International Journal of Mathematical, Engineering and Management Sciences
Vol. 11, No. 3, 1227-1264, 2026
https://doi.org/10.33889/IJMEMS.2026.11.3.051

MF-SRGAN: A Super-Resolution Generative Adversarial Network for Multi-Focus Image Fusion

Shatabdi Basu
Department of Computer Science and Engineering,
Manipal University Jaipur, Jaipur, Rajasthan, India.
E-mail: shatabdi.basu@jaipur.manipal.edu

Sunita Singhal
Department of Computer Science and Engineering,
Manipal University Jaipur, Jaipur, Rajasthan, India.
Corresponding author: sunita.singhal@jaipur.manipal.edu

Dilbag Singh
Center of Biomedical Imaging, Department of Radiology,
New York University Grossman School of Medicine, New York, United States.
E-mail: dilbag@ieee.org

(Received on April 13, 2025; Revised on August 17, 2025 & February 23, 2026; Accepted on April 22, 2026)

Abstract
In computational imaging, multi-focus image fusion is a critical process that aims to produce a single image that covers all-in-focus areas from numerous partially focused input images. In this paper, we present a novel approach using a Super-Resolution Generative Adversarial Network (SRGAN) specifically designed for multi-focus image fusion. First, we create a new multi-focus image dataset from the publicly accessible COCO dataset. This process generates a complete collection of annotated image pairs with different focus areas. The generator is designed using a residual learning architecture and upsampling layers. The generator creates a high-resolution fused image with features and texture preservation by processing two input images. Using PatchGAN-based implementation, the discriminator ensures that the fused images maintain global and local coherence through adversarial training. Putting emphasis on intensity, structural similarity, and perceptual qualities, we combine content loss with adversarial loss to achieve balanced learning. Extensive trials on public multi-focus image datasets show that our SRGAN-based model achieves superior fusion quality and texture consistency by outperforming five current state-of-the-art approaches in both quantitative and visual evaluations. The proposed method achieves real-time performance, meets the requirements of contemporary image fusion applications, and demonstrates its efficacy in generating high-quality fused images.

Keywords- Multi-focus image fusion, Deep learning, Generative adversarial network, Super resolution, Perceptual image quality, Computer aided manufacturing.

1. Introduction

Multi-Focus Image Fusion (MFIF) is the process by which several images, each of which is focused on a distinct region of a given scene, are combined to produce a single, totally in-focus image (Luo et al., 2025). In imaging, depth-of-field constraints frequently hinder the simultaneous capture of all objects at different depths within a scene in sharp focus. Scenarios such as microscopic imaging (Li et al., 2024b; Song et al., 2006), medical imaging (Ghandour et al., 2024; Jie et al., 2024), remote sensing (Nagarathinam et al., 2024; Zhang, 2022), and visual power patrol inspection (Singh et al., 2023; Wu et al., 2023) make this apparent. MFIF addresses this issue by combining many partially focused images into a single, complete image. This process improves the depth of field and ensures clarity throughout all regions of interest (Hu et al., 2025).

However, the fusion of multi-focus images presents some challenges arising from various factors (Duan et al., 2024). It is difficult to accurately identify focused regions in MFIF, especially when the transitions between the focused and blurred areas are subtle. Fusion performance may deteriorate further due to artifacts, loss of contrast, and sensitivity to variations in noise and lighting. Various approaches for MFIF have been presented in the literature by Wang et al. (2024b). These include traditional spatial and transform-domain methods, as well as advanced computational approaches such as optimization algorithms, deep learning frameworks, and hybrid strategies.

Spatial domain techniques in MFIF directly process pixel intensities to detect and merge focused areas. Pixel-level methods analyze local intensity or activity measures to identify focused pixels (Bouzos et al., 2019; Liu et al., 2021; Ma et al., 2019; Zhang et al. 2021b). Block-based methods (Tian et al., 2018; Wu et al., 2019) and region-based techniques (Chen et al., 2020; Qiu et al., 2019) attempt to incorporate spatial context, but suffer from loss of fine details, block artifacts, and difficulties in managing regions with gradual focus transitions.

Transform-domain fusion methods perform MFIF in alternative representations, such as the wavelet domain (Manchanda and Gambhir, 2019), Fourier domain (Li and Jiang, 2020), shearlet domain (Wei et al., 2018), sparse representation frameworks (Zhang et al., 2020a), and gradient-based domains (Li et al., 2021). These approaches can capture multi-scale image features but rely on predefined transforms and handcrafted fusion rules. This limits adaptability and leads to suboptimal performance across diverse image structures. These constraints have led to the development of data-driven deep learning-based MFIF methods.

In recent years, deep learning algorithms (Ma et al., 2022; Qu et al., 2022) have shown that they can automatically learn complex focus-related features. Liu et al. (2017) were the first to deploy a deep convolutional neural network for MFIF by generating decision maps that differentiate between focused and defocused areas. Bhalla et al. (2022) proposed a Fuzzy CNN that combines fuzzy logic with convolutional learning to enhance focus detection and minimize fusion artifacts. Ma et al. (2021) proposed an unsupervised encoder-decoder system that identifies discriminative features and utilizes spatial frequency metrics for fusion, eliminating the need for ground-truth images. Shao et al. (2024) introduced STCU-Net, a U-Net-based architecture that combines Transformer modules and CNNs with depth-information learning to capture local structural details and global contextual dependencies simultaneously. Li et al. (2024a) presented a denoising diffusion probabilistic model that systematically improved fused outputs through iterative noise elimination.

Deep learning-based MFIF approaches have been successful, but they still have certain problems. They need large, labeled datasets and have trouble preserving spatial consistency at focus boundaries. These constraints make them less reliable in real-world scenarios with complex focus patterns. Generative Adversarial Networks (GANs) have emerged to overcome these limitations by utilizing adversarial training to enhance the fusion process. These networks allow the model to learn robust, data-driven fusion rules that preserve fine details, enhance spatial consistency, and attain superior fusion quality despite limited or unlabeled training data.

This paper proposes a GAN-based framework that combines super-resolution and adversarial learning to produce fused images with enhanced structural and perceptual details. Designed to strike a balance between sharpness, contrast, and information preservation, the model is suitable for complex fusion tasks requiring diverse focus areas.

1.1 Research Questions

We formulated a set of research questions to investigate the effectiveness of our proposed method. These questions are designed to evaluate the architectural contributions and general performance of the model in visual and quantitative aspects. **Table 1** shows the research questions and their underlying motivation.

Table 1. Key research questions with corresponding motivations.

Research questions	Motivation
RQ1: How effectively can a GAN model based on super-resolution preserve details in multi-focus image fusion compared to existing methods?	The existing CNN and transformer based fusion methods often encounter trade-offs between perceptual clarity and the preservation of structural details. SRGAN's offer a promising alternative in this area although its role in multi-focus image fusion is underexplored.
RQ2: How much do architectural advancements in the SRGAN framework improve visual and quantitative fusion quality?	The original purpose of the baseline SRGAN architecture was to generate images with super-resolution. Adapting the framework for image fusion through refined convolutional layers and dual inputs may show improvements. However, the individual contributions of architectural changes remain unclear.
RQ3: Are the improvements achieved by the proposed method statistically significant and consistent across image fusion metrics?	Visual improvements are inadequate without strong statistical validation. Fusion models often exhibit inconsistent performance across datasets due to variations in focus distribution. Hence, there is need to confirm that the observed improvement of proposed method is robust and statistically significant.

1.2 Contributions and Layout

This paper presents a GAN-based framework for multi-focus image fusion, designed to overcome the limitations of existing fusion approaches in maintaining both high-frequency textures and global consistency. The chief contributions of this paper are as follows:

- A Multi-Focus Super-Resolution Generative Adversarial Network (MF-SRGAN) is proposed by redefining multi-focus image fusion as a super-resolution-guided adversarial reconstruction challenge. Unlike prior GAN-based fusion methods that depend on decision maps or direct pixel-level blending, the proposed method performs end-to-end dual-input feature extraction followed by residual refinement and super resolution reconstruction.
- The framework uses residual learning and pixel-shuffle upsampling to improve the preservation of high-frequency boundaries. A pre-trained VGG19-based perceptual loss is utilized to ensure structural and textural consistency in feature space instead of traditional pixel-wise supervision.
- A new multi-focus image fusion dataset has been created using a systematic approach derived from the COCO dataset. The proposed approach for dataset generation uses binary mask segmentation and controlled Gaussian defocus simulation to generate realistic pairs of partially focused images with smooth boundary transitions. This technique offers reliable supervision for adversarial learning and improves the generalization capacity of MF-SRGAN across various focus patterns.

The remainder of this paper is arranged as follows. Section 2 offers an overview of GAN-based image fusion methods in the existing literature and explores the functionality of super-resolution GANs. Section 3 presents the customized super-resolution GAN model developed specifically for multi-focus image fusion. The details of the newly developed dataset, training details, comparative evaluations with state-of-the-art image fusion methods using visual and quantitative metrics, results of ablation studies, sensitivity analysis, significance testing and computational efficiency are outlined in Section 4. Section 5 summarizes the findings, provides responses to the research questions, highlights limitations, and explores future directions. Section 6 provides the conclusion of this paper.

2. Related Works

2.1 GAN-Based Image Fusion Methods

Although conventional fusion methods successfully produce reasonably good-quality fused images, they display noise and artifacts. These methods have low computational efficiency along with difficulty in generalizing the model. In recent years, deep learning techniques have come to the forefront to achieve end-to-end image fusion through a single model that integrates feature extraction, transformation, and fusion steps. Their ability to capture both low-level features such as edges and textures and high-level features such as structures is crucial in the field of computer vision (Chai et al., 2021).

GAN was first proposed by Goodfellow et al. (2014) for the purpose of image generation. The basic methodology involves two neural networks: a generator that creates realistic samples from random noise and a discriminator that can differentiate between real and generated samples. Simultaneously, the two networks are trained in a maximum minimum game in which the generator tries to fool the discriminator while the discriminator seeks to improve its classification accuracy. This adversarial process, by means of backpropagation, enables the generator to approximate the true data distribution without the explicit requirement of likelihood estimation.

Several GAN-based methods (Guo et al., 2019; Huang et al., 2020; Li et al., 2023a; Li et al., 2023b; Wang et al., 2021; Zhang et al., 2021a) have recently been proposed to use adversarial learning and deep feature extraction to improve multi-focus image fusion. These techniques seek to minimize artifacts by improving focus region recognition, structural consistency, and boundary region sharpness. **Table 2** presents a comparative analysis of methods that use GAN variants and their features and limitations.

Table 2. Comparative analysis of existing GAN-based methods.

Reference	Method	Features	Open challenges
Guo et al. (2019)	Conditional Generative Adversarial Network (cGAN)	1. Use of a Siamese network in the generator to generate a confidence map. 2. Implementation of Convolutional Conditional Random Fields (ConvCRFs) to refine the confidence map.	1. Training in a limited dataset. 2. Heavy reliance on confidence maps where any inaccuracy can lead to suboptimal fusion.
Huang et al. (2020)	Adaptive Constraints GAN (ACGAN)	1. Used adaptive weight blocks to ensure fusion without decision maps. 2. Use of perceptual loss to preserve object boundaries.	1. Lack of structural consistency in complex scenes. 2. Suboptimal performance for images with minimal contrast variations.
Zhang et al. (2021a)	Unsupervised GAN	1. Used a repeated blur technique to find focus areas at the pixel level. 2. Implemented a joint gradient constraint to improve texture details.	1. Noise in low-light regions may get amplified due to network's reliance on gradient constraints. 2. Issues with intricate focus transitions in images with highly varying textures.
Wang et al. (2021)	Squeeze-and-Excitation (SE) Residual Network	1. Combines gradient penalty and reconstruction loss to enhance edge details. 2. Mitigated the defocus spread effect by generating focus maps for foreground regions.	1. Increase in computational complexity due to the inclusion of multiple parallel branches. 2. Needs generalization to real-world datasets with variations in lighting and textures.
Li et al. (2023a)	Siamese Conditional Generative Adversarial Network (SCGAN)	1. Used Wasserstein Divergence (DIV) Optimization to stabilize the discriminator. 2. Designed structured sparse loss function to improve edge clarity.	1. Complex training process due to joint distribution learning and Siamese structure. 2. Challenges in achieving consistent performance over diverse datasets.
Li et al. (2023b)	Gradient and Intensity Joint Proportional Constraint GAN (GIPC-GAN)	1. Employed a novel constraint for the simultaneous preservation of structural and brightness features of the source images. 2. Use of two encoding paths for better feature extraction.	Increase in computational costs due to dense connections and multilayer structure.

2.2 Super-Resolution Generative Adversarial Networks

Traditional GANs in MFIF frequently encounter issues such as unstable training resulting from adversarial loss, limitations in accurately capturing fine details of focused areas, and a likelihood of producing artifacts in the fused image. They may also lack the ability to adequately preserve the structural integrity of the original input images, resulting in unsatisfactory fusion outcomes. The SRGAN framework proposed by Ledig et al. (2017) addressed the issue by focusing on the generation of high-resolution images with photorealistic features by employing perceptual loss functions to capture texture and fine details.

SRGAN is particularly well suited for MFIF because it can improve the high-frequency content at the fusion boundaries. This ensures sharper transitions between focused regions. Transformer-based fusion methods depend on global attention mechanisms and often require extensive pre-training. SRGAN is better than transformer-based fusion methods because it provides a targeted framework that preserves fine-grained textures, local focus regions, and structural boundaries. These qualities are particularly important in multi-focus fusion tasks. The framework enhances multi-focus image fusion by preserving sharpness and structural integrity while reducing artifacts. A novel perceptual loss function integrates content loss, calculated from the feature maps of the VGG network, and adversarial loss. The combination of these two losses ensures visually combining and structurally precise outputs. The objective function of SRGAN is represented as a min-max optimization problem (Goodfellow et al., 2014) as shown below:

$$\min_{\theta_G} \max_{\theta_D} E_{I^{hr}\sim P_{train}(I^{hr})}\left[\log D_{\theta_D}\left(I^{hr}\right)\right] + E_{I^{lr}\sim P_G(I^{lr})}\left[\log\left(1 - D_{\theta_D}\left(G_{\theta_G}\left(I^{lr}\right)\right)\right)\right] \quad (1)$$

Here, $\min_{\theta_G} \max_{\theta_D}$ represents the adversarial training objective, θ_G represents the parameters of the generator G that seek to minimize the objective of fooling the discriminator, and θ_D represents the parameters of the discriminator D, which aims to maximize the objective of differentiating real images from fake ones.

The term $E_{I^{hr}\sim P_{train}(I^{hr})}$ depicts the expectation over high-resolution images I^{hr} drawn from the real data distribution $P_{train}\left(I^{hr}\right)$. The discriminator D aims to maximize the log probability of accurately recognizing high-resolution images as real ($D_{\theta D}(I^{hr}) \approx 1$).

The term $E_{I^{lr}\sim P_G(I^{lr})}$ depicts the expectation for low-resolution images (I^{lr}) processed through the generator G to produce super-resolved images ($G_{\theta_G}\left(I^{lr}\right)$). The discriminator D seeks to maximize the log-probability of classifying these generated images as fake $\left(1 - D_{\theta_D}\left(G_{\theta_G}\left(I^{lr}\right)\right)\right)$.

