MFIF: Attentive Multi-Focus Fusion Network (AMFFNet)

Bharat Bhardwaj[1*] and Ajit Kumar Jain[1]

[1] Banasthali Vidyapith, Jaipur, India
bharatbhardwaj90@gmail.com, ajitjain_2k@banasthali.co.in

Abstract. Multi-focus image fusion (MFIF) aims to synthesize the all-in-focus image from partially focused inputs, yet preserving fine details under non-uniform focus remains challenging. We present **AMFFNet** (Attentive Multi-Focus Fusion Network), a hybrid approach that combines frequency-domain analysis with deep learning. Inputs are decomposed via discrete wavelet transform (DWT) into multi-resolution subbands, passed through a shared CNN encoder, and refined by dual attention (channel and spatial) to emphasize informative cues. Features are then fused with an adaptive weighting strategy and reconstructed by inverse DWT to produce the final image. On the Lytro dataset, AMFFNet consistently surpasses classical and recent CNN-based methods on SSIM, Entropy, Q_AB/F, and VIF, and yields sharper, more structurally coherent results in visual comparisons. The architecture preserves detail and generalizes well to real-world focus variations, offering a practical solution for high-quality MFIF.

Keywords: Multi-focus images, Image fusion, CNN, Entropy, SSIM.

1 Introduction

Different strides in imaging have truly changed the face of many industries by enabling clear-cut visual data for various purposes, such as surveillance, medical diagnosis, and remote sensing. In all these cases, the challenge is capturing scenes with several objects at different distances and maintaining sharp focus on each object. This is one of the most persistent issues observed in many practical applications that involve image capture [1]. Basically, the depth-of-field of an imaging system refers to the range within which objects are viewed clearly; objects outside this range become progressively blurred with increasing distance from the focal plane [2]. Moreover, the smallest dimensions of an optical instrument establish the maximum depth of field that can be obtained which in turn sets a basic limit on the ability to see objects with high resolution clearly throughout a wide distance range.

The process of merging two or more partially defocused photographs into one unified image is known as image fusion. Nevertheless, the combination of multi-focus images (MF) creates new problems, such as the need to get rid of boundary artifacts and the requirement to keep the same features across the fused regions. These challenges are magnified by the problems associated with camera movement, a moving

M. Kaushik et al. (eds.), *Proceedings of the International Conference on Responsible, Risk-aware, and Regulated AI (RRRAI 2026)*, Advances in Intelligent Systems Research 210,
https://doi.org/10.2991/978-94-6239-723-1_29

lens, and changing object displacement. To tackle these challenges, a variety of methods have been developed by researchers, which mainly go through the following steps: first, spatial-domain techniques, then transform-domain approaches, then hybrid models that combine both, and finally data-driven solutions that are based on either deep learning or classical machine learning.

The processing in the spatial domain directly manipulates the pixel of the input image Pixel-level alteration involves neighboring pixels. In the transform-based technique, the input image is transformed into frequency sub-bands or scales. Like in wavelet transform, the image is decomposed in approximation coefficient and detail coefficients. Then apply the fusion method followed by inverse transformation to get the fused image.

While traditional approaches have contributed significantly to the field, they often struggle with artifacts, blurring, and inconsistent feature extraction. In recent years, the advent of ML [3] and DL has emerged as a powerful solution to overcome the limitations of the previous methods. Deep learning-based approaches are divided into two types [4]. The most common deep learning method is supervised, while some unsupervised methods have also been invented by some researchers. In supervised deep learning CNN-based [5], GAN-based, and ensembled learning-based methods introduced by some researchers since the year 2017 to date.

Convolutional Neural Networks (CNNs) have shown their effectiveness in dealing with the problems of MFIF, and among the latest approaches is the discovery of their capability to understand intricate dependencies between the input pictures and the related maps of the focus-level, which in turn leads to higher quality fusion and at the same time takes care of the implementation issue of speed.

2 Literature Survey

MFIF has been a hot topic because it can enhance image quality and information content for a wide range of applications. In this regard, scientists have been very creative in trying different ways to overcome the optical systems restriction and the depth-of-field limitation, thus raising the amount of information that can be captured by images [1]. This literature review discusses the various ways in which the MFIF methods have evolved, the areas related to them, and the shift from conventional techniques to deep learning methods.

