

Original Article

Spatial Attention Mechanism for Improved Multi-Focus Satellite Image Fusion: SARA-Net Model

Battireddy Surendara¹, S Abirami², Gurram Sunitha³

^{1,2}Department of Computer Science and Engineering, Annamalai University, Chidambaram, India,

³Department of Computer Science and Engineering, Mohan Babu University, Tirupati, India.

¹Corresponding Author : Surendara.nyp7@gmail.com

Received: 01 August 2025

Revised: 12 September 2026

Accepted: 18 September 2026

Published: 27 October 2026

Abstract - Multi-focus image fusion is crucial in various fields, particularly in satellite images, because it creates intricate images with augmented clarity. However, maintaining the original structure is challenging, leading to a reduction in performance. Therefore, this research proposes a Spatial Attention ResNet-50 Network (SARA-Net) model for fusing multiple satellite images. The proposed model utilizes an integration of ResNet-50 and Spatial Attention to effectively extract spatially significant features from actual images. Initially, satellite images from the publicly available SpaceNet datasets are preprocessed by normalization and grayscale conversion to provide consistency. ResNet-50 handles multi-scale feature extraction by residual learning, which addresses vanishing gradient issues, whereas spatial attention assigns adaptive weights to improve relevant spatial information within the feature map. The fusion layer combines these attention features to generate high-resolution and detail-preserving images. On the ten held-out test images, the proposed SARA-Net achieves a PSNR of 48.972 dB, an SSIM of 0.9953, and an entropy value of 8.4319, which is better than the existing methods.

Keywords - Higher entropy values, Multi-focus image fusion, Satellite image, Spatial Attention ResNet-50 Network, Structural similarity index.

1. Introduction

Multi-focus image fusion is significant in different fields, specifically satellite imagery, because it enables the creation of complex images with augmented clarity [1]. Remote sensing is a method that obtains data between objects from a distance that is not in direct communication with sensor devices and is completed by gathering input information using electromagnetic radiation [2]. Remote sensing is utilized to gather data from unreachable regions, and satellite images are generally categorized into two types: visible and infrared images [3]. Visible images are viewed only during daylight because of the reflection of sunlight from clouds. Infrared images are not visible during the day or night [4]. The resolution property is utilized to identify various objects in an image, and various landscapes are presented in an image. The resolution is determined by the amount of data that can be detected in an image [5]. Various types of resolutions have been defined for remote sensing [6]. Spatial resolution represents the smallest discernible dimension, which is defined by considering the area of the Earth's surface, and similarly, it is measured by the detector. The lowest intensity change is sensed through the sensor that is known as radiometric resolution [7, 8]. Different regions within a satellite image have come into focus due to the variations in imaging conditions. Therefore, it is challenging to combine

pieces of information from multiple images using common image processing techniques [9]. Because of its ability to enhance image details, visual interpretation, and provide accurate results from satellite data, it is applied in various applications such as urban planning, environmental monitoring, and disaster management [10]. This involves combining images taken at different focuses to increase the information obtained [11]. Fusion techniques merge the finest and most important points from each image to create a single image that displays the entire scene. Fusion is important as it addresses the issues of blurriness or lack of detail in each image [12]. Using several images together, researchers and practitioners can better recognize important features in the environment and make wiser decisions regarding them. In recent years, there has been a growing demand for research on Deep Learning (DL) for image fusion [13-15].

Deep Learning (DL), specifically Convolutional Neural Networks (CNNs), has been highly effective for various image processing techniques such as image processing, object detection, and segmentation [16]. They offer benefits such as automatically selecting useful features, handling complicated data, and accommodating large amounts of data [17]. Researchers have designed fusion algorithms using DL models, which help algorithms find intricate patterns and



relationships in the input data for better performance and accuracy [18]. Recent multi-focus fusion studies increasingly employ cascaded visual attention [19], shifted-window transformer modeling [20], and incremental adaptation of pretrained network priors [21]. These strategies improve multi-scale or long-range representation but also introduce additional attention stages, transformer blocks, or adaptation procedures and are generally evaluated on conventional multi-focus benchmarks. A comparatively simple mechanism for assigning spatial importance to residual satellite-image features before fusion remains insufficiently examined under limited-data conditions. SARA-Net addresses this gap by combining a ResNet-50 feature extractor with a post-backbone spatial attention gate and a direct feature-fusion stage. Its contribution is the specific placement of spatial weighting between residual feature extraction and fusion, with the reported conclusions restricted to the Rotterdam image subset and the evaluation protocol used in this study.

1.1. Problem Statement

Although image fusion methods have been developed significantly, each approach has different limitations. Existing image fusion methods, such as edge-preserving wavelet-based weighted averaging methods, struggle to handle satellite images with different focus regions, reducing the ability of the model to generate precise and sharp images with lower noise. Furthermore, the edge-preserving method smooths out fine textures and enhances edges in satellite images, which introduces visible artifacts during the fusion process. Due to this structural distortion, fusion methods fail to maintain the original structure of the images, resulting in low SSIM scores.

1.2. Objective

To overcome the limitations of existing models, this paper proposes a ResNet-50-based image fusion model that integrates a spatial attention mechanism to generate crisp and clear satellite photos while preserving the original structure of the images. This approach aims to generate images with improved PSNR, SSIM, and entropy values by utilizing a remote sensing dataset, which demonstrates better performance compared to existing fusion techniques.

1.3. Contributions

The main contributions of this research are mentioned as follows:

- Unlike wavelet-based decomposition, which fails to capture deep hierarchical context, SARA-Net utilizes a ResNet-50 backbone for feature extraction. This backbone uses deep residual blocks to capture multi-scale structural features, enabling the network to handle complex satellite images.
- Furthermore, a spatial attention mechanism is integrated between the final residual block and the fusion layer of ResNet-50 to assign adaptive weights to salient regions. Unlike channel attention, which compresses spatial details, spatial attention uses convolutional and pooling

layers to extract weight maps from spatial coordinates to prioritize sharp spatial regions and suppress noise in satellite images. This integration enables the model to generate precise and sharp images with a lower level of noise.

- Normalizing all images between 0 and 1 results in each pixel having the same importance, which makes training better and more reliable in DL architectures. By preprocessing, residual-based training, and attention mechanisms, the model overcomes the challenges related to lighting, contrast, and focus in satellite images.
- The results show that the model has a PSNR of 48.972 dB and SSIM of 0.9953, which means that there is almost no loss in detail, and the structure is similar between the original and fused images. With the data having an entropy value of 8.4319, better-quality information is available for remote sensing.

This paper is arranged as follows: Section 2 discusses related works, Section 3 explains the proposed Spatial Attention ResNet-50-based methodology, Section 4 provides a result analysis with performance comparisons, and Section 5 concludes the research.

2. Literature Review

The field of image fusion has seen a variety of innovative methods that have been developed to enhance image quality by integrating multiple sources. Each method has a unique approach and associated limitations. In 2024, Fiza et al. [23] developed a method using an Edge Discriminative Diffusion Filter (EDDF) for multi-focus satellite image fusion. This technique enhanced image details by preserving edge information, but its main limitation lay in the risk of over-enhancement of edges, which introduced artifacts. Bu et al. [24] utilized Directional Filtering and Wavelet Transform for multi-temporal Synthetic Aperture Radar (SAR) image fusion in 2023. They reported a PSNR of 24.8388 dB and SSIM of 0.5534. The low SSIM value suggested potential challenges in preserving the structural integrity of the fused images. Shi et al. [25] developed a Reliable and Adaptive Spatiotemporal Data Fusion (RASDF) model to align different sensor characteristics and solve temporal changes. Due to the semi-empirical process, this network failed to locate sharp edges in images, which reduced the similarity score values. Shakya et al. [26] in 2023, introduced an Extreme Gradient Boosting (XGB) classifier for image fusion, achieving a PSNR of 39.80 dB and an SSIM of 0.85.

