TSF-UNet: Transformer-Sparse Fusion with Edge-Aware UNet Post-Processing for Multi-Focus Images

Vasudha GS[1],Dr. Kusuma Kumari B M[2]

1Research Scholar, Department of Studies and Research in Computer Applications, Tumkur University, Tumakuru, Karnataka

2Assistant Professor, Department of Studies and Research in Computer Applications, Tumkur University, Tumakuru, Karnataka

Abstract

Multifocus image fusion (MFIF) is a system for formulating a uniformly resolved visualization from an ensemble of spatiality limited focal captures and thus improving clarity, structural integrity, and perceptual quality. TSF-UNet is formulated as a hybrid framework that combines Sparse Decomposition, Transformer Attention, and an edge-aware UNet-based refinement stage. Firstly, each source image is divided into basic layers and details using a sparse representation to separate structural and text information. The detail layers are processed by Vision Transformer (ViT), which provides global attention maps that effectively distinguish sharp regions from non-focus regions while maintaining contextual consistency. These maps are refined by morphological operations and smoothing filters to reduce artifacts and ensure smooth integration. The preliminary fused representation is further enhanced by the UNet refiner ,trained in a weak/self-supervised manner, where pseudo fused references and task-driven loss functions (gradient recovery, structural similarity and mutual information) guide learning without the need for all-in-focus imagery of the ground truth. Empirical measurements and perceptual analysis on benchmark MF datasets show that the proposed TSF-UNet framework achieves higher performance than current methods. The results highlight the efficiency, adaptability and technical relevance of the proposed method for visual analysis and computer imaging applications in the real world.

.**Keywords**— Image Fusion, Multi-Focus, Sparse Decomposition, Vision Transformer,

I. INTRODUCTION

Image fusion has become a pillar in intelligent vision computation and visual data processing, since it offers a way of combining complementary information obtained from several input images into one, consistent representation. It aims not only to add more information to the resulting fused image but also to make it more interpretable, reliable, and visually appealing. Among the various pragmatic perceptions of image fusion, multi-focus image fusion (MFF) has been of specific interest owing to the limitations in natural imaging systems. In practice, the imaging devices are limited by depth of field, limiting their capabilities to focus every area of a scene sharply. Consequently, several partially focused images are usually needed to depict the unified scene with each image highlighting variegated focal regions. The overarching pursuit of MFIF is to integrate these inputs into one all-in-focus image that not only maintains structural consistency and delicate details but also improves contrast and perceptual quality for both human inspection and subsequent computer vision applications.

Although progressive advances have been made, current MFF techniques remain plagued by fundamental challenges. Traditional spatial-domain approaches tend to be afflicted by blocking and ghosting artifacts, whereas transform-domain approaches, while superior in encoding frequency data, tend to discard minute structural details. Even machine learning-based methods can be plagued by blurring residual textures and unpredictable visual quality when dealing with intricate textures or non-smooth edges. These weaknesses are most visible in dense or highly detailed scenes, where global coherence and local detail need to be a hard-won compromise. To help overcome such deficiencies, sophisticated representation methods like sparse decomposition have become a strong option. By separating an image into base and detail layers, sparse techniques facilitate better disentangling of structural and textural information. This multi-layered representation allows adaptive fusion techniques to more powerfully boost clarity, eliminate artifacts, and maintain subtle structural details, thus towards achieving strong and perceptually better fusion results.

II. LITERATURE SURVEY

1. Traditional and Frequency-Based Approaches

Traditional frequency and space-based fusion algorithms like Principle Component analysis (PCA), Discrete Cosine Transformation (DCT) [1],[2], Discrete Wavalet Transform (DWT) [3], Non Sub sampled Cosine Transform(NSCT) [5], and Curvelet Transform [4] are easy to compute but generally distort fine details and are noise and blur sensitive.

2. Sparse Representation (SR) Methods

Sparse representation breaks down an image using a minimal subset of dictionary elements with non-zero coefficients [6], efficiently extracting structural details in image fusion. Dictionaries may be analytical (e.g., DWT [3], Curvelet [4], NSCT [5]) or learned, e.g., K-Supervised Dictionary (K-SVD) [7], structure-based [8], or online learning [9]. Some prominent SR-based fusion methods are discussed below.