The generator's optimization step is shown in Equation (2). The objective is to train generator G to generate high-resolution images $G_{\theta_G}\left(I_m^{lr}\right)$). that accurately correspond to the respective high-resolution images of the ground truth I_m^{hr}. Training minimizes the super-resolution loss function $\mathcal{L}^{sr}$ in all training samples to modify the generator's parameters θ_G.

$$\widehat{\theta_G} = \arg\min_{\theta_G} \frac{1}{M}\sum_{m=1}^{M} \mathcal{L}^{sr}\left(G_{\theta_G}\left(I_m^{lr}\right), I_m^{hr}\right) \quad (2)$$

Here, $\arg\min_{\theta_G}$ represents the process of identifying the generator's parameters that optimize the specified objective function. M is the number of training samples in the current batch. $\mathcal{L}^{sr}$ is super-resolution loss function consisting of perceptual loss, which combines content loss and adversarial loss. This loss function is crucial to producing high-quality, visually appealing super-resolution images.

3. Proposed Method

MF-SRGAN is an improved version of super-resolution GAN (Ledig et al., 2017) and consists of a generator and a discriminator. Unlike existing GAN-based fusion methods, MFSRGAN focuses on adversarial training to preserve both global and local textures. In addition, MF-SRGAN upscales fused images to enhance clarity and sharpness, while the other techniques operate at standard resolutions. The subsequent sections provide details of the architecture of the generator and discriminator, and the loss functions used.

3.1 Architecture of Generator

The generator architecture is shown in **Figure 1**. A pair of multifocus images are passed as input to convolution layer blocks Conv1 and Conv2 with a kernel size of 9 × 9 to preserve spatial information. The parametric ReLU function is applied after each convolution for non-linearity.

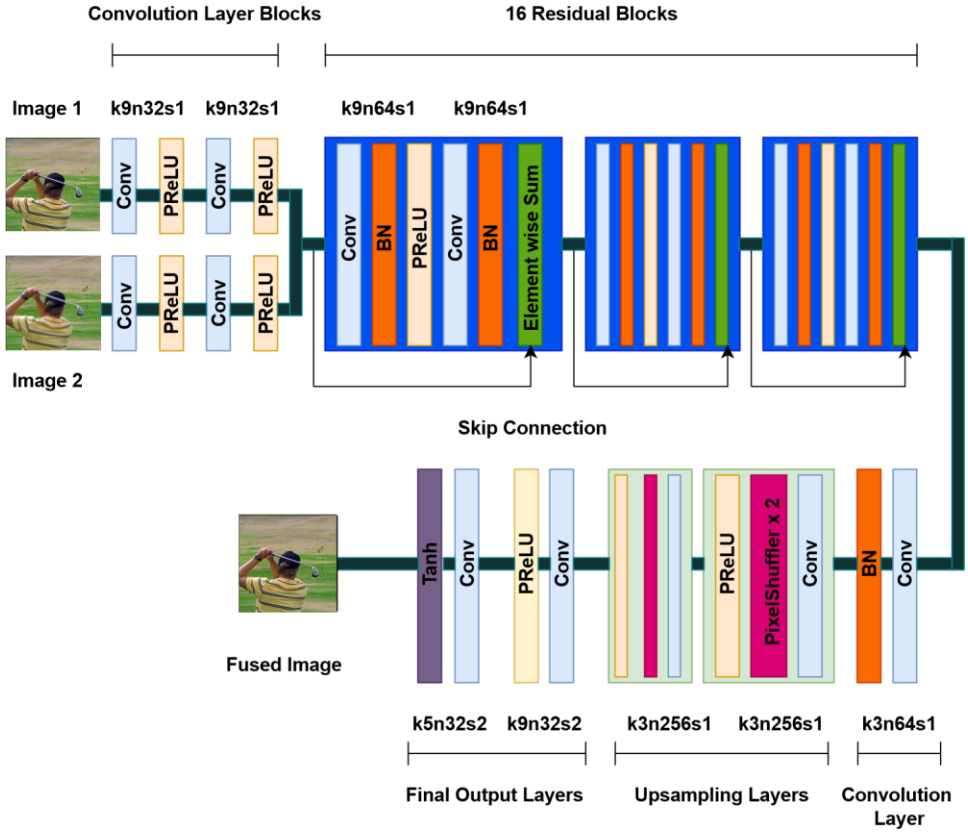


Figure 1. Architecture of MF-SRGAN generator network.

The core of the generator network consists of 16 residual blocks (Block 1-16) with an identical layout first proposed by Gross and Wilber (2016). Each residual block has two convolutional layers with 3 × 3 kernels and 64 feature maps followed by batch normalization and PReLU activation functions. The residual blocks help the model learn complex residual mappings while preventing the vanishing gradient problem. The features extracted from previous layers are further refined using a second convolutional layer with a kernel size of 3 × 3 and batch normalization.

Two blocks of upsampling layers (blocks 17 and 18) are used to increase the spatial resolution of the feature maps. The architecture of this layer has convolutional operations with stride 3 × 3 to introduce learnable parameters, batch normalization for stabilization of training, pixel shuffle for upsampling by a factor of 2 in order to produce a high-resolution image, and a PReLU activation function to introduce nonlinearity.

These upsampling layers, which are based on super-resolution methods, improve the spatial resolution of the fused output. They also play an important role in the retention of fine-grained details from source images. In multi-focus fusion, high-frequency information at focus boundaries is often lost or blurry. The use of pixel-shuffle-based upsampling enhances texture sharpness and elevates spatial clarity in the fused image. The super-resolution component improves fusion quality by allowing feature-level refinement before the final image is produced.

The final output layer produces the fused image by transforming the feature maps produced by the preceding layers. The architecture has two convolutional layers, Conv4 and Conv5, to map the features into the output image space. Conv4 has a kernel of size 9 × 9 to extract more abstract features along with a PReLU activation function. Also, Conv4 reduces the size of the feature maps by down sampling the spatial dimensions. Conv5 performs another set of convolutional operations to transform the feature maps into 3 output channels. The tanh activation function is also applied to ensure that the pixel values of the output image are within a valid range. **Table 3** shows the configuration of each convolutional layer used in the generator.

Table 3. Detailed summary of the generator.

Generator	Layer	Input channels	Output channels	Kernel size	Stride	Activation function
Input convolution layer	Conv1	3	32	9 × 9	1	PReLU
	Conv2	3	32	9 × 9	1	PReLU
Residual blocks	Block 1-16	64	64	9 × 9	1	PReLU
Second convolution layer	Conv3	64	64	3 × 3	1	-
Upsampling layer	Block 17-18	256	64	3 × 3	1	PReLU
Output convolution layer	Conv4	32	32	9 × 9	2	PReLU
	Conv5	32	3	5 × 5	2	Tanh

3.2 Architecture of Discriminator

Figure 2 shows the discriminator architecture where several discriminator blocks precede a final convolutional layer.

The initial convolutional block, Conv1, has a convolutional layer with a kernel size of 3 × 3 and a LeakyReLU activation function. This layer preserves the spatial dimensions of the input image. The subsequent seven blocks (Conv2 - Conv8) all consist of a convolutional layer, a batch normalization layer, and a LeakyReLU activation function layer. These sequences of discriminator blocks have filters of increasing size (64, 128, 256, 512). Each block increases the number of feature maps while reducing the spatial dimensions by 2. The final layer, Conv9, is a 3 × 3 convolutional layers with a single output channel that reduces the feature map to a single-channel output to perform the final classification, indicating a real or fake image. **Table 4** presents the convolutional layer settings used in the discriminator.

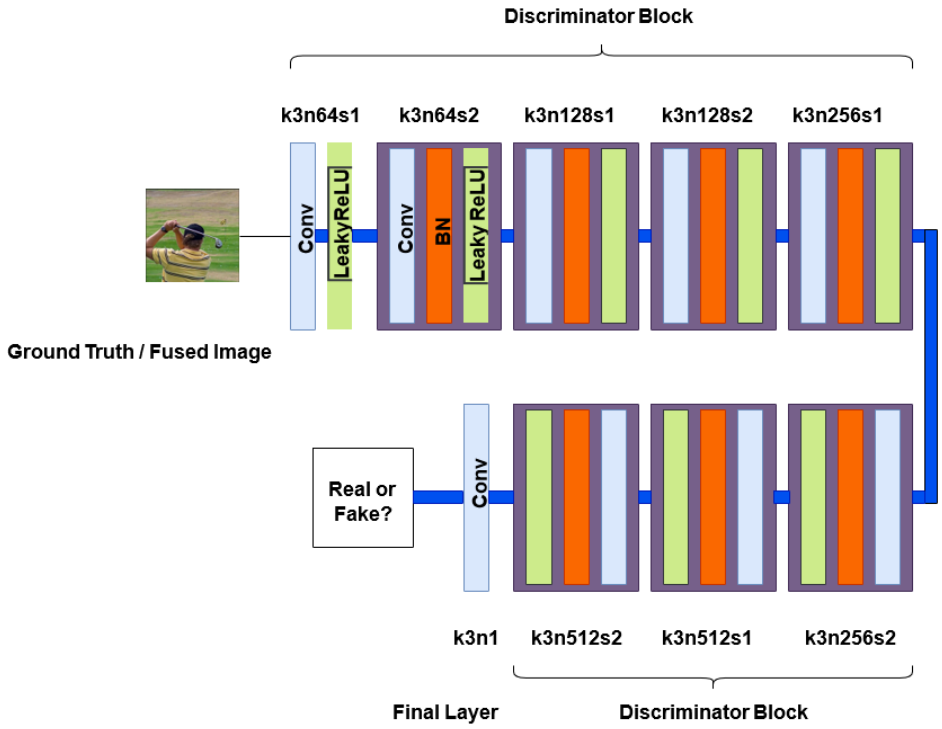


Figure 2. Architecture of discriminator network.

Table 4. Detailed summary of the discriminator.

Layer	Input channels	Output channels	Kernel size	Stride	Activation function
Conv1	1	64	3 × 3	1	LeakyReLU
Conv2	64	64	3 × 3	2	LeakyReLU
Conv3	64	128	3 × 3	1	LeakyReLU
Conv4	128	128	3 × 3	2	LeakyReLU
Conv5	128	256	3 × 3	1	LeakyReLU
Conv6	256	256	3 × 3	2	LeakyReLU
Conv7	256	512	3 × 3	1	LeakyReLU
Conv8	512	512	3 × 3	1	LeakyReLU
Conv9	512	1	3 × 3	1	-

3.3 Loss Functions

Loss functions in GANs are crucial for balancing the adversarial game between the generator and the discriminator. These functions guide the generator to provide realistic outputs while allowing the discriminator to differentiate between real and generated data. The MF-SRGAN network uses generator loss and discriminator loss.

3.3.1 Loss Function of Generator

The generator loss is intended to train the generator to create high-quality and perceptually precise high-resolution images and is a combination of content loss ($\mathcal{L}_{content}$) and adversarial loss ($\mathcal{L}_{adv}$). The mathematical equation is shown as follows:

$$\mathcal{L}_G = \mathcal{L}_{content} + \lambda \mathcal{L}_{adv} \quad (3)$$

Here, $\lambda = 10^{-3}$ is a scaling factor to balance the influence of adversarial loss.

The content loss guides the generator to prioritize high-level structural fidelity and ensures that the fused image preserves relevant details (e.g., edges, textures) from both the input images. We compute the content loss as the L1 norm of the difference between the feature representations of the generated fused image and the fully focused ground truth image as extracted by pretrained VGG19 network which is shown in Equation (4). The feature extractor uses the first 18 layers of VGG19 to capture hierarchical texture and structure information. These layers are known to be good at maintaining perceptual similarity.

$$\mathcal{L}_{content} = \frac{1}{N}\sum_{i=1}^{N} || \Phi\left(f(I_a, I_b)\right)_i - \Phi\left(I_{gt}\right)_i || \quad (4)$$

Here, $\| \cdot \|$ represents the L_1 norm or the mean absolute error. $f(I_a,I_b)$ is the fused image generated by the model. I_{gt} is the ground truth image. $\Phi(f(I_a,I_b)$ represents the feature map of the fused image, while $\Phi(I_{gt})$ is the feature map of the ground truth image. The function $\Phi(.)$ is the output of the first 18 layers of a VGG19 network pretrained on ImageNet. It turns images into a high-dimensional feature space that captures perceptual content.

The adversarial loss ensures that the generator produces fused images that the discriminator is unable to differentiate from real, fully focused images. This is accomplished by training the generator to maximize the discriminator's probability of categorizing the fused image as a real, fully focused image. The generator's goal is to minimize the adversarial loss, which is defined as follows:

$$\mathcal{L}_{adv} = -E_{I_a,I_b \sim P_{data}}\left[\log D\left(f(I_a, I_b)\right)\right] \quad (5)$$

Here, $E_{I_{a,Ib} \sim P_{data}}$ represents the expectation for all pairs of input images I_a, I_b sampled from the true data distribution P_{data}. The negative sign signifies that this loss is minimized throughout training to encourage the generator's production of more realistic outputs. $f(I_a,I_b)$ is the fused image created by the generator. The discriminator D assigns a probability score to $f(I_a,I_b)$, determining if it is real (closer to 1) or fake (closer to 0). The term $\log D(f(I_a,I_b))$ represents the logarithm of the discriminator's score, highlighting the confidence of classification.

3.3.2 Loss Function of Discriminator

The discriminator loss for the fused image ensures that the discriminator acquires the ability to differentiate between real, fully focused images and the generated fused images. It is a combination of two components: loss for real images and loss for generated fake images.

For real images, the discriminator should output a value close to 1, signifying accurate identification of the image as real. The loss for real images is computed using Mean Squared Error (MSE) as shown below:

$$\mathcal{L}_{real} = \frac{1}{N}\sum_{i=1}^{N}\left(D\left(I_{gt}^{(i)}\right) - 1\right)^2 \quad (6)$$

Here, $D\left(I_{gt}^{(i)}\right)$ represents the output of the discriminator for the $i-th$ real image, and N is the batch size.

For generated (fused) images, the discriminator should output a value close to 0, signifying accurate identification of the image as fake. The loss for generated or fake images is shown below:

$$\mathcal{L}_{fake} = \frac{1}{N}\sum_{i=1}^{N}\left(D\left(f\left(I_a^{(i)}, I_b^{(i)}\right)\right) - 0\right)^2 \quad (7)$$

Here, $D\left(f\left(I_a^{(i)}, I_b^{(i)}\right)\right)$ represents the output of the discriminator for the *i-th* fake image. *N* is the batch size, i.e., the number of images processed in a single forward/backward pass during model training.

The total discriminator loss is computed as the average of the losses for real and fake images, as shown in the following equations:

$$\mathcal{L}_{\mathcal{D}} = \frac{1}{2}\left(\mathcal{L}_{real} + \mathcal{L}_{fake}\right) \tag{8}$$

$$\mathcal{L}_{\mathcal{D}} = \frac{1}{2N}\sum_{i=1}^{N}\left[\left(D\left(I_{gt}^{(i)}\right) - 1\right)^2 + \left(D\left(f\left(I_a^{(i)}, I_b^{(i)}\right)\right) - 0\right)^2\right] \tag{9}$$

Here, 1/2 represents a scaling factor to ensure that the inputs of the real and fake terms are equal. The term $\frac{1}{N}\sum_{i=1}^{N}$ represents the average of a batch of size *N*. The term $D\left(I_{gt}^{(i)}\right)$ is the output of the discriminator for the *i−th* real image. The term $D\left(f\left(I_a^{(i)}, I_b^{(i)}\right)\right)$ represents the output of the discriminator for the *i−th* fake image generated by *f*.