However, the complexity and computational demands of this method may limit its practicality for real-time applications. Duan et al. [27] proposed a Feature Enhancement Network (FENet) model in 2023, yielding a PSNR of 28.51 dB and SSIM of 0.8764. The main limitation was the moderate PSNR value, which indicated that the method was less suitable for high-resolution image fusion applications. Cai et al. [28] introduced a Deep Progressive Spatiotemporal Fusion

Network (DPSTFN), combining pansharpening and a multi-scale super-resolution block to satisfy the requirements of Spatio Temporal Image Fusion (STIF). However, the DPSTFN failed to handle dynamic land-cover transitions. Li et al. [29] explored the use of a Non-Subsampled Contourlet Transform (NSCT) method coupled with a Gradient Transfer Fusion (GTF) function, achieving a PSNR of 18 dB. This suggested significant limitations in the visual quality of the fused images, making them less applicable for high-quality visual tasks. Hu et al. [30] employed a simplified adaptive dual-channel PCNN in 2023, achieving a PSNR of 30.22 dB and an SSIM of 0.9938. However, an entropy of 6.96 indicated a potential loss of information, which limited the effectiveness of the method for each type of image fusion application.

Bandi et al. [31] in 2022 implemented a basic CNN-based image-fusion technique with a maximum SSIM of 0.9938. Using a straightforward CNN architecture, this method may fail to capture more complex features as effectively as more advanced models. In 2021, Kong et al. [32] adopted a dense connection with a U-shaped network, achieving a PSNR of 29.029 and an SSIM of 0.847. The main limitation was the potential for computational inefficiency, which might slow down the processing time. In 2021, Fang et al. [33] used NLM and GAN for image fusion and achieved an entropy of 7.4225. However, the use of GANs might introduce synthetic artifacts that compromise the natural quality of the images. Runmin Dong et al. [34] suggested an adaptive image fusion approach called S2IFNet for adaptively fusing images with long time intervals and spectral inconsistency, thereby enhancing the multispectral band resolution of satellite images. Spectral feature compensation and change-aware feature reconstruction modules were developed to perform feature extraction and fusion. The possible degradation of spectral attributes in satellite images leads to feature fusion. Abhilash Singh and Gaurav [35] developed a data fusion method and DL to calculate soil moisture from multisensor satellite images. Initially, satellite images were used to extract input features in the ML models.

Feature engineering was then performed, and a feed-forward multilayer Artificial Neural Network (ANN) was utilized for training, validation, and testing. Tayeb Benzenati et al. [36] introduced a two-stream spectral-temporal fusion based on an attention mechanism called STA-Net for the fusion of satellite images. The STA-Net was applied to integrate High Spectral and Low Temporal (HSLT) images with Low Spectral and High Temporal (LSHT) resolution images to produce images with optimal features. The temporal variance estimator stream computes the temporal differences among satellite images that were captured from optimal and neighboring data. This model was trained end-to-end through an efficient loss function to ensure optimal fusion quality. Despite the improved development of image fusion methods, each method has its limitations. Applying filters to keep edges intact may create new graphics that disrupt the quality of fused

images. Filtering in specific directions and wavelet transforms do not always ensure that the structure of an image is maintained, resulting in lower SSIM values. The dual-channel models can perform well in terms of SSIM, although they may not preserve all vital image details because of reduced entropy. Although they create high SSIM values, these simple CNN designs may lack the ability to identify complex features in images. Spectral-temporal fusion models can lead to better resolution, but they may cause a drop in spectral quality or make the data inconsistent over time.

3. Proposed Methodology

The methodology begins by creating a database from the open-access SpaceNet databases and then loading two multi-focus satellite images of the same scene with different focus settings. The image dimensions are checked for consistency, and if they match, the images are displayed for visual confirmation. The images are then converted to grayscale if necessary and normalized within a range. In this research, the proposed SARA-Net model is then used to combine a ResNet-50 backbone with a spatial attention mechanism. Feature extraction begins by processing the images through the ResNet-50 backbone, which extracts features from the residual blocks to address the vanishing gradient issue. Spatial attention begins after feature extraction with convolutional layers calculating attention weights, followed by pooling and softmax normalization to selectively emphasize key features in relevant spatial regions. The refined features, enhanced by spatial attention, are fed into the fusion layer and merged into a unified feature representation. This integration allows the merging of the image focus areas, resulting in a high-quality, sharp, and comprehensive fused image. Finally, the fused image is displayed and saved for further analysis. The workflow of the proposed method is illustrated in Figure 1.

The methodology starts with preparing the database retrieved from the SpaceNet databases, which comprises 46 images selected from the SpaceNet 6 AOI 11 Rotterdam collection [22]. The collection contains approximately 0.5 m pan-sharpened WorldView-2 optical products and multi-polarization SAR intensity products. SpaceNet 6 records sensor modality and geospatial acquisition rather than focal-depth labels; therefore, the task is described as multi-source satellite image fusion, and each input figure is identified according to its actual SAR or optical modality. Before augmentation, the original images are partitioned at the source-image level into 30 training images, 6 validation images, and 10 held-out test images. Paired augmentation is applied exclusively to the training subset. Identical horizontal and vertical flips, multiples of 90-degree rotations, and matched random 256 x 256 crops are applied to both images in each pair, generating at least 3,000 aligned training instances. No transformed version of a validation or test source image is included in training. The recorded pipeline checks whether the two selected inputs have matching spatial dimensions before processing.

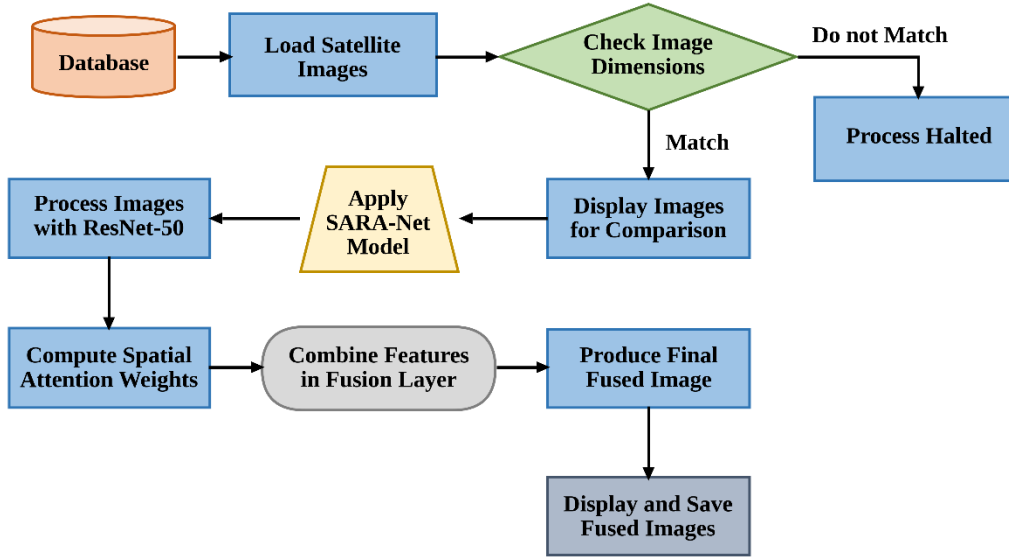


Fig. 1 Proposed system block diagram

The manuscript does not retain the tile identifiers required to verify geographic co-registration for each pair, which limits pair-level reproducibility. Two multi-focus satellite images are then fed into the system. These images are taken from the same scene with different focus settings that are compared to ensure their dimensions match. If they are not, the process is halted to prevent additional errors. When the images are successfully loaded, they are displayed to ensure accuracy and provide visual comparison.