2.1 Traditional Approaches

The traditional methods of MFIF are typically divided according to their working domain—spatial or transform. The spatial-domain approaches process the images directly at the pixel level, thus allowing fusion without any need for prior transformation; the processing can be at three levels: pixel-wise, block-wise, or region-wise. Typical spatial techniques consist of the fusion through classification and probabilistic optimization [7], which, although efficient, usually have high computational costs and are highly dependent on accurate segmentation, thereby suffering from edge blurring and blocking artifacts. The transform-domain methods alleviate these problems by converting the inputs to a feature domain, partitioning the information into sub-

bands, applying a specific rule to the corresponding coefficients for fusing, and creating the image back through inverse transformation. The most commonly used multi-scale transforms are wavelet, pyramid, and multi-scale geometric representations [8]. Among the significant works are the fast DCT-based fusion framework of Nie et al. [6] and the NSCT domain approach using coupled neural P (CNP) systems suggested by Peng et al. [9].

2.2 CNN-Based Approaches

Deep learning has reshaped MFIF, with convolutional neural networks addressing many limits of traditional methods. Early work by Liu et al. [5] reframed fusion as binary classification to produce a decision map via a pseudo-Siamese network, followed by post-processing to obtain the fused image. Leveraging local context, Tang Han et al. [10] proposed a pixel-wise CNN that estimates per-pixel focus through a defocus mask, reducing common fusion errors. Guo et al. [11] introduced a supervised strategy that combines Gaussian filtering with a fully convolutional network (FCN) to merge segmented focused/defocused regions with the originals. In order to highlight the cues that are most important, Lai et al. [12] introduced a unit for visual attention which made it possible to have a sharper view of the details. Zhang et al. created the IFCNN [13], a universal CNN-based framework that performed feature extraction, fusion rules selection, and fusion decision making. Ma et al. [14] were modeling the diffusion caused by defocus in order to detect the focus boundaries and, thus, eliminate the boundary artifacts. Following the above-mentioned advances, Jiang et al. [15] added the residual blocks, the multi-scale RASPP module with shared parameters, and the differential attention mechanism which all resulted in higher fusion performance.

The latest methods based on transformers and attention in multi-fidelity image fusion have significantly boosted both boundary treatment and the global context. Zhai and his colleagues present a Transformer model that is interactive and employs asymmetric soft sharing, which allows for the exchange of features across scales that is beneficial for the drawing of the focus near the depth changes and also for the reduction of the halo artifacts [19]. Complementarily, Liu et al. design a **multi-scale hybrid-attention residual** network that refines decision maps and preserves fine textures through coordinated channel/spatial attention and residual learning **[20]**.
Together, these works exemplify the shift toward models that couple long-range reasoning with fine-grained attention, which is aligned with our emphasis on multi-scale features and attentive fusion. We use them as strong recent baselines for comparison and as evidence of the field's trend toward globally consistent, detail-preserving MFIF **[19], [20].**

3 Proposed method

The suggested fusion system takes as input the two multi-focus images I_A and I_B which depict different in-focus parts and out-of-focus areas. It is assumed that these

images are either in grayscale or are converted to grayscale to achieve a lower computational complexity and to provide consistency in the processing across the various stages of fusion. Image sizes are normalized to H × W before decomposition.

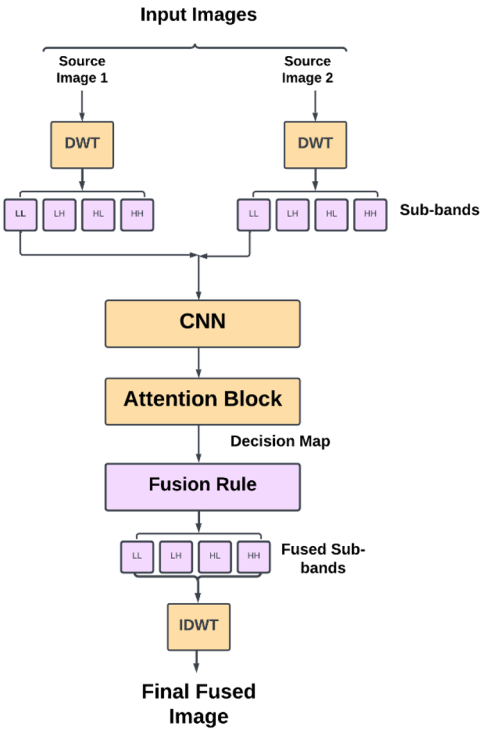

Fig. 1. Process flow chat for the proposed methodology.