4. Experiments and Results

This section provides details of the datasets utilized, including a description of the test dataset and the methodology followed to generate the training dataset. We also specify the training configuration of the proposed model, including hyperparameters, optimization strategies, and implementation details. The environment configuration used for the experiments is also discussed. We present both visual and quantitative evaluation results, comparing the performance of our method with that of a few baseline multi-focus MFIF models. The results of sensitivity analysis and significance testing are also provided. Finally, we present the results of the ablation experiments performed on the proposed model.

4.1 Datasets

We used four datasets for our experiments. The Lytro dataset (Nejati et al., 2015), MFIWHU dataset (Zhang et al., 2021a) and a real-world dataset were used for testing purposes. Another multi-focus dataset was constructed for training purposes by segmenting and processing images from the publicly available COCO dataset. This dataset was selected for its extensive representation of indoor (e.g. kitchens, drawing rooms) and outdoor scenes (e.g. parks, urban areas). The dataset has objects from several categories such as humans, animals, vehicles, and small portable items. This ensures diversity in terms of spatial structure, object scale, and texture.

Each image of the COCO dataset was split into two separate focus areas based on binary masks. Both binary masks were dilated utilizing a structuring element (elliptical kernel of dimensions 21×21) to ensure a smoother transition at the boundaries of the mask and prevent artifacts resulting from dark pixels. The image was split into two regions using binary masks, and random blur was applied using odd-numbered kernel size to introduce variation in blur. For each segmented region, the blurred area was merged with the original image via binary masks to produce two partially focused images. **Figure 3** provides a side-by-side visual comparison of real and synthetically generated blur. The first image is an all-focus scene, while the second image depicts real-world blur obtained under practical imaging conditions. The subsequent images depict the synthetically generated near-focus and far-focus counterparts created using complementary binary masks and Gaussian blur.

Figure 3. Side-by-side comparison of real and synthetically generated defocus blur.

The generated multi-focus image pair is represented by the equations below:

$$I_a(m,n) = M_a(m,n) \cdot I(m,n) + M_b(m,n) \cdot \big(G(m,n;K) \times I(m,n)\big) \quad (10)$$

$$I_b(m,n) = M_b(m,n) \cdot I(m,n) + M_a(m,n) \cdot \big(G(m,n;K) \times I(m,n)\big) \quad (11)$$

Here, $I_a(m, n)$ is the partially focused image A and $I_b(m, n)$ is the partially focused image B. $I(m, n)$ is the original input image from the COCO dataset that serves as the source to produce partially focused images. $M_a(m, n)$ represents the binary mask that defines the region of the image that needs to retain sharpness, while $M_b(m, n)$ is the binary mask that defines the region of the image that needs to be blurred. The term $G(m, n; K) \times I(m, n)$ represents the convolution of the Gaussian blur kernel G of size $K \times K$ with the original image I, which simulates a defocus effect for the image.

The fully focused ground truth image is represented by equation below:

$$I_{RES}(m,n) = I(m,n) \quad (12)$$

Here, $I(m, n)$ is the original input image from the COCO dataset.

Three corresponding images were generated for each image in the COCO dataset. These were Image A (focused on Mask A), Image B (focused on Mask B), and Image Res (original image as reference image). This process provided 30,000 images for the training process from the 10,000 images initially selected from the COCO dataset. The dataset generation pipeline is available at https://github.com/Shatabdi22/MF-SRGAN-DatasetGeneration

The Lytro dataset contains 20 image pairs of categories such as Golf, Scuba, Gym, Lock, Model, etc. The MFI-WHU dataset comprises 120 image pairs specially designed for research in multi-focus image fusion. Each pair consists of images with distinct focal points, generated by employing Gaussian blurring to simulate diverse focus settings. Both Lytro and MFI-WHU were used as test datasets without any augmentation. The real-world data set consisted of 5 pairs of indoor and 5 pairs of outdoor images. These images were acquired under controlled conditions using a fixed camera arrangement to ensure spatial alignment while altering the focal depth.

4.2 Training Details

Algorithm 1 gives a brief overview of the training process.

Algorithm 1 The Training Process

Network Initialization:
Set loss functions (Content Loss and Adversarial Loss);
for *100 epoch* **do**
    for *each iteration* in the training dataset **do**
        Step 1: Train the Generator
            Select input multi-focus image pairs (I_a,I_b);
            Generate the fused image I_f;
                Compute the Content Loss between I_f and the ground truth image I_{gt};
            Compute the Adversarial Loss using the Discriminator's output;
        Compute the total Generator Loss;
        Update the Generator G;
        Step 2: Train the Discriminator
            Compute the Discriminator Loss for the real image I_{gt};
            Compute the Discriminator Loss for the generated image I_f;
        Compute the total Discriminator Loss;
        Update the Discriminator D;
    end for
    Save model checkpoints periodically;
end for

The training time epochs are set to n epochs =100. The batch size is configured to 4 at a time during each iteration of the training. The Adam optimizer is used to calculate the adaptive learning rates, which helps to achieve faster convergence by adjusting the learning rate of each parameter. The exponential decay rate for the first moment estimate β_1 is set to 0.5, while the exponential decay rate for the second moment estimate, β_2, is set to 0.999. The learning rate of the generator and discriminator is set to 0.0002. The entire training process took about 16 hours.

The hyperparameters of the loss functions in our model for multi-focus image fusion are determined based on existing literature (Ledig et al., 2017) and experimental validation. The generator loss function integrates content loss and adversarial loss, weighted by a parameter $\lambda = 10^{-3}$ that maintains the balance between perceptual similarity and realism in the fused images produced. To compute content loss, we apply L1 loss to the feature maps of the generated and ground-truth images, which are obtained by a pre-trained feature extractor (VGG). The adversarial loss, computed as the MSE loss, evaluates the discriminator's ability to categorize the generated images as real. In this implementation, real and fake labels are chosen as $y_{\text{real}} = 1$ and $y_{\text{fake}} = 0$, respectively.

4.3 Experimental Comparisons

4.3.1 Fusion Performance Metrics

Assessing the quality of fused images in multi-focus image fusion is an essential and complex task that necessitates a combination of visual and quantitative evaluation techniques. The qualitative/visual evaluation emphasizes human perception, examining the visual appeal and efficacy of the fusion in relation to the human visual system. It ensures that the fused image maintains perceptual clarity, with both focused areas and transitions appearing natural to the human eye.

Quantitative evaluation is crucial for delivering an objective assessment of fusion performance. The process entails the analysis of the fused image through mathematical and statistical methods to verify its effective integration of information from all source images and its accurate representation of the scene. The combination of both approaches ensures a comprehensive assessment, in which visual metrics evaluate perceptual significance, and quantitative metrics offer statistical confirmation of fusion accuracy and information retention.

To ensure a comprehensive evaluation, we utilize four categories of metrics typically employed in multi-focus image fusion (Liu et al., 2012): information theory-based metrics, image feature-based metrics, image structural similarity metrics, and human perception inspired metrics.

In this paper, we select six performance metrics such as Entropy, Mutual Information (information theory-based metrics), Standard Deviation and Spatial Frequency (image feature-based metrics), Correlation Coefficient (image structural similarity metric), and Visual Information Fidelity (human perception-inspired metric). The description and calculation formula for the metrics is provided in **Table 5**.

Table 5. Computation of performance metrics.

Metric	Description	Formula	Best value
Entropy Li et al. (2023b)	Measures the amount of information in a fused image	$EN = \sum_{i=0}^{255} P_i \log\left(\frac{1}{P_i}\right)$ where, P_i shows the normalized frequency of intensity level i in an 8-bit grayscale image.	Higher
Mutual Information Feng et al. (2022)	Measures the consistency of content between source and fused images	$MI = MI^{BF} + MI^{HF}$ where, $MI^{BF} = \sum_{f=0}^{L}\sum_{b=0}^{L} p^{BF}(b,f)\log_2\frac{p^{BF}(b,f)}{p^{B}(b)p^{F}(f)}$, $MI^{HF} = \sum_{f=0}^{L}\sum_{h=0}^{L} p^{HF}(h,f)\log_2\frac{p^{HF}(h,f)}{p^{H}(h)p^{F}(f)}$ Here, MI^{BF} and MI^{HF} quantify the mutual information between the source images and the fused image. $p^{B}(b), p^{H}(h)$ and $p^{F}(f)$ represent the probability distribution of the source and fused images. $p^{BF}(b,f)$ and $p^{HF}(h,f)$ depict the co-occurrence histograms between the source images and fused image, respectively.	Higher
Standard Deviation Liu et al. (2024)	Measures the dispersion of pixel intensity values relative to the mean	$SD = \sqrt{\frac{1}{S\times R}\sum_{i=1}^{S}\sum_{m=1}^{R}(F_{im}-\bar{F})^2}$ where, $\bar{F} = \frac{1}{S\times R}\sum_{i=1}^{S}\sum_{m=1}^{R}\lvert F_{im}\rvert$ Here, S is height of the image and R is width of the image. F_{im} is the value of intensity at row i and column m. $\bar{F}$ is the mean intensity calculated as the average of all pixel values in an image.	Higher

Table 5 continued…

Spatial Frequency Li et al. (2023c)	Measures the variations in pixel intensity	$SF = \frac{1}{M \times Q} \sum_{m=1}^{M} \sum_{q=1}^{Q} G(m,q)$ where, $G(m,q) = \sqrt{\left(F_x(m,q)\right)^2 + \left(F_y(m,q)\right)^2}$ The term $F_x(m,q) = F(m,q) - F(m,q-1)$ shows the horizontal gradient approximations using finite differences. The term $F_y(m,q) = F(m,q) - F(m-1,q)$ shows the vertical gradient approximations using finite differences. M and Q are the dimensions of the image.	Higher
Correlation Coefficient Wang et al. (2024a)	Measures the linear relationship between source and fused image	$CC = \sqrt{\frac{\sum_{j=1}^{M} \sum_{k=1}^{S} \left(F_t(j,k) - F_f(j,k)\right)^2}{M \times S}}$ where, M and S shows number of columns and rows in the images respectively, $M \times S$ represents the total number of pixels, $F_t(j,k)$ and $F_f(j,k)$ represents the pixel intensity values of source images and fused image at location (j,k).	Closer to +1
Visual Information Fidelity Feng et al. (2022)	Measures the accuracy of visual information preserved in the fused image through human visual perception model	$VIF = \frac{\sum_{n=1}^{N} \sum_{d=1}^{D} \log\left(1 + \frac{g_n^2 \lambda_d}{\sigma_v^2}\right)}{\sum_{n=1}^{N} \sum_{d=1}^{D} \log\left(1 + \frac{\lambda_d}{\sigma_n^2}\right)}$ where, λ_d represents the eigenvalues of the covariance matrix of the source image in the $d-th$ spatial block. σ_v^2 and σ_n^2 shows the variance of noise in the fused and source image respectively. The term g_n indicates the gain factor modeling fusion in subband n. N is the count of subbands and D is the number of spatial blocks.	Higher

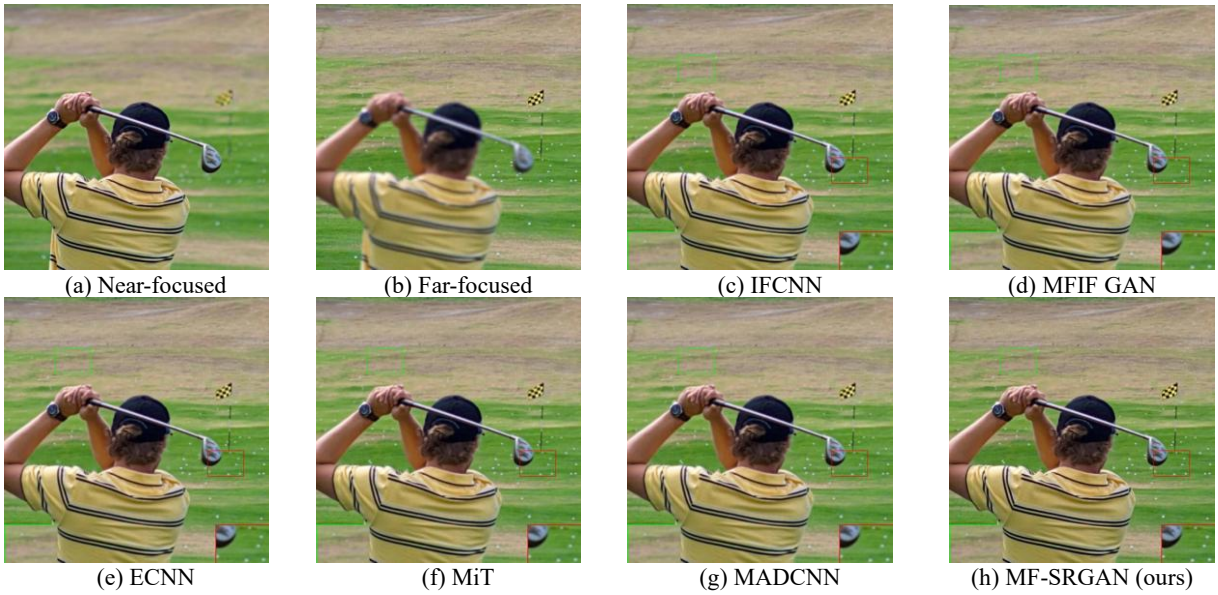
To assess the effectiveness of the proposed model, we compare our MF-SRGAN model with five state-of-the-art fusion techniques i.e., IFCNN (Zhang et al., 2020b), MFIF-GAN (Wang et al., 2021), ECNN (Amin-Naji et al., 2019), MiT (Karacan, 2023), and MADCNN (Lai et al., 2019).

4.3.2 Visual Analysis on Lytro Dataset

We chose two image pairs from the Lytro dataset and visually examined them to confirm the advantages of our MF-SRGAN model over other current methods.

Each pair has near-focused and far-focused images. **Figure 4(a)** is a near-focused image, since the man holding the golf club in the foreground appears clear. But the flag and the grass field in the background are clearly blurred. This shows that the camera was focused on objects closer to the lens. On the other hand, **Figure 4(b)** is called "far-focused" because the background, notably the checkered flag and the distant grass field, is quite clear. However, the man in the foreground looks blurry. This shows that the camera was set up to focus on distant objects rather than the nearby things. These two images show a typical multi-focus image pair, where different parts of a scene are in focus in each image.