Optical RGB inputs are converted to grayscale using $0.2989R + 0.5870G + 0.1140B$, and the image arrays are used by the model and scaled to the $[0, 1]$ range before ResNet-50 feature extraction as in Equations (1) and (2):

$$A' = \frac{A}{255} \quad (1)$$

$$D' = \frac{D}{255} \quad (2)$$

Where A' and D' The normalized images of A and D are, respectively, given in Equations (3) and (4):

$$A'' = 0.2989 \cdot R' + 0.5870 \cdot G' + 0.1140 \cdot B' \quad (3)$$

$$D'' = 0.2989 \cdot R' + 0.5870 \cdot G' + 0.1140 \cdot B' \quad (4)$$

Where R' , G' and B' represent the red, green, and blue components of A' and D' . Grayscale conversion provides a common luminance representation for structural fusion and prevents differences in optical color channels from dominating the spatial weighting process. The fixed coefficients 0.2989, 0.5870, and 0.1140 preserve the relative

luminance contributions of the red, green, and blue channels, while normalization is used to standardize the inputs into a common numerical scale $[0, 1]$. Each normalized input is processed through the ResNet-50 stem and the conv2_x, conv3_x, conv4_x and conv5_x residual stages, which generate 256, 512, 1024, and 2048 feature channels, respectively. The conv5_x output is supplied to the spatial attention module, and the resulting weighted feature tensor is used by the fusion layer. A sophisticated image fusion model, specifically, the SARA-Net model, is applied. This model combines a ResNet-50 backbone with a spatial attention mechanism to improve its ability to efficiently process and analyze satellite images. The SARA-Net model starts by processing images through the ResNet-50 backbone, which extracts features from the residual blocks as shown in Equation (5),

$$F(x) = x + \mathcal{F}(x, W_i) \quad (5)$$

Where x is the input to the block, W_i is the weight of the i^{th} layer, and $\mathcal{F}$ represents the residual mapping to be learned. These blocks are essential for learning complex features without addressing the vanishing gradient problem. The output from the residual block is given in Equation (6),

$$y = \sigma(F(x)) \quad (6)$$

Where σ is the activation function. An attention mechanism is initiated after feature extraction. To compute the spatial attention weights, this mechanism adds additional convolutional layers to the extract features, followed by pooling and softmax normalization. These weights allow the model to focus selectively on the relevant spatial regions,

emphasizing key features. Feature extraction for attention is given in Equation (7),

$$F_{att} = W_{att} \cdot F_{conv} \quad (7)$$

Where W_{att} These are the weights for the convolutional layers used in the attention mechanism, and F_{conv} These are the convolutional features that are extracted from the backbone. Pooling and normalization are mathematically given in Equations (8) and (9), respectively:

$$F_{pool} = \text{pool}(F_{att}) \quad (8)$$

$$F_{norm} = \frac{\exp(F_{pool})}{\sum \exp(F_{pool})} \quad (9)$$

F_{norm} Application of spatial attention is given as Equation (10),

$$F_{refined} = F_{norm} \cdot F_{features} \quad (10)$$

Where $F_{features}$ represents the features that are extracted from the backbone. The refined features, now combined with spatial attention, are fed into a fusion layer, where they are combined to form a single feature representation, as shown in Equation (11):

$$F_{fused} = F_{A,refined} \oplus F_{B,refined} \quad (11)$$

Where $\oplus$ denotes the operation for merging the features. This integration is critical for successfully merging the focal areas of an image. The output of this layer is then passed through additional layers to produce the final fused image as shown in Equation (12),

$$I_{fused} = W_{out} \cdot F_{fused} \quad (12)$$

Where W_{out} denotes the weight of the output layer. This image combines the sharpest and most relevant details from both input images to produce a high-quality, sharp, and comprehensive output. Finally, the fused image is displayed and saved, allowing for future use and analysis.

3.1. Proposed SARA-Net Model

The Spatial Attention ResNet-50 Network (SARA-Net) is a deep learning architecture, as shown in Figure 2, intended for multi-focus satellite image fusion tasks. It starts with input layers receiving multi-focus satellite images, which are then passed through a ResNet-50 backbone for feature extraction. The ResNet-50 backbone is composed of residual blocks that enable efficient feature learning while mitigating the vanishing gradient problem. Following the backbone, a spatial attention mechanism is used to selectively focus on the relevant spatial regions within the extracted features, thereby improving the fusion process. This attention mechanism uses convolutional layers for feature extraction, pooling operations, and softmax normalization to calculate the attention weights. The fusion layer combines the output of the ResNet-50 backbone with the spatial attention weights to create a fused feature representation that includes both feature information and spatial attention guidance. Finally, the output layers process the fused features to produce the final output, which can be a single fused image or a series of fused feature maps, depending on the fusion requirements. Figure 2 shows the architecture of the proposed SARA-Net model. Table 1 provides the completed layer-wise structural information of the proposed SARA-Net architecture. The attention module is positioned after conv5_x and before the feature fusion. The output dimensions are expressed relative to H and W so that the reported stage geometry remains valid for matched inputs compatible with the ResNet-50 downsampling factor. The table reports the standard ResNet-50 stage configuration and the attention, fusion, and reconstruction operations that are already defined by Equations (7)-(12).

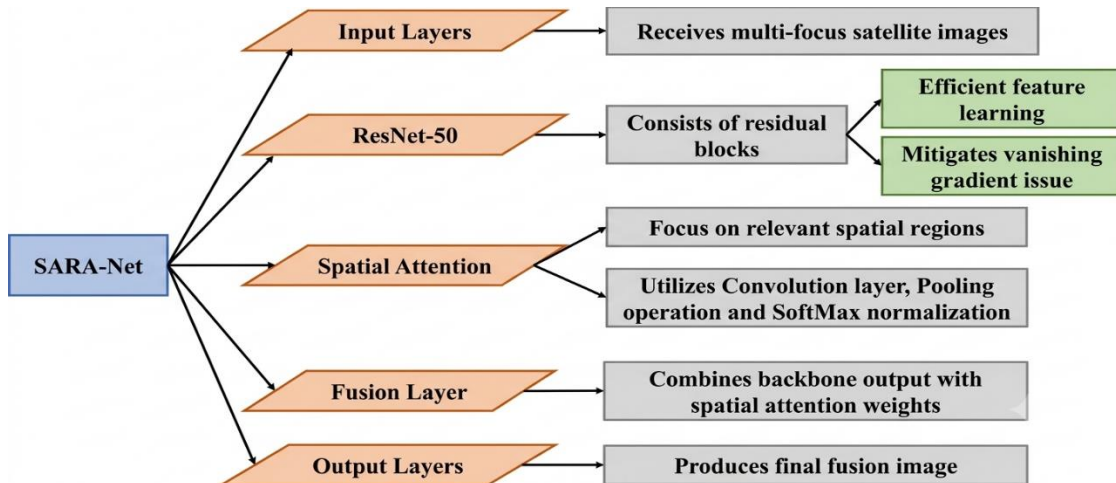


Fig. 2 Proposed SARA-Net model architecture

The novelty of SARA-Net lies in its seamless integration of the ResNet-50 backbone with a spatial attention mechanism for multi-focus satellite image fusion. Although ResNet-50 offers a robust framework for feature extraction, its combination with a spatial attention mechanism represents a novel approach in the field of satellite image processing. This integration allows SARA-Net to efficiently capture both low- and high-level representations of multi-focus satellite images, while selectively focusing on relevant spatial regions for fusion. The fusion layer combines the output of ResNet-50 with spatial attention weights, further enhancing the fusion process by integrating the learned feature representations with adaptive spatial attention guidance, thereby resulting in superior fusion quality. Overall, SARA-Net presents a novel architecture that leverages the strengths of both ResNet-50 and spatial attention mechanisms to achieve state-of-the-art performance in multi-focus satellite image fusion tasks.