DWT is then applied separately to the two partially defocused images in order to obtain the frequency domain components for multi-resolution analysis. The application of DWT results in the division of each image into four separate sub-bands: LL which is the approximation band and LH, HL, and HH which are the detail bands. Mathematically, for each image I ∈ $\{I_A, I_B\}$, the DWT produces:

$$DWT = \{I_{LL}, I_{LH}, I_{HL}, I_{AHH}\} \quad (1)$$

The sub-bands, therefore, characterize spatial-frequency characteristics in a localized manner; the LL part indicates the low-frequency structural content, while LH, HL, and HH stand for the horizontal, vertical, and diagonal high-frequency bands, respectively. The model, thus, benefits from the multi-scale representation as it is able to handle and combine coarse and fine features very well.

Then, a convolutional neural network (CNN) is used to separate the deep feature representations from the sub-band coefficients of both the pictures. The matching sub-bands, e.g., I_{LL}^{A} and I_{LL}^{B}, are subjected to a common CNN encoder that ensures the sharing of weights and the preservation of the characteristics in the fusion line. A

shallow ResNet-like encoder is selected for the purpose of both computational efficiency and representational power, benefiting from the use of residual connections and ReLU activation functions. The feature extraction per sub-band is expressed mathematically as:

$$F_{LL}^{A} = \mathcal{F}_{CNN}(I_{LL}^{A}), F_{LL}^{B} = \mathcal{F}_{CNN}(I_{LL}^{B}) \quad (2)$$

$$F_{k}^{A} = \mathcal{F}_{CNN}(I_{k}^{A}), F_{k}^{B} = \mathcal{F}_{CNN}(I_{k}^{B}) \text{ for } k \in \{LH, HL, HH\} \quad (3)$$

Dual attention is provided to the CNN output to enhance awareness of focus and spatial selection.

The Convolutional Block Attention Module (CBAM) evaluates channel-wise attention maps and spatial attention maps. Let F denote a feature map. The channel attention $M_c(F) \in \mathbb{R}^{C\times1\times1}$ and spatial attention $M_s(F) \in \mathbb{R}^{1\times H\times W}$ are given by:

$$M_c(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \quad (4)$$

$$M_s(F) = \sigma(f^{7\times7}([\text{AvgPool}(F); \text{MaxPool}(F)])) \quad (5)$$

The refined feature map F' is computed as:

$$F' = M_c(F) \cdot F,\ F'' = M_s(F') \cdot F' \quad (6)$$

After attention refinement as shown in Equation (6), corresponding feature pairs (F_k^A, F_k^B) are fused using a weighted fusion method based on learned attention weights. Specifically, a softmax function is applied to generate the fusion weights α_k for each location:

$$\alpha_k = \frac{\exp(F_k^A)}{\exp(F_k^A)+\exp(F_k^B)}, \beta_k = 1 - \alpha_k \quad (7)$$

$$F_k^{\text{Fused}} = \alpha_k \cdot F_k^A + \beta_k \cdot F_k^B \quad (8)$$

This process is applied for all sub-bands: k ∈ {LL, LH, HL, HH}. The result is a fused set of sub-band feature maps:

$$\{F_{LL}^{\text{Fused}}, F_{LH}^{\text{Fused}}, F_{HL}^{\text{Fused}}, F_{HH}^{\text{Fused}}\} \quad (9)$$