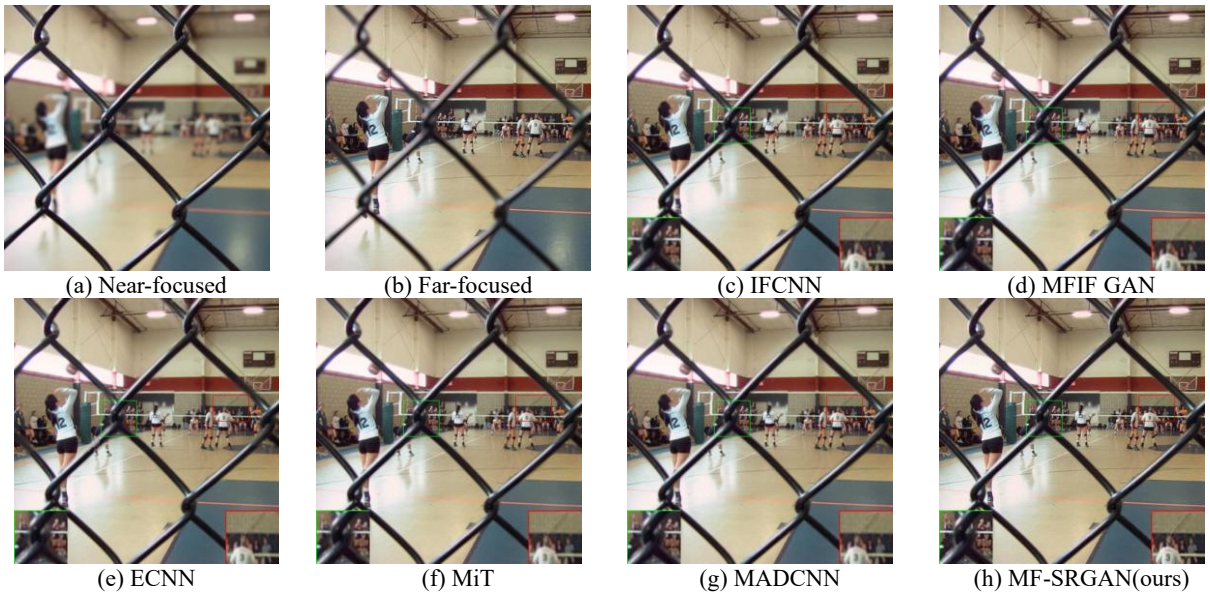
The results of fusion are shown in **Figures 4** to **6**. We chose two details among these two sets of image pairs for analysis. These details were marked with green and red rectangles and magnified. We have shown the zoomed-in portions from the "Gym" image pair as a representative set for further visual analysis.



(a) Near-focused (b) Far-focused (c) IFCNN (d) MFIF GAN

(e) ECNN (f) MiT (g) MADCNN (h) MF-SRGAN (ours)

Figure 4. Visual comparison of MF-SRGAN with five state-of-the-art methods on the "Golf" image pair from the Lytro dataset. Green and red boxes show zoomed regions of the field texture and the golf club, respectively, used to assess detail preservation after fusion.



(a) Near-focused (b) Far-focused (c) IFCNN (d) MFIF GAN

(e) ECNN (f) MiT (g) MADCNN (h) MF-SRGAN(ours)

Figure 5. Visual comparison of MF-SRGAN with five state-of-the-art methods on the "Gym" image pair from the Lytro dataset. Green and red boxes show zoomed regions of the chain link fence and player in jersey, respectively, used to assess detail preservation after fusion.

As observed in the enlarged part of **Figure 4**, our model shows a good focus and sharpness of the golf club in the boundary region. Clear fusion artifacts are visible in the transition regions between the nearby field and the golf club for the other five methods, emphasizing the challenges in seamlessly merging the focal planes. The color tones in the green highlighted regions of our method are well preserved, demonstrating a natural mixing of hues without discernible distortions. For the other methods, the grass and field areas exhibit irregular color blending or fading.

Figure 5 shows the fusion results of the six methods in the "Gym" image pairs. Initial inspection of the images shows that the proposed method provides better visual coherence between foreground and background.

Figure 6. Zoomed-in visual comparison of MF-SRGAN with five state-of-the-art methods on the "Gym" image pair from the Lytro dataset.

To enable better analysis, **Figure 6** provides zoomed-in visual comparisons of the green highlighted area (fence) and the red highlighted area (player in jersey). It can be observed from the green highlighted area in **Figure 6** that our model effectively preserves the fine details of the chain link fence with the contrast between the fence and the background being sharply defined to ensure clear separation. In the other five methods, the chain link fence either dominates or interferes, resulting in ambiguous blending and uneven

focus distribution. The player in the red highlighted region stands vividly against the background with recognizable details of the jersey. In the other methods, the details are partially defocused, reducing the visual realism of the fused image.

4.3.3 Comparative Analysis on Lytro Dataset

The six metrics mentioned in this section are used for the quantitative evaluation of 20 pair of images from the Lytro dataset with the MF-SRGAN model and five different methods. The results are shown in **Figures 7** to **12**. The top three averages of each metric have been highlighted with red, green, and blue fonts, respectively, on the chart.

As shown in **Figure 7**, MF-SRGAN achieves the highest average Spatial Frequency (SF) value with a 4.75% improvement over MFIF-GAN. This shows that the proposed model improves texture sharpness better than its counterparts.

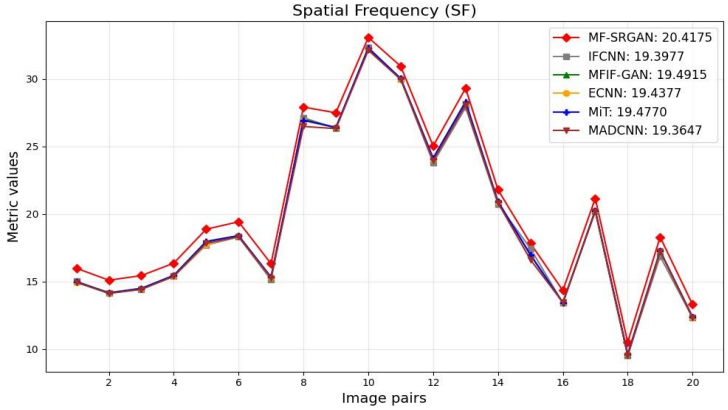


Figure 7. Comparison of spatial frequency metric on twenty image pairs of Lytro dataset.

The entropy values for all methods are shown in **Figure 8**, and MF-SRGAN achieves the highest Entropy (EN) of 7.5339. With a 0.105% improvement over the next-best-performing model of MFIF-GAN, the proposed model shows a slight but consistent advantage in retaining image information.

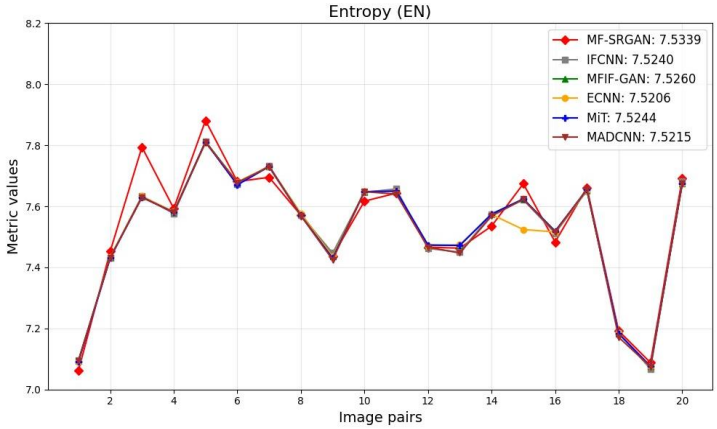


Figure 8. Comparison of entropy metric on twenty image pairs of Lytro dataset.

Figure 9 compares the values of Mutual Information (MI) for different models. With an average MI value of 6.1251, MF-SRGAN shows better preservation of common information between source and fused images. The proposed model has a 1.8% improvement over MADCNN, which is the next best-performing model.

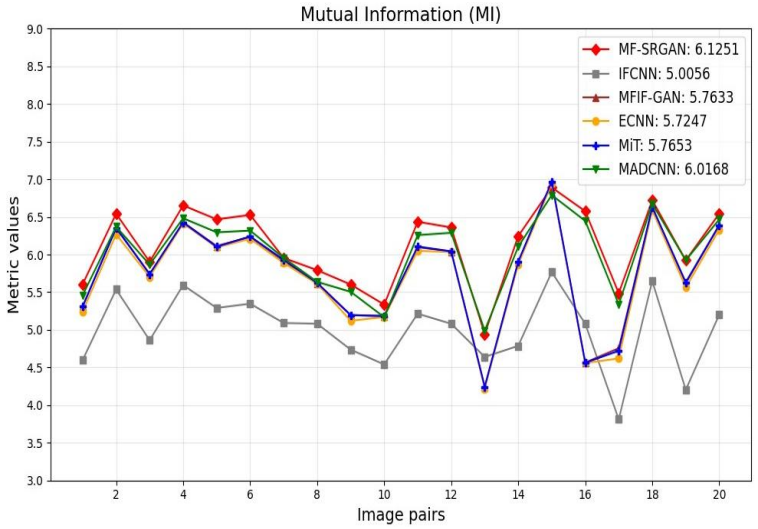


Figure 9. Comparison of mutual information metric on twenty image pairs of Lytro dataset.

Our model is second in the Standard Deviation (SD) metric as shown in **Figure 10**, where it is 0.0382 lower than the top-ranked method of MADCNN.

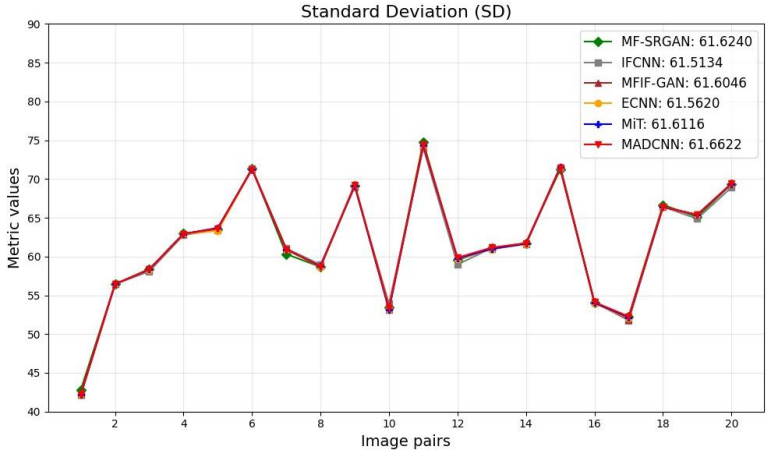


Figure 10. Comparison of standard deviation metric on twenty image pairs of Lytro dataset.

The graph in **Figure 11** shows the values of the Correlation Coefficient (CC) for six methods. MF-SRGAN achieves the highest CC value of 0.9790 with a 0.307% improvement over IFCNN, which is the second best-performing model. This result demonstrates that the proposed model achieves superior preservation of structural details and essential features in the fused image.

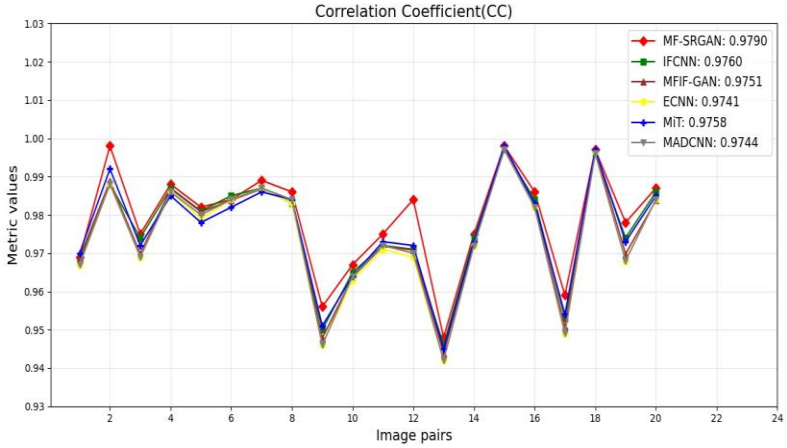


Figure 11. Comparison of correlation coefficient metric on twenty image pairs of Lytro dataset.

The Visual Information Fidelity (VIF) metric measures how effectively the fused image preserves the perceptual information compared to the source images. As shown in the graph of **Figure 12**, MF-SRGAN has the highest VIF score with an improvement of 1.21% over MADCNN. This indicates a better preservation of structural and perceptual information compared to other models.

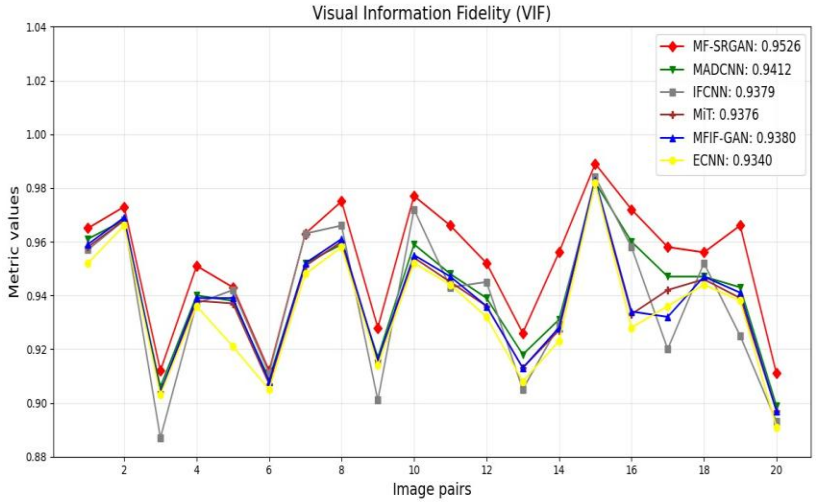


Figure 12. Comparison of visual information fidelity metric on twenty image pairs of Lytro dataset.

The results in **Table 6** indicate that the MF-SRGAN model exhibits the best fusion performance compared to the other five methods for the Lytro dataset. The improvements noted in SF, MI, and VIF demonstrate that MF-SRGAN successfully preserves high frequency components at focus boundaries. Unlike the traditional methods, the super resolution framework in MF-SRGAN improves feature representations before reconstruction. This makes it easier to reconstruct boundary areas where focused and defocused areas intersect.

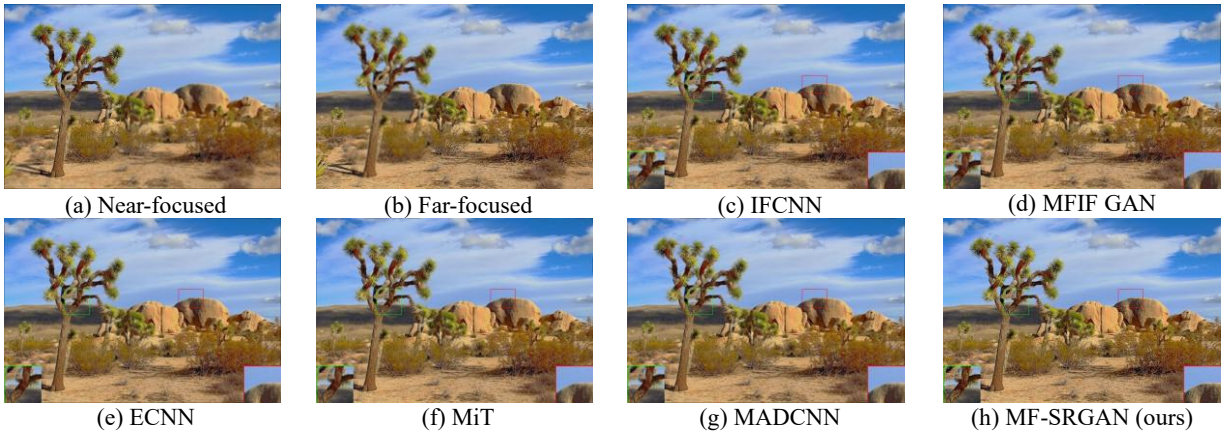
Table 6. Average score of six performance metrics on Lytro dataset.