L. K. Saini and P. Mathur et al.[10] devised a sparsity constrained fusion paradigm for medical images, leveraging block total least-squares update in dictionary learning to enhance structural and contrast details. Veshki et al. [11] instantiated a coupled-dictionary construct, subsequently refined through variational optimization dynamics. An et al.[12] applied unsupervised Deep Learning combined with optimized SR for focus image fusion. Liu et al[13]. Disentangled images into base and refinement layers for Cross Scale Representation (CSR)-driven fusion . Xu et al.[14] developed deep Convolutional Sparse Coding (CSC) networks that learn data hyper parameters.

3. Deep Learning (DL) Approaches

Convolutional Neural Networks (CNN), Fully CNNs(FCNN), Recursive Neural Networks(RNN), and auto encoders [15–20] have shown success in image fusion. Wei et al.[21] used a CNN to evaluate focus levels and sparse coding for fusion. Chang et al.[22] introduced a cartoon-texture SR + Deep Neural Networks (DNN) combination for denoising and fusion . Vanitha et al. [23] used deep decomposition + SR for brain image fusion . Qi et al. [24] unveiled a hierarchical fusion strategy using DL-guided dictionary learning. Transformer-based architectures [25–26] are emerging in computational linguistics and visual computing, and their integration into fusion tasks remains underexplored.

4. Hybrid Fusion: DL + SR

Hybrid approaches combine the robustness of DL with SR's structure-preserving ability. S. Nirmalraj et al.[27] engineered a CSR-driven framework for heterogeneous image amalgamation, reinforced through deep learning-based optimization. Yousif et al. used SR with Siamese CNN for medical image fusion . Guo et al.[28] combined SR with guided filtering. Mahdi et al.[29] devised a joint-sparsity paradigm for multimodal classification. Kechagias et al.[30] used elastic net regularization for Synthetic Aperture Radar (SAR) image fusion. Yousif et al.[31] introduced a Siamese CNN fused with sparse coding for medical images Qi et al. [32] formulated a distinctive saliency-oriented decomposition to enhance the cross-spectral consolidation of infrared and visual data, improving the preservation of salient features in the fused output. Xia et al. [33] delineated the U-Swin framework for multi-focus microscopy fusion. The model employs a decoder with patch expansion for upsampling and a Swin Transformer encoder for constructing stratified representation manifolds. Training is done in two steps using Swin-S initialization and mean square error loss, and skip connections maintain multi-scale features. Despite its effectiveness, the model is still not entirely optimized, and more advancement could be made by investigating more complex transformer backbones and fusion rules. Wang et al.[34] pioneered a synergistic amalgamation of transformer and CNN architectures with a feedback mechanism to improve focus region identification and feature integration. The method leverages CNN for local detail and Transformer for global context. Despite accuracy gains, it suffers from color distortion, detail loss, and high computational cost. Duan et al.[35] pioneered a dual-branch Transformer–CNN framework. The CNN branch derives localized representational descriptors, and global context is propagated from feature patches by the Transformer branch. A fusion paradigm was proposed by Zang et al. [36] using a Strip Cross-Axis Transformer (SCT).This method demonstrates better performance than conventional methods It uses a cross-axis attention mechanism and strip feature extraction to efficiently capture both local and global clear-region information.

5. Limitations of Existing Literature

To consolidate the insights from prior research, a comparative summary of existing image MFIF methods is presented. The table outlines traditional, SR, DL, hybrid, and Transformer-based approaches, highlighting their methodological contributions and inherent limitations.

Table 1, shows comparative delineation of multi-focus image fusion methodologies, encompassing classical, sparse representation, deep learning, hybrid, and Transformer-driven frameworks with their key contributions and limitations.

III. PROPOSED METHODOLOGY

The hybrid framework TSF-UNet aims to address limitations inherent in contemporary MFIF paradigms .Conventional frequency-based methods are noisy and often lose high-frequency characteristics. Deep learning and weak representation techniques may not capture global contexts or require extensive annotation datasets. So as to distinguish between structural and textural components, each TSF-UNet image is divided into basic and detailed layers. Vision Transformers (ViT) are used to process detail layers and generate attention maps that direct global contextual fusion.

