Dambo Mapping for Environmental Monitoring Using Landsat TM and SAR Imagery: Case Study In The Zambian Copperbelt," *Int. J. Remote Sens.*, vol. 27, pp. 4839–4845, 2006.

[10] L. Zhang, L. Zhang and B. Du, "Deep Learning for Remote Sensing Data," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, 2016.

[11] M. Schmitt and X. Zhu, "Data Fusion and Remote Sensing - An Ever-Growing Relationship," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 4, pp. 6–23, 2016.

[12] H. C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura and R. M. Summers, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[13] S. Hochreiter and J. Schmidhuber, "Long Short-term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] Y. Zhang, F. Liu and H. Wang, "Dual-branch Convolutional Networks for SAR and Multispectral Data Fusion," *Remote Sens. Lett.,* vol. 11, no. 4, pp. 345–357, 2020.

[15] L. Xiong, Q. Zhao and Y. Chen, "SAR-to-Optical Image Translation Using Conditional GANs," *IEEE Trans. Geosci. Remote Sens.,* vol. 61, pp. 1–10, 2023.

[16] M. Liu, J. Zhou and Y. Tang, "Sentinel-1 and Sentinel-2 Data Fusion using U-Net for Land Cover Classification," *ISPRS J. Photogramm. Remote Sens.,* vol. 187, pp. 113–124, 2022.

[17] B. Li and Z. Wang, "Urban Structure Enhancement using Residual Learning on SAR-Optical Fused Imagery," *Comput. Environ. Urban Syst.,* vol. 76, pp. 93–102, 2019.

[18] R. Ding, X. Hu and C. Wang, "Benchmark Study of Fusion Methods for Multisensor Imagery: IHS, PCA, Wavelet, and HIS+NSCT," *Remote Sens.,* vol. 17, no. 2, pp. 225–240, 2025.

[19] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza and J. Chanussot, "Multimodal Fusion Transformer for Remote Sensing Image Classification," *IEEE Trans. Geosci. and Remote Sens.*, vol. 61, pp. 1-20, 2023.

[20] *Statistical Yearbook of Vietnam 2023*, General Statistics Office, Hanoi: Statistical Publishing House, 2024.

[21] K. Lewińska, S. Ernst, D. Frantz, U. Leser and P. Hostert, "Global Overview of Usable Landsat and Sentinel-2 Data for 1982–2023," *Data in Brief*, vol. 57, pp. 111054, 2024.

[22] A. Doerry, "Introduction to Synthetic Aperture Radar," in *2019 IEEE Radar Conf.*, Boston, MA, USA, pp. 1–90, 2019.

[23] R. Torres, P. Snoeij, D. Geudtner, et al., "GMES Sentinel-1 Mission," *Remote Sens. Environ.*, vol. 120, pp. 9–24, 2012.

[24] Y. Zhang, Y. Ban and Y. Hu, "Fusion of Sentinel-1 and Sentinel-2 Time Series for Urban Land Cover Classification using A Dual-Branch CNN," *ISPRS J. Photogramm. Remote Sens.*, vol. 164, pp. 230–243, 2020.

[25] Y. Zhang, J. Liu and D. Li, "Remote Sensing Image Fusion using Deep Learning: A Review," *Inf. Fusion*, vol. 52, pp. 20–34, 2019.

[26] Y. Liu, L. Wang and M. Xu, "SAR and Optical Image Fusion via Deep Residual Networks for Urban Mapping," *ISPRS J. Photogramm. Remote Sens.*, vol. 178, pp. 110–124, 2021.

[27] Z. Cheng, H. Sun, M. Takeuchi and J. Katto, "Deep Convolutional AutoEncoder-based Lossy Image Compression," in *Proc. the 2018 Pic. Cod. Symp.*, San Francisco, CA, USA, pp. 253-257, 2018.

[28] V. Turchenko, E. Chalmers and A. Luczak, "A Deep Convolutional Auto-Encoder with Pooling–Unpooling Layers in Caffe," *Int. J. Comput.*, vol. 18, no. 1, pp. 8–31, 2019.

[29] Y. Xu, K. Fu and L. Zhang, "A Multi-scale Feature Fusion Network for Cloud Removal in Remote Sensing Images," *ISPRS J. Photogramm. Remote Sens.,* vol. 161, pp. 182–194, 2020.

[30] Z. Alzamili, K. Danach and M. Frikha, "Deep Learning-Based Patch-Wise Illumination Estimation for Enhanced Multi-Exposure Fusion," *IEEE Access*, vol. 11, 120642 - 120652, 2023.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[32] K. Guo, X. Li, H. Zang and T. Fan, "Multi-Modal Medical Image Fusion Based on FusionNet in YIQ Color Space," *Entropy*, vol. 22, no. 12, pp. 1423, 2020.

[33] Z. Wang and A. C. Bovik, "A Universal Image Quality Index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, 2002.

[34] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Syst. Techn. J.*, vol. XXVII, no. 3, pp. 379-423, 1948.

[35] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th Edn., Pearson, New York, 2018.

[36] A. M. Eskicioglu and P. S, Fisher, "Image Quality Measures and Their Performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, 1995.

[37] C. Xydeas and V. Petrovic, “Objective Image Fusion Performance Measure,” *Electron. Lett.*, vol. 36, no. 4, pp. 308–309, 2000.

[38] Y. Luo, J. Zhang and H. Li, “DAFCNN: Dual-Attention Fusion Convolutional Neural Network for SAR And Multispectral Image Fusion,” *Remote Sens. Lett.,* vol. 14, no. 5, pp. 435–450, 2023.

[39] Y. Yuhendra and M. Minarni, “Optical SAR Images Fusion: Comparative Analysis of Resulting Images Data,” *MATEC Web of Conf.*, vol. 215, pp. 01002, 2018.