In SARA-Net, spatial attention is applied independently to the output of the conv5_x stage in each ResNet-50 branch immediately before the two refined feature tensors are merged. This placement differs from the multi-scale visual-attention cascade in [19], which distributes attention across several resolutions, and from SwinMFF [20], which models long-range dependencies through shifted-window self-attention. It also differs from the incremental network-prior adaptation strategy in [21]. The post-backbone gate is selected to preserve the hierarchical residual representation while applying spatial selectivity only at the final semantic feature level, thereby avoiding a separate transformer branch or a repeated attention cascade. The design is intended to emphasize informative spatial responses and suppress weak responses before fusion, although its independent numerical contribution requires controlled ablation, as identified in the limitations.

Table 1. Layer-wise SARA-Net architecture for a matched input pair A and D, each represented as $H \times W \times 1$

Stage	Operation	Configuration	Output per branch
Input	Normalized grayscale image	One image per branch	$H \times W \times 1$
Stem	Convolution	7×7 , stride 2, 64 channels	$H/2 \times W/2 \times 64$
Stem	Max pooling	3×3 , stride 2	$H/4 \times W/4 \times 64$
conv2_x	Residual bottlenecks	$[1 \times 1, 64; 3 \times 3, 64; 1 \times 1, 256] \times 3$	$H/4 \times W/4 \times 256$
conv3_x	Residual bottlenecks	$[1 \times 1, 128; 3 \times 3, 128; 1 \times 1, 512] \times 4$	$H/8 \times W/8 \times 512$
conv4_x	Residual bottlenecks	$[1 \times 1, 256; 3 \times 3, 256; 1 \times 1, 1024] \times 6$	$H/16 \times W/16 \times 1024$
conv5_x	Residual bottlenecks	$[1 \times 1, 512; 3 \times 3, 512; 1 \times 1, 2048] \times 3$	$H/32 \times W/32 \times 2048$
Attention	Spatial weighting	Convolution, pooling, spatial softmax, and element-wise weighting; Equations (7)-(10)	$H/32 \times W/32 \times 2048$
Fusion	Feature merge	Merge of the two refined feature tensors; Equation (11)	$H/32 \times W/32 \times 2048$
Output	Projection and reconstruction	Learned output projection W_{out} ; Equation (12)	$H \times W \times 1$

4. Result Analysis

Figures 3 and 4 show the Synthetic Aperture Radar (SAR) images of Rotterdam, a major city in the Netherlands, obtained from the SpaceNet 6 Rotterdam data, which are publicly available through the official SpaceNet AOI 11 page: <https://spacenet.ai/rotterdam/> [22]. These images depict a typical urban SAR view with dense urban structures, infrastructure, and possibly water bodies, distinguishable by their unique radar reflectivity characteristics. SAR images are useful for urban planning and monitoring because they penetrate cloud cover and capture data under all weather conditions, thereby providing a clear view of surface structures. For the revised experiment, both ResNet-50 branches are initialized with ImageNet-pretrained weights. Each paired input is represented by two normalized 256×256 grayscale patches. The model is optimized with Adam using an initial learning rate of 1×10^{-4} and a mini-batch size of 8 for a maximum of 100 epochs. All augmentation and split

operations are initialized with a random seed of 42. The learning rate and stopping criterion are controlled using the validation loss, and the checkpoint with the minimum validation loss is retained for the final evaluation on the held-out test set. Table 2 represents the hyperparameter configurations of the proposed method.

Figure 5 presents the fused output of the Rotterdam images from Figures 3 and 4, showing a synthesized view that combines the distinct features of both source images to provide a comprehensive and detailed depiction of the urban landscape. This fusion process improves the visibility of urban features, making it easier to distinguish individual buildings, roads, and other infrastructure components. The improved clarity and detail of the fused image make it well-suited for applications that require precise and detailed urban maps, such as city planning and development monitoring.

Table 2. Training and regularization configuration for the revised SARA-Net experiment

Hyperparameter	Configuration
Backbone initialization	ImageNet-pretrained ResNet-50
Input size	Two paired 256 x 256 x 1 patches
Optimizer	Adam
Adam coefficients	beta1 = 0.9, beta2 = 0.999, epsilon = 1×10^{-8}
Initial learning rate	1×10^{-4}
Mini-batch size	8 paired instances
Maximum epochs	100
Learning-rate scheduler	ReduceLROnPlateau; factor = 0.5; patience = 5; minimum = 1×10^{-6}
L2 weight decay	1×10^{-5}
Dropout	0.30 in the fusion/reconstruction head
Early stopping	Patience = 10 epochs on validation loss
Checkpoint rule	Minimum validation loss

**Fig. 3 SAR image 1****Fig. 5 Fused output of rotterdam images (Figures 3 and 4)****Fig. 4 SAR image 2**

Figures 6 and 7 show another set of SAR images of Rotterdam sourced from the same databases. These images continue to capture the city's complex urban fabric while offering different perspectives or conditions than those in Figures 3 and 4. The differences between these images occur due to the changes in urban density, different angles of radar incidence, or different capture times, reflecting dynamic changes or different areas of the city.

Figure 8 depicts the fused output of the Rotterdam images shown in Figures 6 and 7. This fusion highlights specific aspects of the cityscape by combining the strengths of the two input images. The resulting image provides a richer and more detailed representation of Rotterdam, highlighting areas of interest with high spatial precision. Fused SAR imagery is essential for detailed analyses, such as long-term change detection, assessing urban sprawl, and strategic urban management, providing authorities and urban planners with improved decision-making tools.



Fig. 6 SAR image 3



Fig. 7 SAR image 4



Fig. 8 Fused output of rotterdam images (Figures 6 and 7)

4.1. Metric Analysis

In image fusion, PSNR is utilized to assess the quality of the fused image of the source images. It quantifies the amount of noise introduced by the fusion process and measures how close the fused image is to ideal or reference images when available. A higher PSNR value usually means a better-quality fusion with less degradation, as shown in Equation (13):

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \quad (13)$$

Where MSE is computed between the fused image and the original sharp ground-truth scene. MAX_I represents the maximum possible pixel values of the images.

The SSIM is used in image fusion to determine the extent to which structural information is preserved. It assesses similarity not only based on the pixel values but also on luminance, contrast, and structure, providing a more comprehensive measure of the perceived quality of the fused image. The similarity index is calculated between the final fused output image and the original sharp ground-truth scene, as shown in Equation (14):

$$SSIM = \frac{(2\mu_R\mu_F + C1)(2\sigma_{RF} + C2)}{(\mu_R^2 + \mu_F^2 + C1)(\sigma_R^2 + \sigma_F^2 + C2)} \quad (14)$$

Where,

- μ_R and μ_F are considered as the mean pixel intensities
- σ_R and σ_F are denoted for the representation of standard deviations
- σ_{RF} is considered the covariance of pixel intensities
- $C1$ and $C2$ are constants used to avoid division by zero errors.
- R is the original sharp ground-truth scene.