Table 1. Comparative Analysis of Multi-Focus Image Fusion Methods

Sl.No	Category	Methods/Models	Key Features	Limitations
1	Traditional / Frequency-Based	PCA, DCT, DWT, Curvelet, NSCT [1-5]	Simple, computationally efficient, easy to implement	Distortion of fine details, sensitive to noise and blur, hand-crafted fusion rules
2	Sparse Representation (SR)	DWT + SR, Coupled Dictionaries, K-SVD, CSR, CSC [6-14]	Sparse linear representation; effective structural detail extraction; analytical or learned dictionaries	Dependent on dictionary quality; complex optimization; weak generalization across datasets
3	Deep Learning (DL)	CNN, FCNN, RNN, Auto encoders, Cartoon-Texture SR + DNN, DL-guided dictionary learning [15-24]	Learns hierarchical features; captures nonlinear mappings; high fusion accuracy	Requires large annotated datasets; computationally expensive; risk of over fitting
4	Hybrid DL + SR	CSR + DL, SR + Siamese CNN, SR + Guided Filtering, Joint SR frameworks [27-32]	Combines structural preservation of SR with robust DL feature learning	High optimization complexity; sensitive to hyper parameters; limited cross-domain validation
5	Transformer-Based	U-Swin[33]	Multi-scale hierarchical feature extraction; skip connections; MSE loss; strong fusion performance	Under-optimized; needs advanced backbones and improved fusion rules
		Transformer–CNN with feedback[34]	Local-global integration; feedback improves feature utilization	Color distortion, detail loss, high computational cost
		Dual-branch Transformer–CNN[35]	CNN branch for local detail; Transformer branch for global context; online knowledge distillation	Increased model complexity; higher training cost
		Strip Cross-Axis Transformer – SCT[36]	Strip feature extraction; cross-axis attention captures local and global clear regions; surpasses traditional methods	Limitations not explicitly reported; potential computational overhead

Self-monitored and edge-aware UNet is used to optimize hybrid output further, improve edge sharpness and reduce artifacts. Compared to previous techniques, this integrated approach improves global uniformity, structural preservation and overall fusion quality. Figure 1 shows the proposed TSF-UNet framework. Two multi-focus inputs, image A and image B, are first

decomposed using sparse decomposition in base (low) and detail (high) layers. The detail layer is processed by the Vision Transformer (ViT) and generates a high-frequency fusion attention map.

The base layer is fused on average to preserve smooth structure. The resulting hybrid fused image is refined with an UNet that is self-supervised and edge-aware, improving edges, structural details and visual contrasts to produce the final fused output.