Metric	IFCNN	MFIF-GAN	ECNN	MiT	MADCNN	MF-SRGAN
SF	19.3977	19.4915	19.4377	19.4770	19.3647	**20.4175**
EN	7.5240	7.5260	7.5206	7.5244	7.5215	**7.5339**
MI	5.0056	5.7633	5.7247	5.7653	6.0168	**6.1251**
SD	61.5134	61.6046	61.5620	61.6116	**61.6622**	61.6240
CC	0.9760	0.9741	0.9741	0.9742	0.9744	**0.9790**
VIF	0.9345	0.9380	0.9340	0.9376	0.9412	**0.9526**

4.3.4 Visual Analysis on MFI-WHU Dataset

The visual analysis of the MF-SRGAN model with other existing algorithms has been carried out on three image pairs from the MFI-WHU data set. The results are shown in **Figures 13** to **15**.

As observed in the zoomed-in portions of **Figure 13**, our model shows sharp edge preservation and fine details of the tree that easily differentiate it from its background. The five existing methods demonstrate blurring and loss of high-frequency details in the region surrounding the tree trunk and branches. In addition, edge artifacts result in inconsistent focus and soft edges. The magnified sections of the rock formation capture the natural texture of the surface and are devoid of any noticeable boundary artifacts. The images of the rock from existing methods exhibit a loss of texture fidelity. This results in unnaturally smooth surfaces and reduced sharpness in all methods.

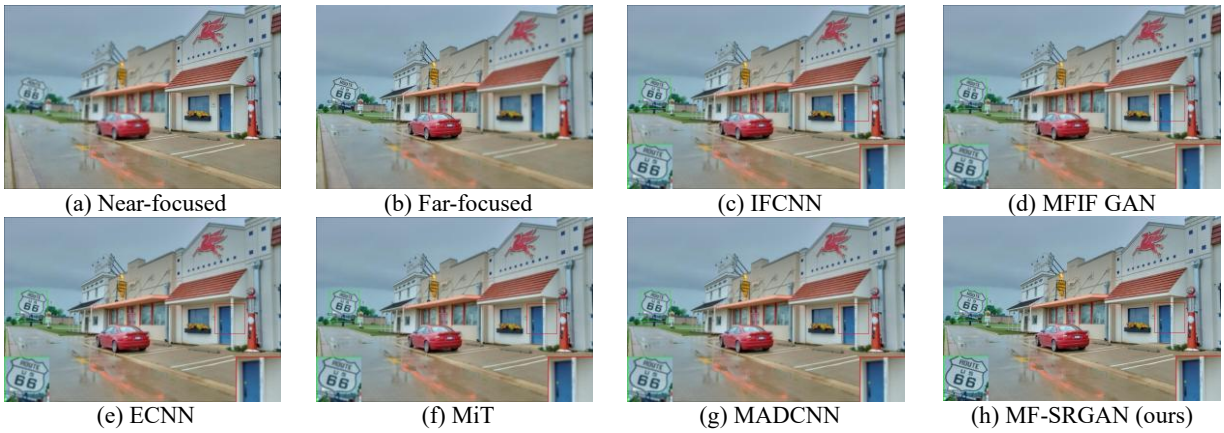


(a) Near-focused (b) Far-focused (c) IFCNN (d) MFIF GAN

(e) ECNN (f) MiT (g) MADCNN (h) MF-SRGAN (ours)

Figure 13. Visual comparison of MF-SRGAN with five state-of-the-art methods on the "Desert Plant" image pair from the MFI-WHU dataset. Green and red boxes show zoomed regions of the tree and rock formation, respectively, used to assess detail preservation after fusion.

Figure 14 shows the fusion results of the six methods in the "Diner" image pairs. The result of the proposed method makes the overall visual balance between the foreground and background appear more natural.



(a) Near-focused (b) Far-focused (c) IFCNN (d) MFIF GAN

(e) ECNN (f) MiT (g) MADCNN (h) MF-SRGAN (ours)

Figure 14. Visual comparison of MF-SRGAN with five state-of-the-art methods on the "Diner" image pair from the MFI-WHU dataset. Green and red boxes show zoomed regions of the route 66 sign and blue door, respectively, used to assess detail preservation after fusion.

(a) IFCNN (b) MFIF-GAN

(c) ECNN (d) MiT

(e) MADCNN (f) MF-SRGAN

Figure 15. Zoomed-in visual comparison of MF-SRGAN with five state-of-the-art methods on the "Diner" image pair from the MFI-WHU dataset.

For a more detailed assessment, **Figure 15** displays zoomed-in comparisons of the areas marked in green (Route 66 sign) and red (blue door and wall boundary). The green highlighted area in **Figure 15** shows the

route 66 sign from our model with easily readable text. The edges of the sign are sharp, and the background seamlessly integrates without loss of detail. The other five methods demonstrate blurring and loss of sharpness in the route 66 sign characterized by soft text edges. This feature disrupts the clarity expected in flat areas with high contrast. The door area indicated in the red highlighted box exhibits color accuracy and edge sharpness in our model. The magnified portion of the door area shows over-smoothing and inconsistent focus blending, resulting in an unnatural appearance across all other methods.

4.3.5 Comparative Analysis on MFI-WHU Dataset

To further evaluate the efficacy of the fusion of the MF-SRGAN model, we compared it with five other existing methods on 20 image pairs from the MFI-WHU dataset. The results are presented in **Figures 16** to **21**. The top three averages of each metric have been highlighted with red, green, and blue fonts, respectively, on the chart.

Figure 16 shows the Spatial Frequency values for the comparative models. This metric measures the degree of texture and edge information in an image. MF-SRGAN shows a 0.67% improvement over the second-best ECNN model and has the highest SF score in the MFI-WHU dataset. This shows that the proposed model can preserve finer details and leads to sharper and visually enriched fused images.

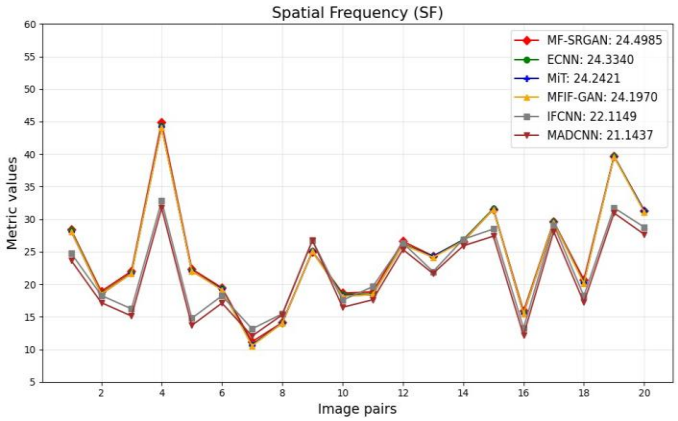


Figure 16. Comparison of spatial frequency metric on twenty image pairs of MFI-WHU dataset.

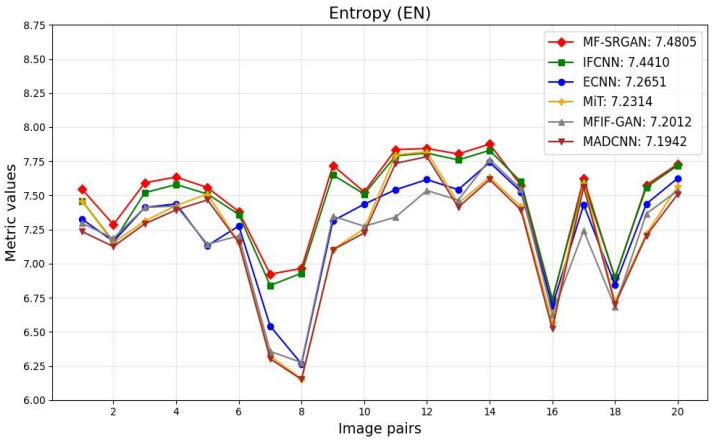


Figure 17. Comparison of entropy metric on twenty image pairs of MFI-WHU dataset.

Figure 17 shows that MF-SRGAN achieves the highest entropy value of 7.4805, surpassing the nearest performing model of IFCNN by 0.53%. This implies that MF-SRGAN preserves better feature diversity and contrast, thereby creating more informative fused images compared to other models.

Figure 18 shows the values of the Mutual Information metric for the comparative methods. This metric measures the degree of shared information between the fused and source images. MF-SRGAN outperforms the second-best ECNN model by 2.25% by achieving the highest MI value.

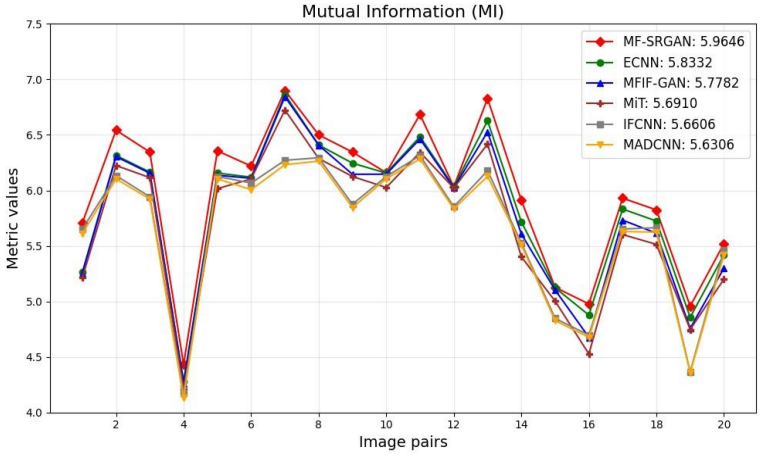


Figure 18. Comparison of mutual information metric on twenty image pairs of MFI-WHU dataset.

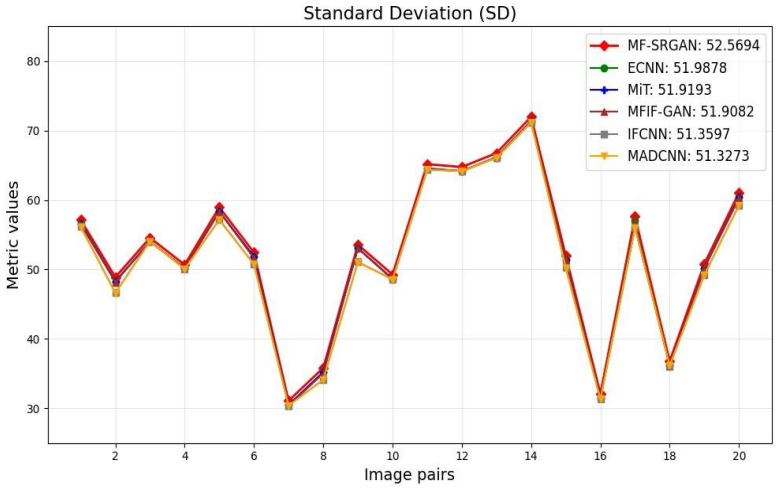


Figure 19. Comparison of standard deviation metric on twenty image pairs of MFI-WHU dataset.

The values of the Standard Deviation metric for the six models, which evaluates the intensity and contrast variations in the fused images, are shown in **Figure 19**. MF-SRGAN surpasses ECNN by 1.12% in the MFI-WHU dataset with the best SD value of 52.5694. This shows that MF-SRGAN improves the variation in pixel intensities and produces visually richer fused images compared to the other five methods.

As observed in **Figure 20**, the value of the correlation coefficient for MF-SRGAN is 0.0061 lower than the top-ranked model of ECNN.

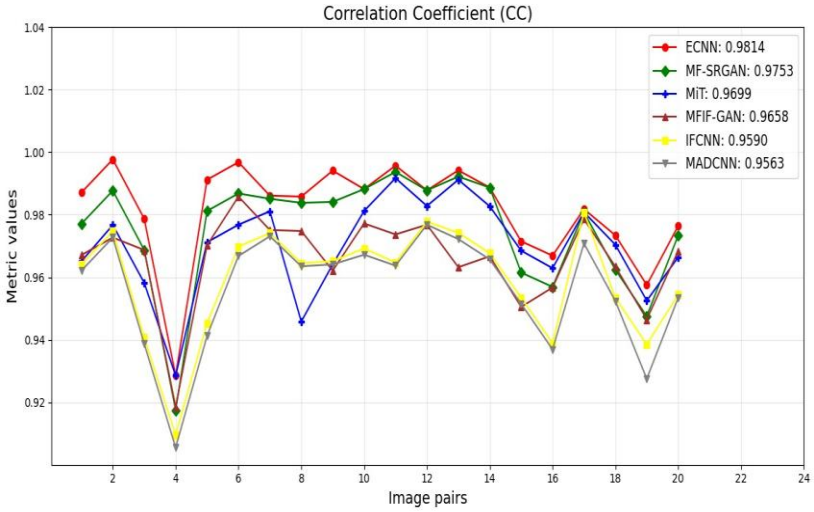


Figure 20. Comparison of correlation coefficient metric on twenty image pairs of MFIWHU dataset.

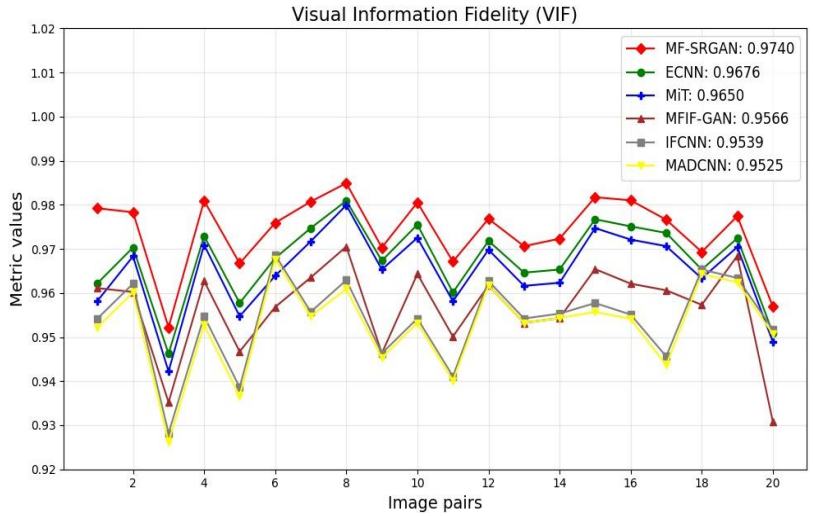


Figure 21. Comparison of visual information fidelity metric on twenty image pairs of MFI-WHU dataset.

Figure 21 shows that MF-SRGAN achieves the highest Visual Information Fidelity score of 0.9740, which is 0.66% better than ECNN. MF-SRGAN is therefore a preferable option for high-fidelity image fusion since it preserves finer image details and improves perceptual similarity to the reference image better than competing methods.