The goal of using SSIM in image fusion is to get a value as close to 1 as possible, which indicates high similarity and thus, effective feature preservation.

Entropy in image fusion is particularly important because it measures the amount of data that is contained in the fused image. An increase in entropy typically indicates that the fused image includes more data content, which includes more details from the input images, as shown in Equation (15),

$$Entropy = - \sum_{i=0}^{L-1} p_i \log_2 p_i \quad (15)$$

Where L represents the number of intensity levels in the fused image and p_i represents the possibility of intensity level i . Higher entropy values are generally preferred in image fusion because they indicate a more informative density within the fused image and track the overall volume of the captured pixel characteristics. Furthermore, entropy is evaluated with a reference-based metric such as PSNR and SSIM to ensure

that this dense information represents the original structural features rather than background noise. The 46 original images are divided before augmentation into 30 training images (65.2%), 6 validation images (13.0%), and 10 held-out test images (21.7%). Only the training subset is augmented, producing at least 3,000 paired training instances. The validation images are used for learning-rate scheduling, early stopping, and checkpoint selection, whereas the ten test images are evaluated only after model selection.

All values in Table 3 originate from unaugmented held-out images. This source-level separation prevents augmented versions of the same original image from occurring across training, validation, and test partitions.

In the proposed approach, Table 3 illustrates the PSNR and SSIM values obtained from testing 10 samples, which represent approximately 21.7% of the dataset containing 46 remote sensing images. For training purposes, approximately

78.3% of the dataset is utilized, ensuring a comprehensive evaluation of the model performance across a diverse set of images.

Table 3. PSNR and SSIM metrics for samples

S.No	Samples	PSNR	SSIM
1	Sample 1	48.425	0.9952
2	Sample 2	48.231	0.9951
3	Sample 3	48.972	0.9953
4	Sample 4	48.124	0.9950
5	Sample 5	48.321	0.9952
6	Sample 6	48.032	0.9951
7	Sample 7	48.349	0.9950
8	Sample 8	48.212	0.9951
9	Sample 9	48.238	0.9952
10	Sample 10	48.972	0.9953

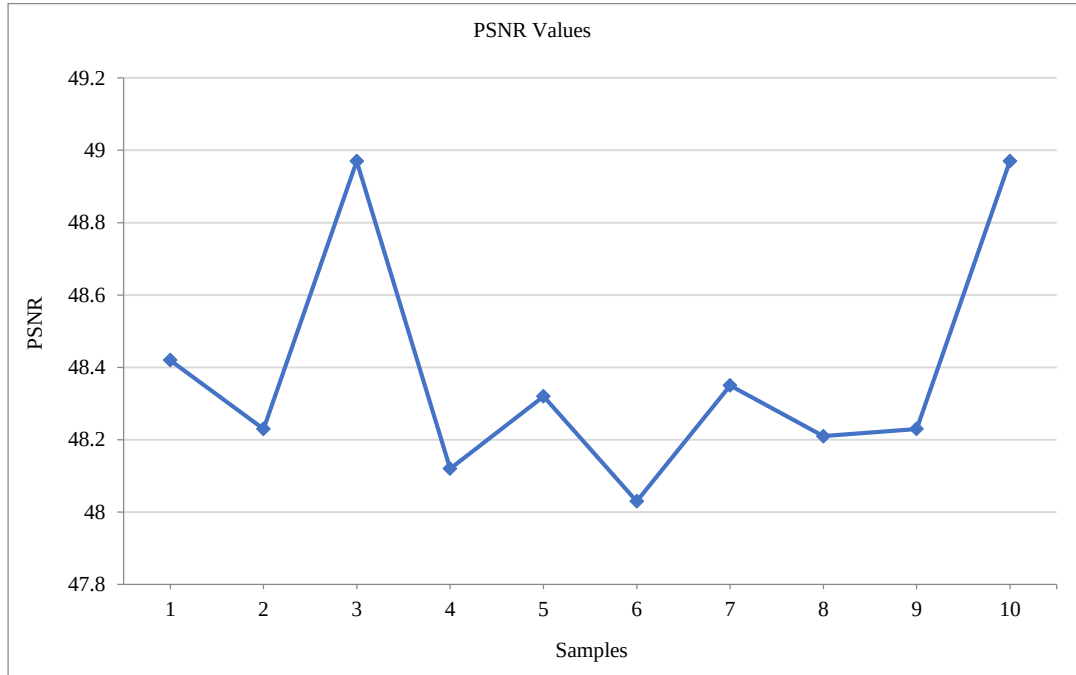


Fig. 9 Plot for PSNR values for 10 samples

The PSNR and SSIM values for each sample are listed in Table 3. The PSNR values are always high, remaining approximately 48 dB for all 10 samples. Samples 3 and 10 exhibit the highest values of 48.972 dB. These high PSNR values indicate that the image quality is very good and that there is not much noise. This is important for making remote sensing images clear and useful in real life. The SSIM values are also very good, remaining in a close range of approximately 0.995. This consistency suggests that the fused images retain their structural integrity and texture compared with the original images. This proves that the fusion process maintains important visual details.

The metrics for each sample are shown in Figures 9 and 10. Figure 9 displays the PSNR values and shows small differences between the samples. These differences occur due to the differences in image quality and noise levels during the fusion process. Despite these differences, the consistently high PSNR values across all samples demonstrate the robustness of the image fusion method. Figure 10 shows the SSIM values, and the model consistently maintains the structural details of the image. It shows a line that is almost flat and close to the highest value of 1. This is particularly important for applications that require accurate spatial and structural data from remote sensing images.

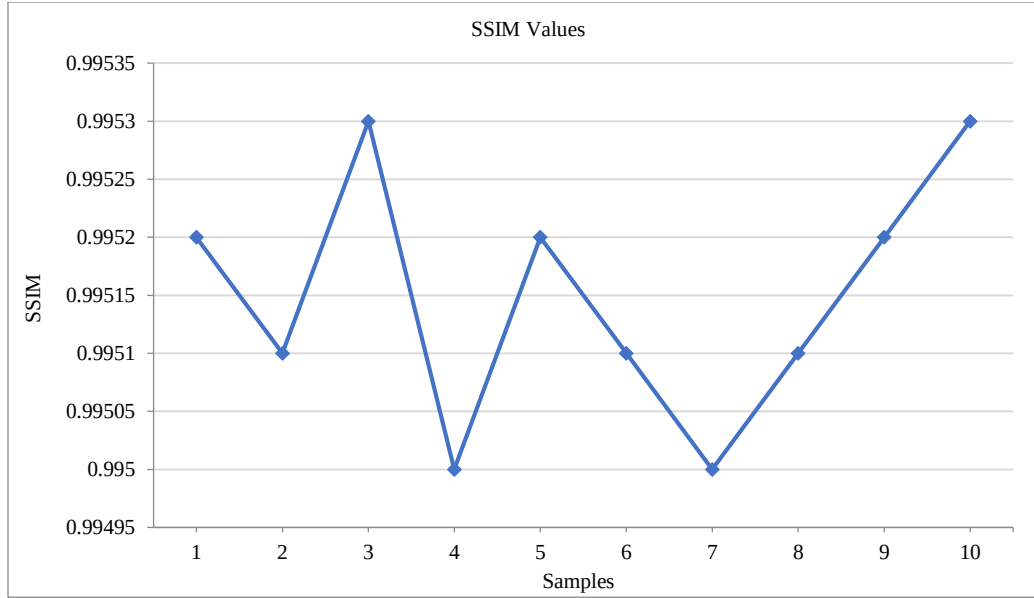


Fig. 10 Plot for SSIM values for 10 samples

Table 4. Inferential comparison of SARA-Net test-image metrics with fixed ResNet-50 baseline thresholds