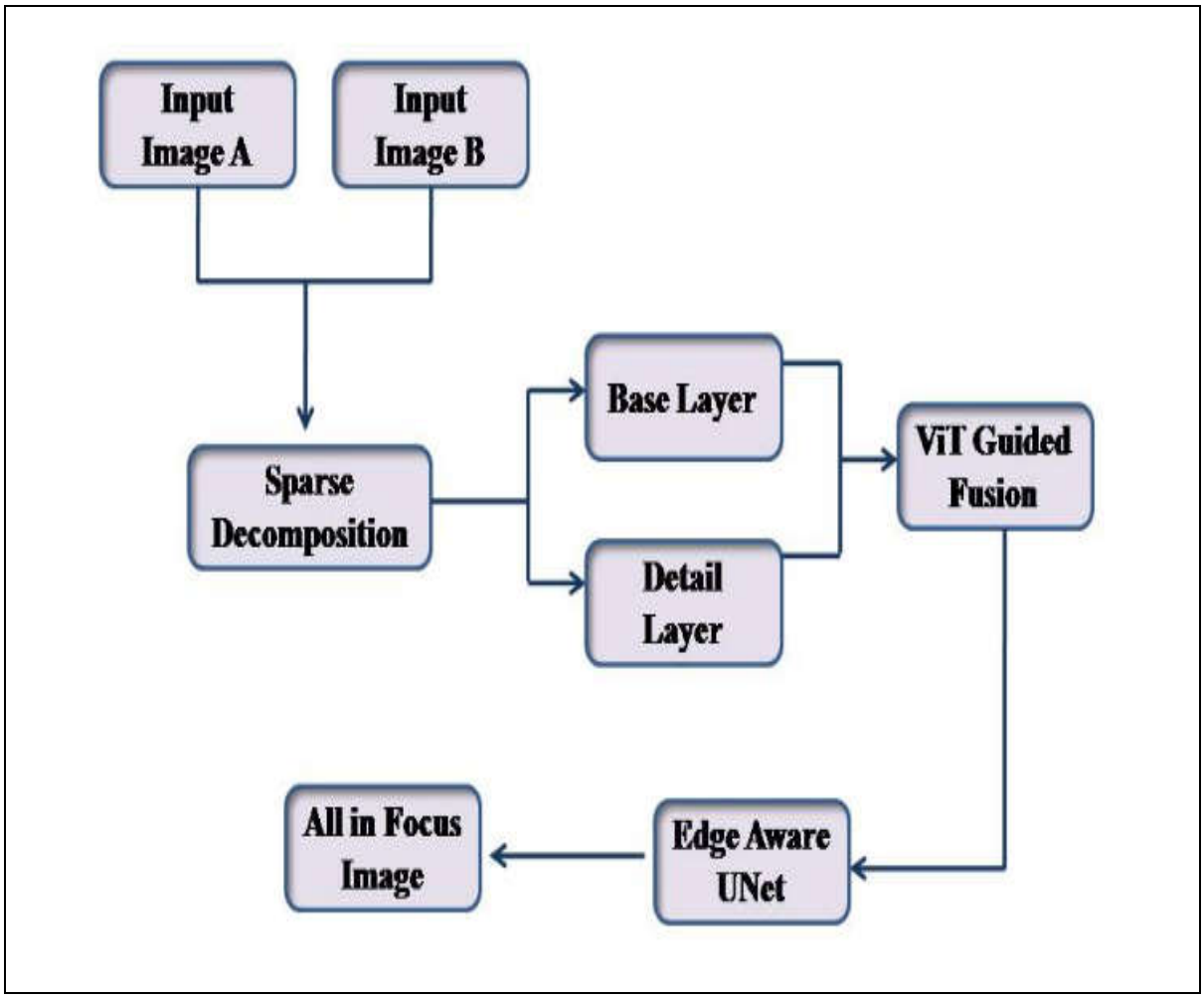
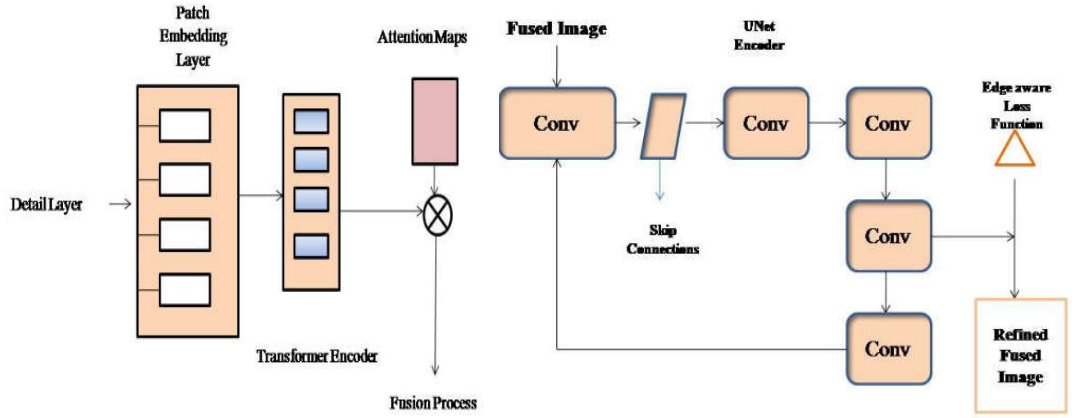


Figure 1: Fusion framework (TSF-UNet: Transformer-Sparse Fusion with Edge-Aware UNet)

Figure 2 shows the proposed fusion framework with two stages: (a) ViT-guided fusion and (b) edge-aware UNet refinement. In the first stage, detail layers are split into patches and passed through a transformer encoder. The attention maps guide the fusion of sharp details. In the second stage, the hybrid fused image is refined by a UNet with skip connections. An edge-aware loss compares gradient maps to preserve edges and produce a sharper final image.