The scores in **Table 7** indicate that the fused images produced by MF-SRGAN display enhanced sharpness, richer details, and better retention of relevant features from the source images of MFI-WHU dataset while

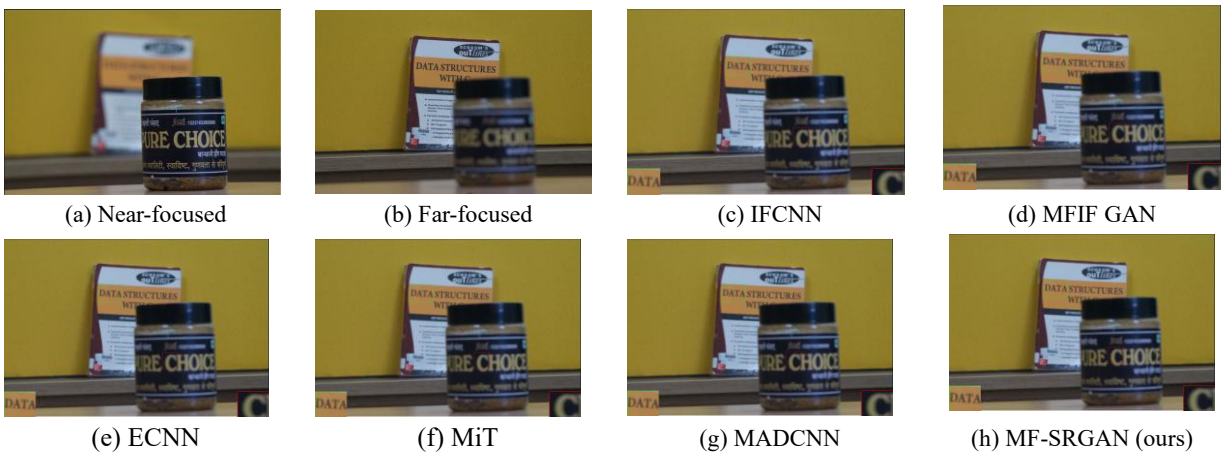
maintaining superior perceptual quality. The steady gains in SF and MI indicate that edge energy and structural content are better preserved at focus boundaries. Higher VIF scores show that focus areas have better perceptual sharpness.

Table 7. Average score of six performance metrics on MFI-WHU dataset.

Metric	IFCNN	MFIF-GAN	ECNN	MiT	MADCNN	MF-SRGAN
SF	22.1149	24.1970	24.3340	24.2421	21.1437	**24.4985**
EN	7.4410	7.2012	7.2651	7.2314	7.1942	**7.4805**
MI	5.6606	5.7782	5.8332	5.6910	5.6306	**5.9646**
SD	51.3597	51.9082	51.9878	51.9193	51.3273	**52.5694**
CC	0.9590	0.9658	**0.9814**	0.9699	0.9563	0.9753
VIF	0.9539	0.9566	0.9676	0.9650	0.9525	**0.9740**

4.3.6 Visual Analysis on Real-World Images

The visual analysis of the MF-SRGAN model with other existing algorithms has been carried out on two image pairs from the real-world data set. The results for the pair of indoor images are shown in **Figures 22** to **23**. As shown in **Figure 22**, the proposed MF-SRGAN produces images with visually sharper structures and more natural foreground-background transitions than the comparative methods.



(a) Near-focused (b) Far-focused (c) IFCNN (d) MFIF GAN

(e) ECNN (f) MiT (g) MADCNN (h) MF-SRGAN (ours)

Figure 22. Visual comparison of MF-SRGAN with five state-of-the-art methods on the indoor image pair from the real-world dataset.

Figure 23 shows the zoomed-in comparisons of areas highlighted in green (the text "DATA") and red (the character "C"). The fine text structures in "DATA" are preserved with sharper stroke boundaries. The character contours in the red highlighted region are well defined with minimal smoothing. This shows retention of high-frequency details and consistent focus blending. The five comparative methods show attenuation of edge sharpness in green and red highlighted areas. Text strokes appear slightly diffused and softened, suggesting inadequate preservation of discriminative details during fusion.

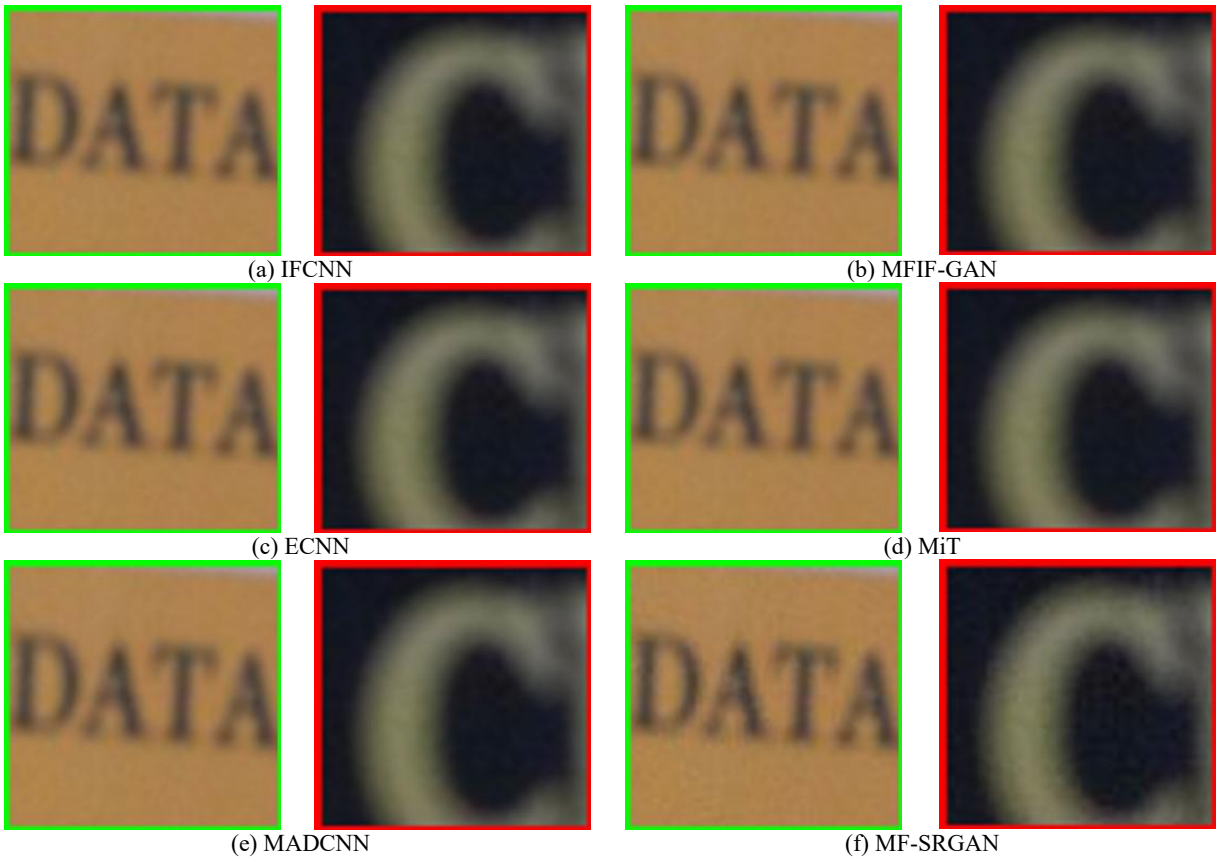


Figure 23. Zoomed-in visual comparison of MF-SRGAN with five state-of-the-art methods on the indoor image pair from the real-world dataset.

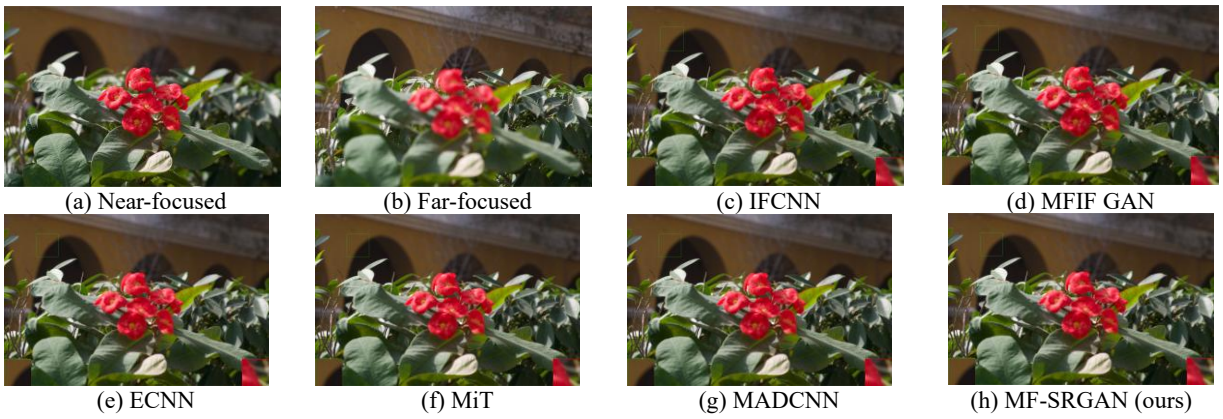


Figure 24. Visual comparison of MF-SRGAN with five state-of-the-art methods on the outdoor image pair from the real-world dataset.

The results for the outdoor image pair are shown in **Figure 24**. The proposed method preserves the fine texture details of the flower petals and the structural contours of the building arch in the background. The

comparative approaches show substantial blur retention in the backdrop arches. The images show slight loss of high-frequency features in the floral region. This indicates incomplete focus integration.

4.3.7 Comparative Analysis on Real-World Images

We conducted a comparative assessment on 10 pairs of real-world images to confirm the generalization capacity of the proposed model. **Table 8** shows the average quantitative results for six performance metrics, comparing MF-SRGAN with five other methods. According to the table, MF-SRGAN achieves the highest values in all metrics except the correlation coefficient. This indicates superior information retention and preservation of details. These results show the effectiveness of the proposed model in real-world scenarios.

Table 8. Average score of six performance metrics on real-world dataset.

Metric	IFCNN	MFIF-GAN	ECNN	MiT	MADCNN	MF-SRGAN
SF	14.6742	14.9631	15.1057	15.2625	13.9833	**16.4268**
EN	6.9183	7.1065	7.2312	7.2156	6.7334	**7.2527**
MI	2.6514	2.7655	2.9614	2.8433	2.5742	**4.0024**
SD	46.5391	46.8346	47.1771	47.0621	46.1821	**48.3927**
CC	0.7536	0.7672	**0.8036**	0.7827	0.7486	0.7941
VIF	0.7883	0.7968	0.8257	0.8179	0.7627	**0.9036**

4.3.8 Statistical Significance Testing

T-tests are important for ascertaining if the variations in performance metrics between methods are statistically significant or occurred by random chance. We conducted paired t-tests between MF-SRGAN and the next best-performing method for each metric. The results for the Lytro dataset are presented in **Table 9** and for the MFI-WHU dataset are shown in **Table 10**. Statistically significant differences ($p < 0.05$) are in bold.

Table 9. Statistical significance testing (t-test) results for the Lytro dataset.

Metric	Comparison	p-value	t-statistic	Result
SF	MF-SRGAN vs MFIF-GAN	**4.7536E-23**	59.4357	✓
EN	MF-SRGAN vs MFIF-GAN	0.448	0.7747	×
MI	MF-SRGAN vs MADCNN	**1.7124E-06**	6.8001	✓
CC	MF-SRGAN vs IFCNN	**0.0008**	3.9427	✓
VIF	MF-SRGAN vs MADCNN	**8.4965E-08**	8.3707	✓

Table 10. Statistical significance testing (t-test) results for the MFI-WHU dataset.

Metric	Comparison	p-value	t-statistic	Result
SF	MF-SRGAN vs ECNN	**0.002**	3.5649	✓
EN	MF-SRGAN vs IFCNN	**0.0001**	4.8065	✓
MI	MF-SRGAN vs ECNN	**1.3947E-05**	5.7937	✓
SD	MF-SRGAN vs ECNN	**1.8676E-24**	70.5267	✓
VIF	MF-SRGAN vs ECNN	**1.1247E-08**	9.5361	✓

Compared to the best baseline approaches, the T-tests in the Lytro dataset show that MF-SRGAN achieves statistically significant gains ($p < 0.05$) in SF, MI, CC and VIF. Although no significant improvement is visible for Entropy ($p = 0.448$), the proposed method still outperforms other metrics.

The results of paired t tests on the MFI-WHU dataset show that MF-SRGAN delivers statistically significant improvements ($p < 0.05$) in metrics such as SF, EN, MI, SD and VIF.

4.3.9 Ablation Experiments

We performed a series of ablation experiments to analyze the impact of various architectural and loss function choices on the performance of our proposed model. The ablation experiments were conducted on ten pairs of multi-focus images from the Lytro dataset.

The variations were as follows:

- *Reduced residual blocks*: The number of residual blocks was reduced from 16 to 8 to investigate how deep feature extraction affects fusion quality.
- *Only content loss* ($L_{content}$): The model was trained solely with content loss and excluding adversarial loss.
- *No upsampling layer:* The upsampling layer was eliminated from the generator architecture to examine its necessity.
- *Only adversarial loss* (L_{adv}): The model was trained only with adversarial loss and elimination of content loss.

Figures 25 to **30** shows the experimental result comparison between the proposed method and module variations on six metrics.

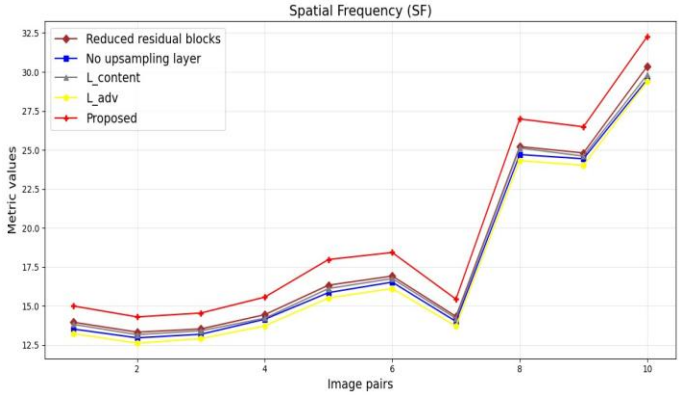


Figure 25. Result of the ablation experiment on spatial frequency for ten pairs of images from Lytro dataset.

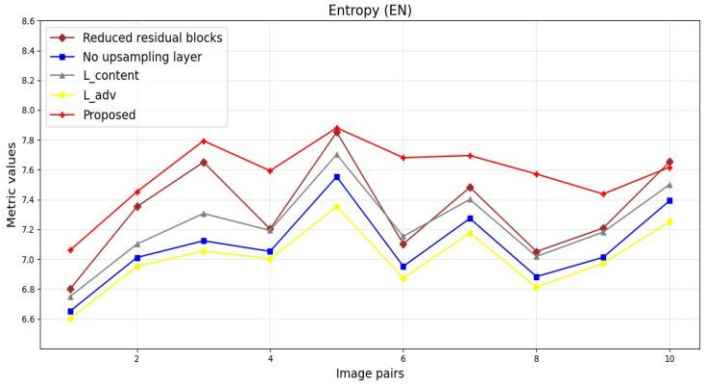


Figure 26. Result of the ablation experiment on entropy for ten pairs of images from Lytro dataset.

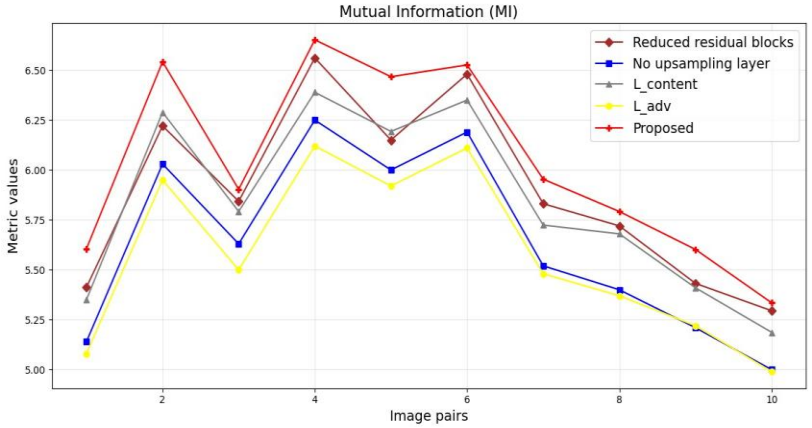


Figure 27. Result of the ablation experiment on mutual information for ten pairs of images from Lytro dataset.