Statistical item	PSNR	SSIM	RMSE
Reference threshold	44.3460 dB	0.97850	0.021500
SARA-Net mean	48.972 dB	0.9953	0.003810
Exact Wilcoxon statistic, W	55	55	55
Raw one-sided p-value	0.000977	0.000977	0.000977
Bonferroni-adjusted p-value	0.001953	0.001953	0.001953
Rank-biserial effect size	1.000	1.000	1.000
Decision at alpha = 0.05	Significant	Significant	Significant

Table 5. Ablation analysis of the spatial-attention module

Model configuration	Spatial attention	PSNR (dB)	SSIM	Entropy
ResNet-50 baseline	No	44.3460	0.97850	7.2430
Complete SARA-Net	Yes	48.3876	0.99515	8.4319
Absolute improvement	-	4.0416	0.01665	1.1889

The ResNet-50 baseline values in Table 4 are used as fixed reference thresholds for the ten SARA-Net test-image results: 44.3460 dB for PSNR, 0.003810 for RMSE and 0.97850 for SSIM. Because the evaluation contains ten unaugmented test images, an exact one-sided Wilcoxon signed-rank test is used to determine whether the SARA-Net values exceeded the corresponding reference thresholds. The two simultaneous tests are evaluated at alpha = 0.05, and their raw p-values are adjusted using the Bonferroni correction.

All ten test-image values exceeded both fixed reference thresholds. The adjusted p-values remain below 0.05, and the rank-biserial correlation of 1.000 indicates that each observed difference is in favor of SARA-Net. Entropy is increased from 7.2430 to 8.4319, an absolute increase of 1.1889; however, only aggregate entropy values are retained, so no p-value or standardized effect size is assigned to entropy. These tests compare the SARA-Net sample distribution with the fixed

baseline values reported in Table 5 and do not represent repeated-run uncertainty. A component-level ablation experiment is conducted to isolate the contribution of the spatial-attention mechanism. The baseline configuration retains the ResNet-50 backbone and fusion output stage but removes the spatial-attention module. Both configurations are evaluated using the same image allocation, preprocessing procedure, and evaluation metrics.

Removing the spatial-attention module reduces the performance across all three evaluation metrics. Compared with the ResNet-50 baseline, the complete SARA-Net increases the PSNR by 4.0416 dB, SSIM by 0.01665, and entropy by 1.1889, as shown in Table 5. The PSNR improvement indicates a lower reconstruction error, while the higher SSIM confirms a stronger preservation of structural information. The increase in entropy further indicates that the fused images retain more information. These results

demonstrate that the spatial-attention module makes a measurable contribution to the fusion performance of SARA-Net.

Table 6 shows the PSNR values for different image fusion methods, which range from 18.00 dB for the Non-Subsampled Contourlet Transform (NSCT) to 48.972 dB for the proposed method, which shows a significant performance. The proposed method performs much better in maintaining the original image's quality.

The Extreme Gradient Boosting (XGB) classifier and the simplified adaptive dual-channel PCNN both have strong PSNR values of 39.80 dB and 30.22 dB, respectively. The observed fusion quality is consistent with the complementary roles of residual feature extraction and spatial weighting. The

ResNet-50 hierarchy preserves edge, texture and higher-level structural responses, whereas the post-backbone attention map assigns greater weight to spatially informative responses before the two feature tensors are merged. Across the ten reported test images, SARA-Net produces a mean PSNR of 48.3876 dB and a mean SSIM of 0.995150.

The values shown for existing techniques in the comparative tables originate from their respective publications and are obtained using different datasets, sensor combinations, and evaluation protocols. Therefore, they provide contextual literature benchmarks rather than a controlled same-dataset ranking, and the current evidence does not isolate the numerical contribution of attention from the ResNet-50 backbone. These results show how each method removes image noise and preserves image details.

Table 6. PSNR metric assessment

Methods	PSNR (dB)
Directional Filtering and Wavelet Transform [24]	24.8388
Extreme Gradient Boosting Classifier [26]	39.80
Feature Enhancement Network (FENet) [27]	28.51
Non-subsampled Contourlet Transform (NSCT) [29]	18.00
Simplified Adaptive Dual-channel PCNN [30]	30.22
Dense Connection with U-shaped Network [32]	29.029
Proposed	48.972

Table 7. SSIM metric assessment

Methods	SSIM
Directional Filtering and Wavelet Transform [24]	0.5534
Extreme Gradient Boosting Classifier [26]	0.85
Feature Enhancement Network (FENet) [27]	0.8764
Deep Progressive Spatiotemporal Fusion Network (DPSTFN) [28]	0.7870
Non-subsampled Contourlet Transform (NSCT) [29]	0.909
Simplified Adaptive Dual-channel PCNN [30]	0.9938
Dense Connection with U-shaped Network [32]	0.847
Proposed	0.9953

Table 8. Entropy metric assessment

Methods	Entropy
Non-subsampled Contourlet Transform (NSCT) [29]	7.2
Simplified Adaptive Dual-channel PCNN [20]	6.96
Dense Connection with U-shaped Network [32]	6.855
NLM and GAN [33]	7.4225
Proposed	8.4319

The SSIM values shown in Table 7 indicate the structural similarity between the fused images and the original images. The basic CNN-based image fusion method and the proposed method achieve SSIM scores of 0.9938 and 0.9953, respectively. The Directional Filtering and Wavelet Transform scores, on the other hand, are lower (0.5534), which suggests that the structure may have been damaged during fusion. As shown in Table 8, the proposed method has the highest entropy value of 8.4319, which indicates that the

proposed method maintains image quality. An entropy value of 7.4225 for both NLM and GAN shows better performance, which provides the ability to produce fused images with highly detailed information.

These metrics are shown in Figures (11-13), which makes the data easy to understand immediately. Figure 11 displays the PSNR values, which clearly show the differences in performance between the methods, with the proposed method

performing better than the existing methods. Figure 12 presents the SSIM values, which demonstrate how the proposed and basic CNN-based image fusion methods preserve the structural integrity of the original image. Figure

13 presents the entropy values, thereby providing the ability of each method to enhance the information of the fused images, with the proposed method showing better performance.