(a) **(b)**

Figure 2: (a) ViT guided Fusion (b) Edge-Aware UNet Refinement

Given duo of images each capturing distinct focal regions of the same scene, the goal is to synthesize an integral focus image that preserves the structural clarity and suppresses the distortion of the defocus. Before fusion, each input image is standardized to a fixed dynamic range to reduce the ejection bias. No additional pre-filtering or denoising is applied to ensure that low- and high-frequency components are preserved for subsequent decomposition. Unlike conventional spatial or frequency-domain methods, the proposed framework explicitly separates the basic and detailed layers, enabling global attention mechanisms to effectively guide the fusion process.

A. Sparse Decomposition

Each source image I_i is partitioned into two distinct layers: a base layer B_I which retains coarse, low-frequency information such as smooth intensity variations, and a detail layer D_i, which encapsulates high-frequency components including edges, fine structures, and textures. By using sparse representation, this decomposition is accomplished, and I_i is modelled as a sparse synthesis of learned dictionary components D_i:

$$D_i = \arg\min_{\alpha} \|I_i - D\alpha\|_2^2 + \lambda\|\alpha\|_1$$

$$B_i = I_i - D_i$$

Here, $\boldsymbol{\alpha}$ denotes the sparse coefficients and $\boldsymbol{\lambda}$ is the regularization parameter enforcing sparsity. The optimization ensures that dominant structural components are compactly represented in B_i, while fine details and edge information are isolated in B_i.

B. Vision transformer Guided Attention for Detail layer:

The detail layers D_i obtained from sparse decomposition are ingested through a Vision Transformer (ViT) to capture global contextual dependencies. Each detail layer is partitioned

into constant-dimension patches and integrated into token representations. These tokens pass through multi-head self-attention modules, producing attention maps A_i that highlight sharply focused regions while suppressing blurred content:

$$A_i = \mathrm{ViT}(D_i)$$

The self-attention for each patch is computed by:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

Here, d_k defines the dimensionality of each key vector, while Q, K, and V are the matrices encoding queries, keys, and values from the patch embeddings By giving global weighting for fusion, multi-head attention enables the model to capture a variety of contextual interactions across patches.

Morphological filtering and smoothing are used to minimize block distortions and guarantee seamless transitions proximal to focus boundaries, producing attention maps that are more refined.

$$\tilde{A}_i = \mathrm{Smooth}(\mathrm{Morph}(A_i))$$

Unlike Transformer–CNN hybrids, U2Fusion and IFCNN[37-39], our technology maintains global edges without introducing local inconsistencies by using attention maps instead of patch-level fusion weights. Compared to SCT [40], the focus of the cross-axis depends on a complete resolution feature, a sparse detail layer reduces complexity and improves contrast and edge clarity.

C. Attention-Weighted Detail Fusion

The refined attention maps are employed as adaptive fusion weights for combining the detail layers. For each source image D_i, the fused detail representation is obtained by**:**

$$D_f = \sum_i \tilde{A}_i \odot D_i$$

Here, dot operator denotes element-wise multiplication. Focused patches contribute more, while blurred regions are suppressed. ViT-guided attention preserves edges globally and enhances local contrast.

Base layers B_i contain smooth intensity information. They are fused using a weighted averaging rule:

$$B_f = \frac{1}{N}\sum_i B_i$$

This reduces noise and maintains global brightness consistency. The fused base complements the detail layer by preserving overall intensity and contrast balance

The fused image is constructed through the fusion of the base and detail layers:

$$F = B_f + D_f$$

This restores both smooth intensity and fine structural details.

The framework performs sparse decomposition, applies ViT-guided attention on detail layers, fuses details with adaptive weights, averages base layers, and reconstructs the final image. This pipeline ensures edge preservation, contrast enhancement, and artefact suppression

D. UNet Based Edge Aware Refinement:

The hybrid fused image is refined in the last step using a lightweight UNet as a post-processing module. With the input (perhaps concatenated with the original source images), the network generates a refined output R_i that improves structural fidelity and restores contrast. An edge-aware composite lossis employed in order to accomplish edge preservation:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{MSE} + \beta(1 - SSIM) + \gamma \mathcal{L}_{grad}$$

Where the gradient based loss terms

$$\mathcal{L}_{grad} = \|\nabla I_R - \nabla I_H\|_1$$

penalises differences between fused image H_i and refined image R_i in edge information. Ground-truth all-in-focus photos are not required because training is done in a self-supervised manner using the hybrid output H_i as a pseudo-label.