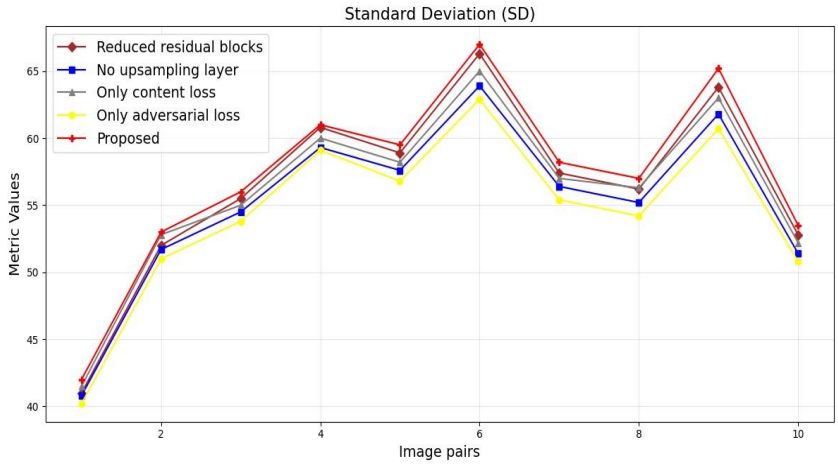


Figure 28. Result of the ablation experiment on standard deviation for ten pairs of images from Lytro dataset.

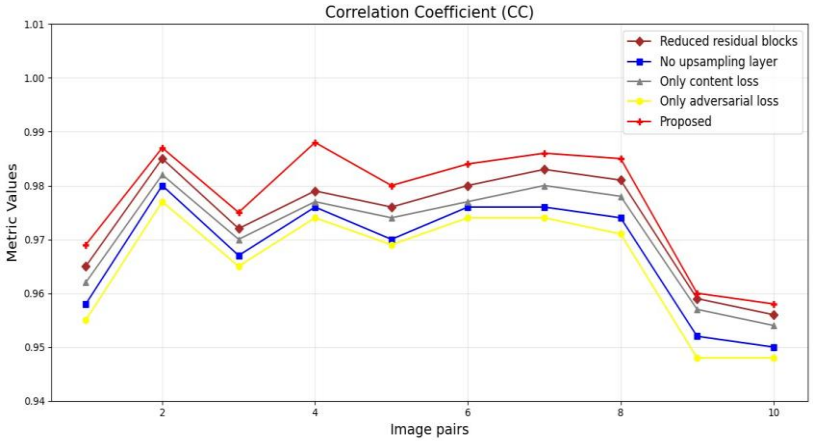


Figure 29. Result of the ablation experiment on correlation coefficient for ten pairs of images from Lytro dataset.

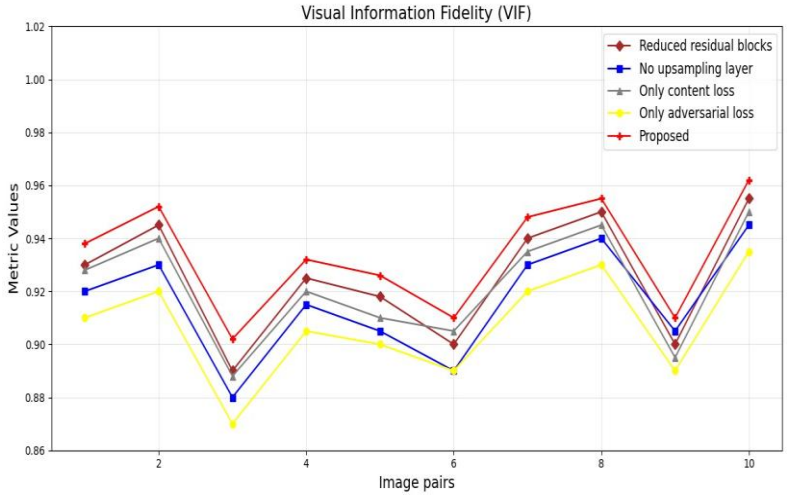


Figure 30. Result of the ablation experiment on visual information fidelity for ten pairs of images from Lytro dataset.

It is clear that decreasing the number of residual blocks led to a decline in image quality due to loss of fine details. This confirms the need for deeper feature extraction for optimal fusion. The experiment with only content loss generated fused images with strong structural consistency. However, the images lacked perceptual sharpness, therefore highlighting the need for adversarial loss to improve visual details. Conversely, training with only adversarial loss led to better outputs but impaired content fidelity and caused structural inconsistencies. Eliminating the upsampling layer of the generator greatly reduced the quality of the image. This underlines the importance of the upsampling layer for high quality reconstruction of the fused image. These variations produced varying degrees of blurring, loss of contrast, and degraded fusion quality at focus boundaries compared to the proposed model.

These results indicate that both content and adversarial losses help to achieve a balance between structure and sharpness. Also, deeper architectures and upsampling layers are needed for effective multi-focus image fusion.

4.3.10 Sensitivity Analysis

We performed sensitivity analysis using the VIF metric, which indicates perceptual quality, and the MI metric, which reflects structural fidelity in the fused image. This analysis evaluated the impact of adversarial loss weight (λ) on fusion performance. We ran a number of tests on the Lytro and MFI-WHU datasets with different (λ) values and then presented the corresponding metric values. The results are shown as bar charts in **Figures 31** to **32**.

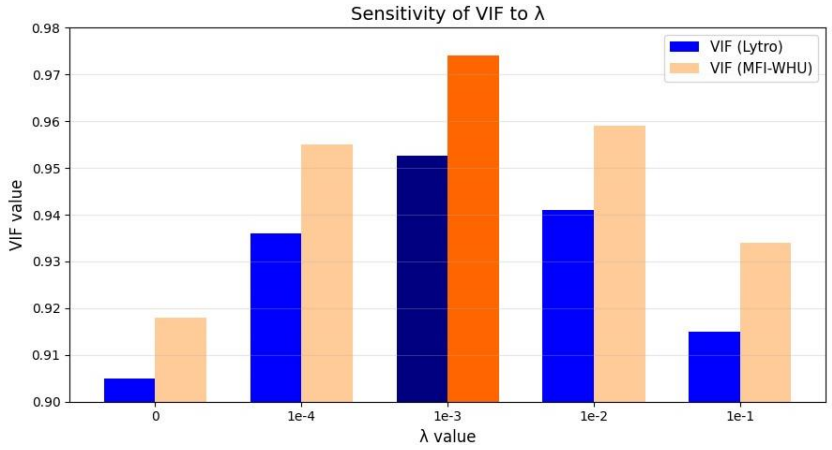


Figure 31. Sensitivity analysis of VIF for Lytro and MFI-WHU datasets.

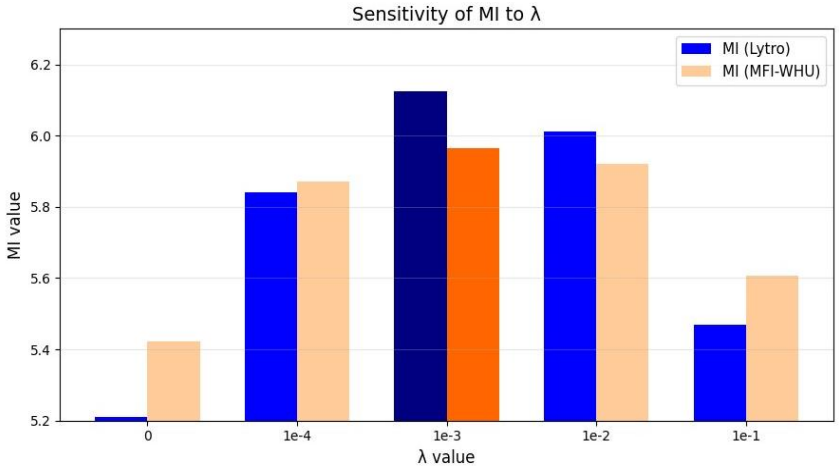


Figure 32. Sensitivity analysis of MI for Lytro and MFI-WHU datasets.

The selected λ is highlighted in dark colors (navy blue and dark orange). The bar charts show that $\lambda = 10^{-3}$ yields the highest VIF and MI scores for both datasets. This makes it the best choice for adversarial loss weight. The graph shows that both perceptual and structural scores improve, reaching a peak at $\lambda = 10^{-3}$, after which the performance goes down. This clearly shows that a moderate adversarial influence is a good way to balance perceptual enhancement and structural preservation.

4.3.11 Computational Efficiency

Given that the source image for MFIF is generally in the RGB color space, our preprocessing pipeline transforms the input image into many formats to ensure compatibility and appropriate data handling during training and inference. We converted images from the RGB color space to the BGR colour space with OpenCV for efficient image processing. For tasks necessitating tensor-based processing, the images are then transformed into PyTorch tensors. The processed images are converted into the RGB color space, facilitating consistent visualization and generation of fused images.

The GPU configuration used in this paper is NVIDIA GeForce RTX 3080 Ti. The CPU configuration employed for traditional methods and additional computations is the AMD Ryzen 9 3900X, which features 12 cores, a clock speed of 3.8 GHz, and 32 GB of RAM. This configuration ensures effective handling of computationally demanding tasks, such as training deep learning models and preprocessing large datasets. The software environment comprises PyTorch, the CUDA Toolkit, and additional libraries for GPU-accelerated operations.

The computational efficiency is evaluated by comparing the average execution time of different methods on ten image pairs of Lytro and MFI-WHU datasets as shown in **Table 11**.

Table 11. Comparison of average execution times on Lytro and MFI-WHU datasets.

Method	Average time for Lytro	Average time for MFI-WHU
IFCNN	3.214 s	3.015 s
MFIF-GAN	4.537 s	4.203 s
ECNN	3.041 s	2.976 s
MiT	5.162 s	5.048 s
MADCNN	3.514 s	3.187 s
MF-SRGAN	3.737 s	3.564 s

CNN-based methods such as IFCNN, ECNN, MADCNN are faster than transformer-based models such as MiT due to lower computational overhead. GAN-based methods such as MFIF-GAN are slightly slower due to adversarial training. The execution time of the proposed method is equal to that of the other deep learning-based methods and offers a balanced trade-off between speed and performance. MF-SRGAN model has approximately 6.4 million parameters (1.7 million in the generator and 4.7 million in the discriminator). During inference, processing a 520×520 image pair requires less than 1 GB of GPU memory and takes a few milliseconds per pair depending on the runtime configuration. Although adversarial training increases the complexity of the training, inference remains computationally efficient. This makes the model suitable for real-time deployment with model compression.

5. Discussions, Limitations, and Future Directions

5.1 Discussions

The proposed MF-SRGAN effectively addresses the limitations of traditional image fusion techniques such as SEWML (Wang et al., 2022), LSDGF1 (You and Yang, 2022)) and deep learning-based methods (Amin-Naji et al., 2019; Karacan, 2023; Lai et al., 2019; Wang et al., 2021; Zhang et al., 2020b) by using a Super-Resolution Generative Adversarial Network. Mo st techniques rely on decision maps or hand-crafted rules for the fusion of images. However, MF-SRGAN employs a dual-path generator with residual learning and upsampling layers. This ensures high-resolution fusion and improved textual consistency. Additionally, the PatchGAN-based discriminator ensures both global and local coherence in the fused output. This reduces issues such as boundary artifacts and information loss. Moreover, the generated dataset derived from COCO helps the model to be adaptable to real-world scenarios, since it is trained on diverse focus patterns.

We provide a response to the research questions raised in Section 1.1 as follows:

(a) RQ1: How effectively can a GAN model have based on super-resolution preserve details in multi-focus image fusion compared to existing methods? Sections 4.3.2 - 4.3.7 provides a detailed response to this question in terms of visual and quantitative analysis. The comparisons show that the proposed method preserves edge sharpness and object boundaries across varying depth regions. Furthermore, as observed from **Tables 6, 7,** and **8**, MF-SRGAN achieves superior performance in most metrics with the

Lytro and MFI-WHU datasets. The use of super-resolution in its methodology ensures better image clarity and contrast retention.

(b) RQ2: How much do architectural advancements in the SRGAN framework improve visual and quantitative fusion quality? Sections 3.1 - 3.2 describes the architectural enhancements done on the baseline SRGAN model to adapt it for multi-focus image fusion. Dual-stream convolutional pathways are used to handle the two input images, and the upsampling layer block is tailored for fusion tasks. These adjustments help maintain spatial consistency and improve texture detail in the fused output. The contribution of each component is shown through ablation studies in Section 4.3.9. This shows that the proposed improvements make the baseline SRGAN model much better in multi-focus scenarios.

(c) RQ3: Are the improvements achieved by the proposed method statistically significant and consistent across the image fusion metrics? Section 4.3.8 shows that the improvements made by MF-SRGAN are statistically significant for most fusion metrics. The p-values for important metrics such as MI, SF, CC, and VIF are significantly below 0.05 for both the Lytro and MFI-WHU datasets. This demonstrates the robustness of the proposed method. Although entropy does not exhibit significance in one comparison, the fact that the results are consistent across datasets and metrics is notable.

5.2 Limitations

The proposed MF-SRGAN model has certain limitations as follows:

(a) Adversarial training, VGG19-based feature extraction, and complex generator architecture increase computational overhead, making real-time performance challenging.
(b) The architecture uses high-resolution image processing, multiple convolutional pathways which lead to high GPU consumption. This limits scalability for large datasets or edge devices.
(c) The effectiveness of MF-SRGAN depends on large training datasets with varied focus patterns. However, any bias in the dataset may affect its generalization to unseen multi-focus image pairs.
(d) The proposed MF-SRGAN requires more training time than traditional fusion techniques. This is due to its complex adversarial training process and loss calculations.
(e) Optimizing MF-SRGAN is challenging, as its performance is dependent on hyperparameter settings, including adversarial balance, learning rates, and loss function weights.
(f) The existing implementation does not use optimization techniques like model pruning, quantization, or lightweight architectural alternatives, which limits its use in real-time or low-power environments.

5.3 Future Directions

Although the proposed MF-SRGAN shows superior performance in multi-focus image fusion, some challenges still remain that need to be addressed for broader applicability.

(a) Evaluation across multiple datasets: The proposed MF-SRGAN can be evaluated on datasets of medical, thermal-visible, and hyperspectral images to ascertain its generalization across varied fusion scenarios. Future research may explore whether the same generator-discriminator configuration works across modalities or needs domain-specific adaptation.
(b) Adaptive weight balancing: Dynamic weight modification of several loss components, such as content, perceptual, and adversarial losses based on image complexity or focus variations, can help to further enhance the fusion performance. Attention driven or reinforcement learning can be explored to adjust loss functions during training.
(c) Model optimization for deployment: The network architecture of MF-SRGAN can be optimized by incorporating efficient convolutional layers and pruning techniques. This can expedite both training and inference while preserving fusion quality. Additionally, lightweight backbone architectures such as MobileNet and neural architecture search (NAS) may be studied to further reduce computing cost.

Quantization methods can also be used to enable real-time deployment on edge devices with limited resources.