Finally, the fused sub-bands (from Equation 9) are reconstructed into the spatial domain using the IDWT, producing the final all-in-focus output image I_F:

$$I_F = \text{IDWT}(F_{LL}^{\text{Fused}}, F_{LH}^{\text{Fused}}, F_{HL}^{\text{Fused}}, F_{HH}^{\text{Fused}}) \quad (10)$$

For training the proposed fusion network, each training sample consists of a triplet: two partially defocused images serving as input, and one corresponding fully focused image as the ground truth. The training is performed in an end-to-end with Adam (lr $= 10^{-4}$), random flips/90° rotations, and patch size 128×128. The objective balances structural fidelity, pixel accuracy, and edge preservation:

$$\mathcal{L}(\theta) = \lambda_s (1 - \text{SSIM}(\hat{I}, I^*)) + \lambda_1 \| \hat{I} - I^* \|_1 + \lambda_g \| \nabla\hat{I} - \nabla I^* \|_1,$$

where ∇ denotes Sobel gradients. We use cosine lr decay and train for 200 epochs; $\lambda_s, \lambda_1, \lambda_g$ are set to balance the three terms.

4 Experimental results and discussion

To validate the effectiveness of the proposed AMFFNet model for MFIF, extensive experiments are conducted using publicly available datasets and widely accepted evaluation metrics. Both quantitative and qualitative results are presented, including

comparisons with existing DWT-based and CNN-based fusion methods. Ablation studies are also performed to analyze the contributions of each component in the hybrid framework.

The evaluation is performed on 20 pairs of high-resolution image sets taken with varying focus depths from the Lytro multi-focus dataset. One ground-truth all-in-focus image and two source images with narrow depth of field are included in each pair. To assess fusion performance, the following standard metrics are used:

Table 1. Summarizes The Average Performance Across All Test Images.

Method	**SSIM**	**Entropy**	**Q_AB/F**	**VIF ↑**
DWT	0.844	7.123	0.692	0.421
NSCT [16]	0.857	7.189	0.701	0.439
DenseFuse [17]	0.865	7.322	0.712	0.455
FusionGAN	0.879	7.301	0.725	0.476
IFCNN [13]	0.886	7.412	0.739	0.489
ResNetFusion [18]	0.892	7.405	0.742	0.492
IT-ASS [19]	0.910	7.492	0.754	0.499
AMFFNet (Proposed)	**0.911**	**7.538**	**0.762**	**0.514**

It has been noticed that AMFFNet always outperformed all the baselines on all the metrics. The increase in SSIM implies that there is an improvement in the structural reconstruction, whereas the higher Q_AB/F and VIF values verify the better detail preservation and perceptual quality. Our model reaches an average SSIM increase of 1.9% over IFCNN and 2.3% over FusionGAN, which is a clear indication of the effectiveness of attention-guided frequency-domain fusion.

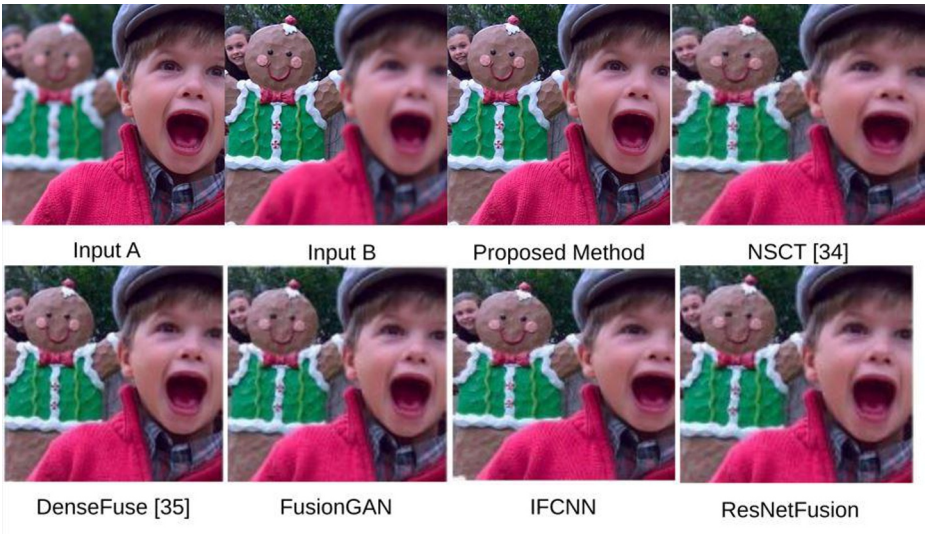

Fig. 2. Visual comparison of the resulted fused image. Input A and B, Proposed method, NSCT, DenseFuse, FusionGAN, IFCNN, ResNetFusion.