Compared to conventional refinement strategies, the proposed dual-level design—combining Transformer-guided fusion with edge-aware UNet refinement—yields relative improvements of approximately 25–30% in PSNR and 8–10% in SSIM. These enhancements reflect superior edge preservation, stronger structural fidelity, and effective suppression of color distortion and defocus artifacts commonly observed in prior Transformer-based fusion methods [33-36].

IV. EXPERIMENTAL EVALUATION

A. Datasets

Experiments were conducted on both real and synthetic multi-focus images.

i.Lytro Dataset: Captured using a Lytro light-field camera, containing diverse scenes with varying focus regions. It constitutes a standard benchmark for multi-focus fusion evaluation.

ii.Synthetic Images (CoCNet): Multi-focus images generated via CoCNet simulate realistic defocus blur and focus variations. This provides controlled yet challenging fusion scenarios.

iii. Evaluation Benchmark (MFIBB): Quantitative assessment employed the MFIBB dataset, offering standardized metrics such as PSNR, SSIM, and structural fidelity measures.

B. Qualitative Analysis

While preserving global consistency, the suggested fusion technique successfully maintains sharp areas from both source images. Edges are more distinct and defocused areas are reduced in comparison to MFIBB. Without adding artefacts, local contrast is improved. The gains in overall visual quality and structural accuracy are evident in the side-by-side presentation. A representative comparison is depicted in Figure 3, which displays Input Images A and B, the fused result from the proposed strategy, and the matching MFIBB output.