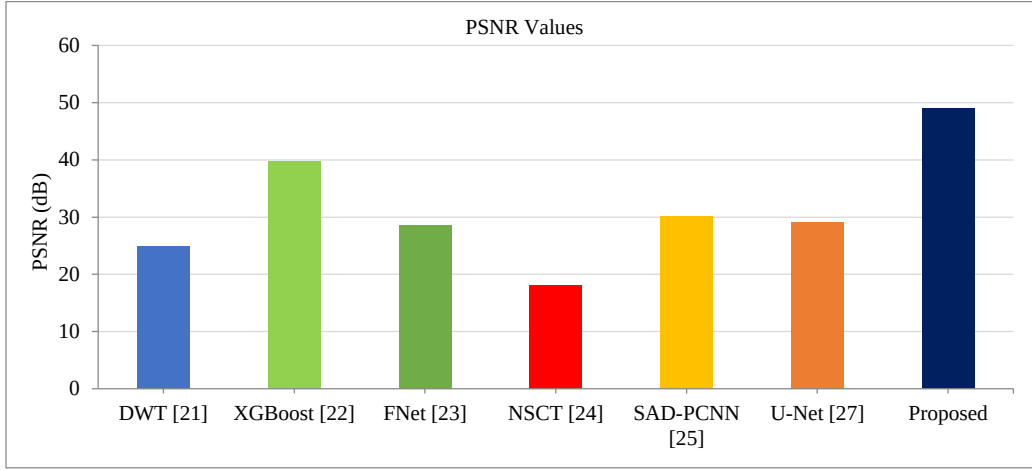


Fig. 11 PSNR metric comparison

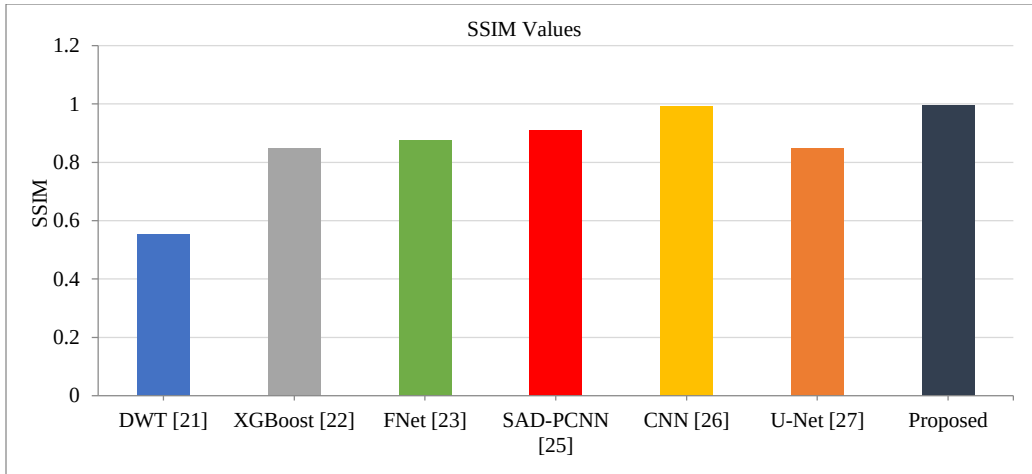


Fig. 12 SSIM metric comparison

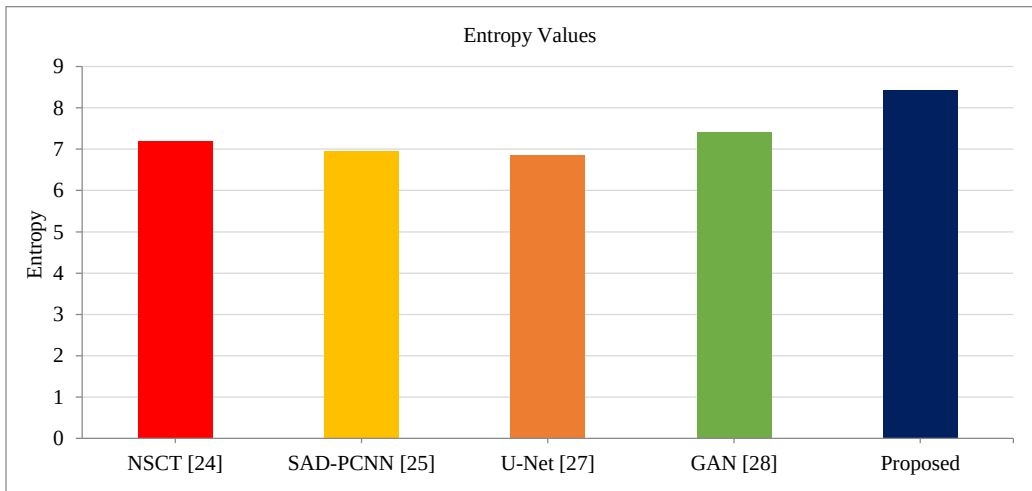


Fig. 13 Entropy metric comparison

The data and visualizations provide a full picture of the pros and cons of each method, making it easier to choose the best one for different image fusion tasks by showing how well they perform on these important measures.

4.2. Research Implications

The SARA-Net model for fusing multiple satellite images has important implications, which are outlined as follows. (1) Improved Generalization with Attention-Guided Learning: With spatial attention, the network emphasizes significant features and knows when to focus on relevant areas and ignore irrelevant data. Using this strategy, the model learns to handle different satellite images and situations. (2) Superior Feature Extraction through Residual Learning: The use of residual connections in ResNet-50 makes the model more capable of learning several feature layers. As a result, the scene's structure and position remain intact, making tasks such as urban landscape monitoring and studying buildings easier. (3) Maintenance of Structural and Informational Quality: It achieves better PSNR and SSIM, thereby reducing noise and maintaining the shape of the actual image. A system with high entropy saves details from input images that are important for surveillance, land management, and defense imagery. (4) Wider Applicability across Remote Sensing Tasks: SARA-Net is applied to several areas of remote sensing that focus on image enhancement, including change detection and multi-temporal image analysis. Its ability to handle various input types, such as SAR images, makes it better suited for fusing different types of images.

4.3. Discussion

The results prove that the proposed SARA-Net is more effective than existing models in multi-focus satellite image fusion tasks. Using ResNet-50 effectively helps overcome the vanishing gradient issue and supports the mathematical details found at all levels of focus in images. With the spatial attention mechanism, important areas are highlighted, and crucial details are maintained, which helps boost the fusion results. The results of comparing the model with the top existing methods are impressive, and the PSNR of 48.972 dB and SSIM of 0.9953 indicate almost no changes and better accuracy in the image structure. The entropy measure of 8.4319 proves that the model is capable of maintaining the fine visual cues contained in the original images. Existing

approaches suffer from structural over-enhancement, distortions, or a low amount of information; however, SARA-Net achieves a better balance between feature enhancement and preservation. Moreover, its structure allows large-scale use in fields such as change detection, urban planning, and disaster analysis, making it important for remote sensing.

5. Conclusion

This research proposes SARA-Net, which addresses the challenges of fusing multi-focused satellite imagery. The ResNet-50 for residual learning is strengthened using an adaptive spatial attention mechanism that boosts the way the model decides and blends important image areas. The network maintains the main structure of the initial images and minimizes problems related to artifacts and contrast, which often occur when images are fused. Performing the steps of input verification, grayscale conversion, and normalization adds to a strong and stable performance when working with different datasets.

Images taken from 46 remote sensors ensure that the proposed model achieves superior results with a PSNR of 48.972 dB, SSIM of 0.9953, and entropy value of 8.4319, indicating that it preserves small image details and limits image degradation. Future work will include the integration of DL with focus map generation for enhanced fusion performance. By including focus map generation in the fusion process, the model can focus on choosing important regions from the entire map automatically, thereby resulting in clearer and more accurate images. Because of this method, the fusion of multi-focus satellite images can be used more often, making it easier to analyze activities such as urban planning, environmental monitoring, and disaster management.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

Funding Statement

Not applicable.

Acknowledgments

Author 1 and Author 2 contributed equally to this work.