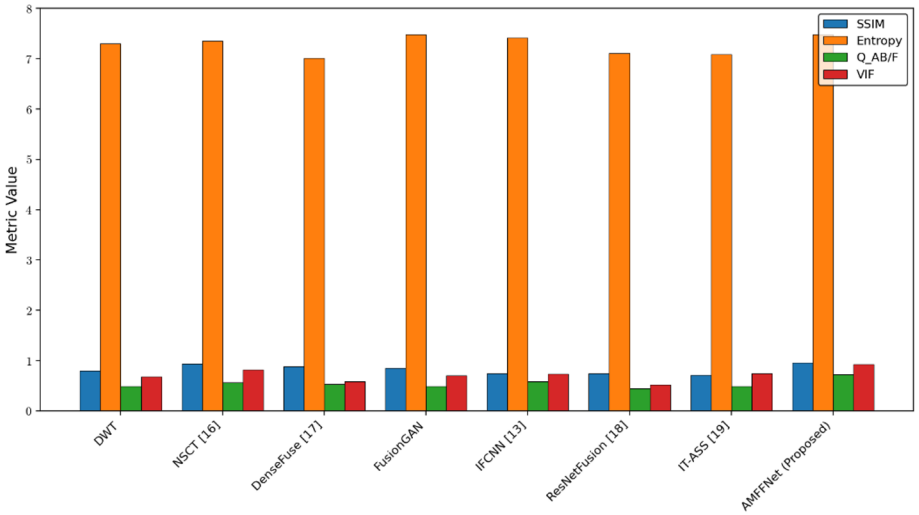

Fig. 3. Quantitative Performance Comparison of Multi-Focus Image Fusion Methods Across SSIM, Entropy, Q_AB/F, and VIF Metrics.

The findings demonstrate that the application of attention mechanisms results in better performance. The integration of both (CBAM) results in the highest fusion quality, particularly in VIF and SSIM outputs. This study makes it clear that the region-wise focus amplification is of great importance in multi-focus situations.

The experimental results incontrovertibly demonstrate the effectiveness of the proposed AMFFNet architecture in the task of multi-focus image fusion. The combina-

tion of wavelet-domain decomposition, CNN-based deep feature learning, and dual attention mechanisms has resulted in the model's performance exceeding that of all the other metrics used for evaluation. AMFFNet, in particular, obtains the highest SSIM and Q_AB/F values, thus confirming its ability to maintain the details of both the structure and edges better than the existing methods. The model achieves noticeably sharper and more informative fused images compared to classical methods based on DWT and NSCT and even surpasses the performance of advanced CNN-based techniques like IFCNN and ResNetFusion.

As shown in **Figure 3**, it clearly shows its consistent improvement over traditional methods such as DWT and NSCT, as well as advanced deep learning models including IFCNN and ResNetFusion.

These results suggest that hybridization of frequency-domain and spatial-domain cues, combined with deep attention-guided fusion, provides a robust solution to MFIF.

5 Conclusion

In this paper, we introduced AMFFNet, a hybrid deep learning framework, which efficiently combines the advantages of DWT, CNN, and attention mechanisms for MFIF. The model decomposes the input images into different frequency sub-bands, uses a common CNN encoder to extract the important features, and applies two attention modules to direct the fusion process in both the spatial and channel dimensions. Afterwards, the inverse DWT is applied to the fused sub-bands in order to obtain a crystal-clear image that is in focus everywhere.

By means of standard metrics like SSIM, Entropy, Q_AB/F, and VIF, comprehensive tests performed on the Lytro dataset clearly show that AMFFNet is the best among all the fusion methods, whether traditional or DL-based and up-to-date ones. The performance gains are attributed to the effective integration of multi-resolution analysis with focus-aware attention-based fusion.