(d) Emerging models: Transformer-based models and uncertainty-aware diffusion frameworks present promising directions to improve preservation of focus boundary and fine-detail fusion. These models can improve spatial attention, texture reconstruction, and uncertainty modeling in ambiguous areas. Future research will also involve benchmarking MF-SRGAN against recent transformer-based fusion models, namely SwinFusion and TransMEF, to thoroughly assess comparative performance.

(e) Unsupervised and semi-supervised learning: There are many challenges in obtaining paired ground truth data in the domain of multi-focus image fusion. Future research can investigate unsupervised and semi-supervised learning frameworks. Using unlabelled or weakly labelled data for training can be made more effective with methods like cycle-consistency and contrastive learning.

6. Conclusions

This paper presented a novel approach for multi-focus image fusion customizing a SRGAN framework. We utilized a newly developed dataset derived from the COCO dataset to efficiently fuse multi-focus input images into a single, high-resolution, all-in-focus image. The proposed MF-SRGAN model addresses key limitations of current GAN-based MFIF methods by improving texture preservation, boundary sharpness, and incorporating perceptual loss to preserve structural consistency. Unlike other methods, MF-SRGAN generates high-quality fused images without depending on decision maps by using a customized super-resolution framework, hence minimizing information loss. Extensive experiments and resultant comparative analysis with ECNN, MiT, MFIF-GAN, MADCNN, and IFCNN reveal that MF-SRGAN achieves superior performance with improvements ranging from 0.3-4.75% across various metrics on the Lytro dataset and 0.53-2.25% on the MFI-WHU dataset. Although MF-SRGAN's complex network architecture causes slower computational speed than IFCNN, ECNN, and MADCNN, its improved fusion quality and capacity to preserve fine details makes it a promising solution for multi-focus image fusion applications. Further optimizations and evaluations on additional datasets could investigate the method's wider application and scalability. Overall, the proposed MFSRGAN shows that integrating super-resolution with adversarial learning can be useful for advanced image fusion tasks.

Conflicts of Interest

The authors confirm that there is no conflict of interest to declare for this publication.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors would like to thank the editor and anonymous reviewers for their comments that help improve the quality of this work.

AI Disclosure

During the preparation of this work the author(s) used generative AI in order to improve the language of the article. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

Amin-Naji, M., Aghagolzadeh, A., & Ezoji, M. (2019). Ensemble of cnn for multi-focus image fusion. *Information Fusion*, *51*, 201-214. https://doi.org/10.1016/j.inffus.2019.02.003

Bhalla, K., Koundal, D., Sharma, B., Hu, Y.-C., & Zaguia, A. (2022). A fuzzy convolutional neural network for enhancing multi-focus image fusion. *Journal of Visual Communication and Image Representation*, *84*, 103485. https://doi.org/10.1016/j.jvcir.2022.103485

Bouzos, O., Andreadis, I., & Mitianoudis, N. (2019). Conditional random field model for robust multi-focus image fusion. *IEEE Transactions on Image Processing*, *28*(11), 5636-5648. https://doi.org/10.1109/TIP.2019.2922097

Chai, J., Zeng, H., Li, A., & Ngai, E.W.T. (2021). Deep learning in computer vision: a critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, *6*, 100134. https://doi.org/10.1016/j.mlwa.2021.100134

Chen, Q., Yang, B., Li, Y., & Pang, L. (2020). Multi-focus image fusion with point detection filter and superpixel-based consistency verification. *IEEE Access*, *8*, 99956-99973. https://doi.org/10.1109/ACCESS.2020.2997370

Duan, Z., Luo, X., & Zhang, T. (2024). Combining transformers with CNN for multi-focus image fusion. *Expert Systems with Applications*, *235*, 121156. https://doi.org/10.1016/j.eswa.2023.121156

Feng, Y., Guo, R., Shen, X., & Zhang, X. (2022). A measure for the evaluation of multi-focus image fusion at feature level. *Multimedia Tools and Applications*, *81*(13), 18053-18071. https://doi.org/10.1007/s11042-022-11976-3

Ghandour, C., El-Shafai, W., El-Rabaie, S., & Abdelsalam, N. (2024). Comprehensive performance analysis of different medical image fusion techniques for accurate healthcare diagnosis applications. *Multimedia Tools and Applications*, *83*(8), 24217-24276. https://doi.org/10.1007/s11042-023-16334-5

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. *Machine Learning*. https://doi.org/10.48550/arXiv.1406.2661

Gross, S., & Wilber, M. (2016). *Training and investigating residual nets.* http://torch.ch/blog/2016/02/04/resnets.html [Accessed: 10/04/2025])

Guo, X., Nie, R., Cao, J., Zhou, D., Mei, L., & He, K. (2019). Fusegan: learning to fuse multi-focus image via conditional generative adversarial network. *IEEE Transactions on Multimedia*, *21*(8), 1982-1996. https://doi.org/10.1109/TMM.2019.2895292

Hu, G., Jiang, J., Sheng, G., & Wei, G. (2025). Msdcnet: multi-stage and deep residual complementary multi-focus image fusion network based on multi-scale feature learning. *Applied Intelligence*, *55*(3), 232. https://doi.org/10.1007/s10489-024-05983-0

Huang, J., Le, Z., Ma, Y., Mei, X., & Fan, F. (2020). A generative adversarial network with adaptive constraints for multi-focus image fusion. *Neural Computing and Applications*, *32*(18), 15119-15129. https://doi.org/10.1007/s00521-020-04863-1

Jie, Y., Li, X., Tan, H., Zhou, F., & Wang, G. (2024). Multi-modal medical image fusion via multi-dictionary and truncated Huber filtering. *Biomedical Signal Processing and Control*, *88*(Part B), 105671. https://doi.org/10.1016/j.bspc.2023.105671

Karacan, L. (2023). Multi-image transformer for multi-focus image fusion. *Signal Processing: Image Communication*, *119*, 117058. https://doi.org/10.1016/j.image.2023.117058

Lai, R., Li, Y., Guan, J., & Xiong, A. (2019). Multi-scale visual attention deep convolutional neural network for multi-focus image fusion. *IEEE Access*, *7*, 114385-114399. https://doi.org/10.1109/ACCESS.2019.2935006

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. *Computer Vision and Pattern Recognition.* https://doi.org/10.48550/arXiv.1609.0480

Li, H., Qian, W., Nie, R., Cao, J., & Xu, D. (2023a). Siamese conditional generative adversarial network for multi-focus image fusion. *Applied Intelligence*, *53*(14), 17492-17507. https://doi.org/10.1007/s10489-022-04406-2

Li, J., Li, B., & Jiang, Y. (2023b). GIPC-GAN: an end-to-end gradient and intensity joint proportional constraint generative adversarial network for multi-focus image fusion. *Complex & Intelligent Systems*, *9*(6), 7395-7422. https://doi.org/10.1007/s40747-023-01151-y

Li, X., Li, X., Cheng, X., Wang, M., & Tan, H. (2023c). MCDFD: multi-focus image fusion based on multi-scale cross-difference and focus detection. *IEEE Sensors Journal*, *23*(24), 30913-30926. https://doi.org/10.1109/JSEN.2023.3330871

Li, X., Zhou, F., Tan, H., Chen, Y., & Zuo, W. (2021). Multi-focus image fusion based on nonsubsampled contourlet transform and residual removal. *Signal Processing*, *184*, 108062. https://doi.org/10.1016/j.sigpro.2021.108062

Li, Y., & Jiang, S. (2020). Multi-focus image fusion using geometric algebra based discrete Fourier transform. *IEEE Access*, *8*, 60019-60028. https://doi.org/10.1109/ACCESS.2020.2981814

Li, Y., Li, X., Wang, J., Chen, G., Xu, J., Tang, Z., Yu, Z., Sun, X., Wang, J., & Yu, H. (2024). Multi-focus image fusion for microscopic depth-of-field extension of waterjet-assisted laser processing. *The International Journal of Advanced Manufacturing Technology*, *131*(3), 1717-1734. https://doi.org/10.1007/s00170-024-13118-5

Liu, W., Zheng, Z., & Wang, Z. (2021). Robust multi-focus image fusion using lazy random walks with multiscale focus measures. *Signal Processing*, *179*, 107850. https://doi.org/10.1016/j.sigpro.2020.107850

Liu, Y., Chen, X., Peng, H., & Wang, Z. (2017). Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, *36*, 191-207. https://doi.org/10.1016/j.inffus.2016.12.001

Liu, Y., Qi, Z., Cheng, J., & Chen, X. (2024). Rethinking the effectiveness of objective evaluation metrics in multi-focus image fusion: a statistic-based approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(8), 5806-5819. https://doi.org/10.1109/TPAMI.2024.3367905

Liu, Z., Blasch, E., Xue, Z., Zhao, J., Laganiere, R., & Wu, W. (2012). Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(1), 94-109. https://doi.org/10.1109/TPAMI.2011.109

Luo, F., Zhao, B., Fuentes, J., Zhang, X., Ding, W., Gu, C., & Pino, L.R. (2025). A review on multi-focus image fusion using deep learning. *Neurocomputing*, *618*, 129125. https://doi.org/10.1016/j.neucom.2024.129125

Ma, B., Zhu, Y., Yin, X., Ban, X., Huang, H., & Mukeshimana, M. (2021). SESF-Fuse: an unsupervised deep model for multi-focus image fusion. *Neural Computing and Applications, 33*(11), 5793-5804. https://doi.org/10.1007/s00521-020-05358-9

Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., & Ma, Y. (2022). SwinFusion: cross-domain long-range learning for general image fusion via Swin transformer. *IEEE/CAA Journal of Automatica Sinica*, *9*(7), 1200-1217. https://doi.org/10.1109/JAS.2022.105686

Ma, J., Zhou, Z., Wang, B., Miao, L., & Zong, H. (2019). Multi-focus image fusion using boosted random walks-based algorithm with two-scale focus maps. *Neurocomputing*, *335*, 9-20. https://doi.org/10.1016/j.neucom.2019.01.048

Manchanda, M., & Gambhir, D. (2019). Multifocus image fusion based on waveatom transform. *Sa¯dhana¯*, *44*(2), 49. https://doi.org/10.1007/s12046-018-1010-z

Nagarathinam, S., Vasuki, A., & Paramasivam, K. (2024). Deep remote fusion: development of improved deep CNN with atrous convolution-based remote sensing image fusion. *The Imaging Science Journal*, *72*(3), 382-402. https://doi.org/10.1080/13682199.2023.2206761

Nejati, M., Samavi, S., & Shirani, S. (2015). Multi-focus image fusion using dictionary based sparse representation. *Information Fusion*, *25*, 72-84. https://doi.org/10.1016/j.inffus.2014.10.004

Qiu, X., Li, M., Zhang, L., & Yuan, X. (2019). Guided filter-based multi-focus image fusion through focus region detection. *Signal Processing: Image Communication*, *72*, 35-46. https://doi.org/10.1016/j.image.2018.12.004

Qu, L., Liu, S., Wang, M., & Song, Z. (2022). TransMEF: a transformer-based multi-exposure image fusion framework using self-supervised multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(2), 2126-2134. https://doi.org/10.1609/aaai.v36i2.20109

Shao, X., Jin, X., Jiang, Q., Miao, S., Wang, P., & Chu, X. (2024). Multi-focus image fusion based on transformer and depth information learning. *Computers and Electrical Engineering*, *119*, 109629. https://doi.org/10.1016/j.compeleceng.2024.10962

Singh, S., Singh, H., Bueno, G., Deniz, O., Singh, S., Monga, H., Hrisheekesha, P.N., & Pedraza, A. (2023). A review of image fusion: methods, applications and performance metrics. *Digital Signal Processing*, *137*, 104020. https://doi.org/10.1016/j.dsp.2023.104020

Song, Y., Li, M., Li, Q., & Sun, L. (2006). A new wavelet based multi-focus image fusion scheme and its application on optical microscopy. In *2006 IEEE International Conference on Robotics and Biomimetics* (pp. 401-405). IEEE. Kunming, China. https://doi.org/10.1109/ROBIO.2006.340210

Tian, J., Liu, G., & Liu, J. (2018). Multi-focus image fusion based on edges and focused region extraction. *Optik*, *171*, 611-624. https://doi.org/10.1016/j.ijleo.2018.06.093

Wang, C., Yan, K., Zang, Y., Zhou, D., & Nie, R. (2024a). Focus-aware and deep restoration network with transformer for multi-focus image fusion. *Digital Signal Processing*, *149*, 104473. https://doi.org/10.1016/j.dsp.2024.104473

Wang, C., Zang, Y., Zhou, D., Mei, J., Nie, R., & Zhou, L. (2024b). Robust multi-focus image fusion using focus property detection and deep image matting. *Expert Systems with Applications*, *237*(Part A), 121389. https://doi.org/10.1016/j.eswa.2023.121389

Wang, J., Qu, H., Wei, Y., Xie, M., Xu, J., & Zhang, Z. (2022). Multi-focus image fusion based on quad-tree decomposition and edge-weighted focus measure. *Signal Processing*, *198*, 108590. https://doi.org/10.1016/j.sigpro.2022.108590

Wang, Y., Xu, S., Liu, J., Zhao, Z., Zhang, C., & Zhang, J. (2021). MFIF-GAN: a new generative adversarial network for multi-focus image fusion. *Signal Processing: Image Communication*, *96*, 116295. https://doi.org/10.1016/j.image.2021.116295

Wei, C., Zhou, B., & Guo, W. (2018). Multi-focus image fusion based on nonsubsampled compactly supported shearlet transform. *Multimedia Tools and Applications*, *77*(7), 8327-8358. https://doi.org/10.1007/s11042-017-4731-9

Wu, P., Jiang, L., Hua, Z., & Li, J. (2023). Multi-focus image fusion: transformer and shallow feature attention matters. *Displays*, *76*, 102353. https://doi.org/10.1016/j.displa.2022.102353

Wu, S., Wu, W., Yang, X., Lu, L., Liu, K., & Jeon, G. (2019). Multifocus image fusion using random forest and hidden Markov model. *Soft Computing*, *23*(19), 9385-9396. https://doi.org/10.1007/s00500-019-03893-9

You, C.-S., & Yang, S.-Y. (2022). A simple and effective multi-focus image fusion method based on local standard deviations enhanced by the guided filter. *Displays*, *72*, 102146. https://doi.org/10.1016/j.displa.2021.102146

Zhang, H., Le, Z., Shao, Z., Xu, H., & Ma, J. (2021a). MFF-GAN: an unsupervised generative adversarial network with adaptive and gradient joint constraints for multifocus image fusion. *Information Fusion*, *66*, 40-53. https://doi.org/10.1016/j.inffus.2020.08.022

Zhang, Q., Li, G., Cao, Y., & Han, J. (2020a). Multi-focus image fusion based on non-negative sparse representation and patch-level consistency rectification. *Pattern Recognition*, *104*, 107325. https://doi.org/10.1016/j.patcog.2020.107325

Zhang, Q., Wang, F., Luo, Y., & Han, J. (2021b). Exploring a unified low rank representation for multi-focus image fusion. *Pattern Recognition*, *113*, 107752. https://doi.org/10.1016/j.patcog.2020.107752

Zhang, X. (2022). Deep learning-based multi-focus image fusion: a survey and a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 4819-4838. https://doi.org/10.1109/TPAMI.2021.3078906

Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., & Zhang, L. (2020b). IFCNN: a general image fusion framework based on convolutional neural network. *Information Fusion*, *54*, 99-118. https://doi.org/10.1016/j.inffus.2019.07.011