References

- [1] Yu Sun et al., "Image Fusion for the Novelty Rotating Synthetic Aperture System based on Vision Transformer," *Information Fusion*, vol. 104, pp. 1-14, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] Aisha Javed et al., "Deep Learning-based Detection of Urban Forest Cover Change along with Overall Urban Changes using Very-High-Resolution Satellite Images," *Remote Sensing*, vol. 15, no. 17, pp. 1-18, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [3] Yuanxin Ye et al., "An Adaptive Attention Fusion Mechanism Convolutional Network for Object Detection in Remote Sensing Images," *Remote Sensing*, vol. 14, no. 3, pp. 1-18, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [4] S. Nagarathinam, A. Vasuki, and K. Paramasivam, "PAN and MS Satellite Image Fusion using Fast Guided Filter and PSO Segmentation," *Sādhanā*, vol. 48, no. 4, pp. 1-16, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [5] Yuanxin Ye et al., "A Robust Multimodal Remote Sensing Image Registration Method and System using Steerable Filters with First-and Second-Order Gradients," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 188, pp. 331-350, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Yizhong Pan et al., "Real Image Denoising Via Guided Residual Estimation and Noise Correction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 4, pp. 1994-2000, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [7] Ming Liu et al., "SIFNet: A Self-Attention Interaction Fusion Network for Multisource Satellite Imagery Template Matching," *International Journal of Applied Earth Observation and Geoinformation*, vol. 118, pp. 1-11, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [8] DaiXun Li et al., "FedFusion: Manifold-Driven Federated Learning for Multi-Satellite and Multi-Modality Fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-13, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [9] Mahmoud M. Hammad et al., "Satellite Image Fusion using a Hybrid Traditional and Deep Learning Method," *Geomatics and Environmental Engineering*, vol. 17, no. 5, pp. 145-162, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [10] Hayder Dibs et al., "Fusion Landsat-8 Thermal TIRS and OLI Datasets for Superior Monitoring and Change Detection using Remote Sensing," *Emerging Science Journal*, vol. 7, no. 2, pp. 428-444, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [11] Mojtaba Safari, Ali Fatemi, and Louis Archambault, "MedFusionGAN: Multimodal Medical Image Fusion using an Unsupervised Deep Generative Adversarial Network," *BMC Medical Imaging*, vol. 23, no. 1, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [12] Simon Anger et al., "High-Resolution Inverse Synthetic Aperture Radar Imaging of Satellites in Space," *IET Radar, Sonar and Navigation*, vol. 18, no. 4, pp. 544-563, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [13] Shuangliang Li, Siwei Li, and Lihao Zhang, "Hyperspectral and Panchromatic Images Fusion based on the Dual Conditional Diffusion Models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-15, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [14] Fahime Arabi Aliabad et al., "Identifying and Monitoring Gardens in Urban Areas using Aerial and Satellite Imagery," *Remote Sensing*, vol. 15, no. 16, pp. 1-25, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [15] Penglei Li et al., "MFFSP: Multi-Scale Feature Fusion Scene Parsing Network for Landslides Detection based on High-Resolution Satellite Images," *Engineering Applications of Artificial Intelligence*, vol. 127, pp. 1-12, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [16] Dana Moukheiber et al., "A Multimodal Framework for Extraction and Fusion of Satellite Images and Public Health Data," *Scientific Data*, vol. 11, no. 1, pp. 1-20, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [17] Yukun Zhang et al., "Cross-Task and Cross-Domain SAR Target Recognition: A Meta-Transfer Learning Approach," *Pattern Recognition*, vol. 138, pp. 1-11, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [18] Juan Xiao et al., "Deep Learning-based Spatiotemporal Fusion of Unmanned Aerial Vehicle and Satellite Reflectance Images for Crop Monitoring," *IEEE Access*, vol. 11, pp. 85600-85614, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [19] Chunfeng Yang et al., "Multi-Scale Visual Attention Cascaded Network for Multi-Focus Image Fusion," *2023 4th International Conference on Information Science and Education (ICISE-IE)*, pp. 89-92, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [20] Xinzhe Xie et al., "SwinMFF: Toward High-Fidelity End-to-End Multi-Focus Image Fusion via Swin Transformer-based Network," *The Visual Computer*, vol. 41, no. 6, pp. 3883-3906, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [21] Xingyu Hu et al., "Incrementally Adapting Pretrained Model using Network Prior for Multi-Focus Image Fusion," *IEEE Transactions on Image Processing*, vol. 33, pp. 3950-3963, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [22] SpaceNet AOI 11 - Rotterdam, SpaceNet, 2020. [Online]. Available: <https://spacenet.ai/rotterdam/>
- [23] Samreen Fiza, and S. Safinaz, "Multi-Focus Image Fusion using Edge Discriminative Diffusion Filter for Satellite Images," *Multimedia Tools and Applications*, vol. 83, no. 25, pp. 66087-66106, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [24] Lijing Bu et al., "Enhancing RABASAR for Multi-Temporal SAR Image Despeckling through Directional Filtering and Wavelet Transform," *Sensors*, vol. 23, no. 21, pp. 1-17, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [25] Wenzhong Shi, Dizhou Guo, and Hua Zhang, "A Reliable and Adaptive Spatiotemporal Data Fusion method for Blending Multi-Spatiotemporal-Resolution Satellite Images," *Remote Sensing of Environment*, vol. 268, pp. 1-24, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [26] Achala Shakya, Mantosh Biswas, and Mahesh Pal, "Fusion and Classification of SAR and Optical Data using Multi-Image Color Components with Differential Gradients," *Remote Sensing*, vol. 15, no. 1, pp. 1-15, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [27] Chenxi Duan, Mariana Belgiu, and Alfred Stein, "Feature Enhancement Network for Cloud Removal in Optical Images by Fusing with SAR Images," *International Journal of Remote Sensing*, vol. 45, no. 1, pp. 51-67, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [28] Jiajun Cai, Bo Huang, and Tung Fung, "Progressive Spatiotemporal Image Fusion with Deep Neural networks," *International Journal of Applied Earth Observation and Geoinformation*, vol. 108, pp. 1-11, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [29] Jinjin Li et al., "Comparative Analysis of Pixel-Level Fusion Algorithms and a New High-Resolution Dataset for SAR and Optical Image Fusion," *Remote Sensing*, vol. 15, no. 23, pp. 1-30, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [30] Pengcheng Hu et al., "Remote Sensing Image Reconstruction Method based on Parameter Adaptive Dual-Channel Pulse-Coupled Neural Network to Optimize Multiscale Decomposition," *IEEE Access*, vol. 11, pp. 78084-78103, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [31] S.R. Bandi, M. Anbarasan, and D. Sheela, "Fusion of SAR and Optical Images using Pixel-based CNN," *Neural Network World*, vol. 4, pp. 197-213, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [32] Yingying Kong et al., "A Fusion Method of Optical Image and SAR Image based on Dense-UGAN and Gram-Schmidt Transformation," *Remote Sensing*, vol. 13, no. 21, pp. 1-17, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [33] Jing Fang et al., "A Noisy SAR Image Fusion Method based on NLM and GAN," *Entropy*, vol. 23, no. 4, pp. 1-17, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [34] Runmin Dong et al., "An Adaptive Image Fusion method for Sentinel-2 Images and High-Resolution Images with Long-time Intervals," *International Journal of Applied Earth Observation and Geoinformation*, vol. 121, pp. 1-13, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [35] Abhilash Singh, and Kumar Gaurav, "Deep Learning and Data Fusion to Estimate Surface Soil Moisture from Multi-Sensor Satellite Images," *Scientific Reports*, vol. 13, no. 1, pp. 1-20, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [36] Tayeb Benzenati, Yousri Kessentini, and Abdelaziz Kallel, "Spectral-Temporal Fusion of Satellite Images via an End-to-End Two-Stream Attention with an Effective Reconstruction Network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 1308-1320, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]