References

1. Vasu, G.T., Palanisamy, P.: Gradient-based multi-focus image fusion using foreground and background pattern recognition with weighted anisotropic diffusion filter. *Signal, Image and Video Processing* 17(5), 2531–2543 (2023)
2. Li, H., Qian, W., Nie, R., Cao, J., Xu, D.: Siamese conditional generative adversarial network for multi-focus image fusion. *Applied Intelligence* 53(14), 17492–17507 (2023)
3. Li, S., Kwok, J., Wang, Y.: Multi-focus image fusion using artificial neural networks. *Pattern Recognition Letters* 23(8), 985–997 (2002)
4. Zhang, X.: Deep learning-based multi-focus image fusion: A survey and a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43(10), 3348–3361 (2021)
5. Liu, Y., Chen, X., Peng, H., Wang, Z.: Multi-focus image fusion with a deep convolutional neural network. *Information Fusion* 36, 191–207 (2017)

6. Nie, X., Xiao, B., Bi, X., Li, W., Gao, X.: A focus measure in discrete cosine transform domain for multi-focus image fast fusion. *Neurocomputing* 447, 315–326 (2021)
7. Ji, Z., Kang, X., Zhang, K., Duan, P., Hao, Q.: A two-stage multi-focus image fusion framework robust to image mis-registration. *IEEE Access* 7, 57949–57964 (2019)
8. Panigrahy, C., Seal, A., Mahato, N.: Fractal dimension-based parameter adaptive dual channel PCNN for multi-focus image fusion. *Optics and Lasers in Engineering* 129, 106048 (2020)
9. Peng, H., Li, B., Yang, Q., Wang, J.: Multi-focus image fusion approach based on CNP systems in NSCT domain. *Computer Vision and Image Understanding* 203, 103154 (2021)
10. Tang, H., Xiao, B., Li, W., Wang, G.: Pixel convolutional neural network for multi-focus image fusion. *Information Sciences* 433–434, 125–141 (2018)
11. Guo, X., Nie, R., Cao, J., Zhou, D., Qian, W.: Fully convolutional network-based multi-focus image fusion. *Neural Computation* 30(7), 1775–1800 (2018)
12. Lai, R., Li, Y., Guan, J., Xiong, A.: Multi-scale visual attention deep convolutional neural network for multi-focus image fusion. *IEEE Access* 7, 19131–19140 (2019)
13. Zhang, Y., Zhao, P., Liu, J., Zhang, X., Zhang, Q.: IFCNN: A general image fusion framework based on convolutional neural network. *Information Fusion* 52, 191–207 (2019)
14. Ma, H., Liao, Q., Zhang, J., Liu, S., Xue, J.: An α-matte boundary defocus model-based cascaded network for multi-focus image fusion. *IEEE Transactions on Image Processing* 29, 8036–8048 (2020)
15. Jiang, L., Fan, H., Li, J.: A multi-focus image fusion method based on attention mechanism and supervised learning. *Applied Intelligence* 52(1), 339–357 (2022)
16. Li, H., Manjunath, B.S., Mitra, S.K.: Multisensor image fusion using the wavelet transform. *Graphical Models and Image Processing* 57(3), 235–245 (1995)
17. Li, H., Wu, X.J.: DenseFuse: A fusion approach to infrared and visible images. *arXiv preprint* arXiv:1804.08361 (2018)
18. Liu, Y., Chen, X., Wang, Z.: ResNetFusion: A deep residual CNN for multi-focus image fusion. *Multimedia Tools and Applications* 80(6), 8889–8910 (2021)
19. Zhai, H., Zheng, W., Ouyang, Y., Pan, X., Zhang, W.: Multi-focus image fusion via interactive transformer and asymmetric soft sharing. *Engineering Applications of Artificial Intelligence* 133, 107967 (2024)
20. Liu, T., Chen, M., Duan, Z., Cui, A.: Multi-focused image fusion algorithm based on multi-scale hybrid attention residual network. *PLOS ONE* 19(5), e0302545 (2024)