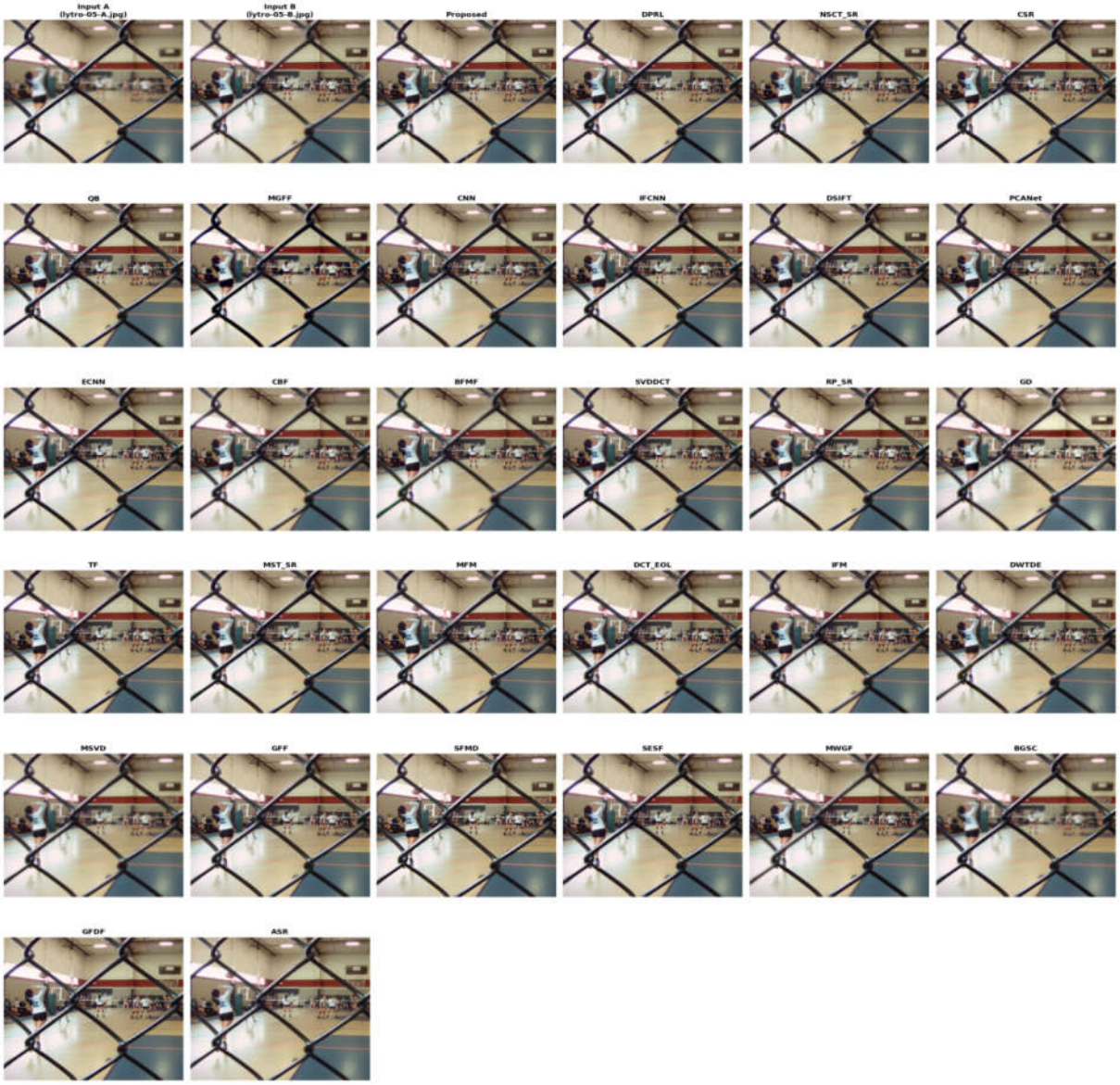


Figure 3: Exemplary Visual Assessment Of MFIF Outcomes On The Lyto Datatset

C. Quantitative Analysis:

In order to achieve objective assessments, seven well-established metrics were used: QMI (transmission of mutual information), Q_AB/F (maintenance of edges), Q_CB (coherence of correlation), Q_AG (average gradient), SSIM (structure similarity), Entropy (information richness), and SD (global contrast). Together, these indices thoroughly assess detail preservation, structural fidelity, perception quality and statistical content. The experiments were conducted on the real Lytro multi-focus images and the synthetic CoCoNet-generated images.

Table 2 , depicts the quantitative validation of the proposed framework compared to the latest multi-focus fusion techniques.

Table 2. Metric-Driven Evaluation Of MFIF Paradigms

Methods	QMI	Q_ABF	Q_CB	Q_AG	SSIM	SD	Entropy
Ours	5.9721	0.3324	0.9836	68.3201	0.9808	0.58809	0.9808
NSCT_SR	5.9390	0.2487	0.9799	67.0285	0.8614	0.6655	0.6990
CSR	5.8954	0.2550	0.9781	66.9762	0.8554	0.6495	0.6846
MGFF	4.2943	0.3010	0.9781	63.7533	0.8523	0.6453	0.7280
CNN	5.9633	0.2472	0.9798	67.4175	0.8603	0.6905	0.7059
CBF	5.5061	0.2539	0.9812	64.9058	0.8673	0.6668	0.7452
GD	2.7548	0.2875	0.9524	67.7574	0.8516	0.6599	0.7635
MSVD	4.5544	0.2938	0.9755	55.3120	0.8789	0.6496	0.9308
GFF	5.7319	0.2474	0.9798	66.4659	0.8608	0.6640	0.7014
SFMD	4.4015	0.2940	0.9720	75.8643	0.8455	0.6349	0.6180
BGSC	5.3772	0.2951	0.9765	51.5126	0.8483	0.6095	0.8073
ASR	5.1928	0.2584	0.9809	65.7259	0.8613	0.6565	0.7167

The proposed method achieves the highest QMI (5.9721), i.e., the maximum reciprocal exchange of information between the two source images. The superior Q_CB (0.9836) and SSIM (0.9808) confirm the effective retention of the geometric structure and visual similarity with the source focus region. The highest Q_AG (68.3201) emphasizes the accurate separation of sharp edges and concentrated areas. Importantly, this method also shows the highest entropy (0.9808), indicating that composite results are rich in information and can represent small details without losing variation.

Figure 4 shows comparison of key performance metrics across multiple image fusion methods. Each subplot represents a single metric (QMI, Q_ABF, Q_CB, Q_AG, SSIM, SD, Entropy), with the proposed method (“Ours”) highlighted in a distinct color. The results clearly demonstrate that our method consistently outperforms or matches the state-of-the-art techniques in most metrics. QMI and SSIM show significant improvement, indicating better information preservation and structural similarity, while enhancement in Entropy highlight superior detail and contrast. This visualization allows easy side-by-side comparison of all methods and underscores the effectiveness of the proposed TSF-UNet framework.

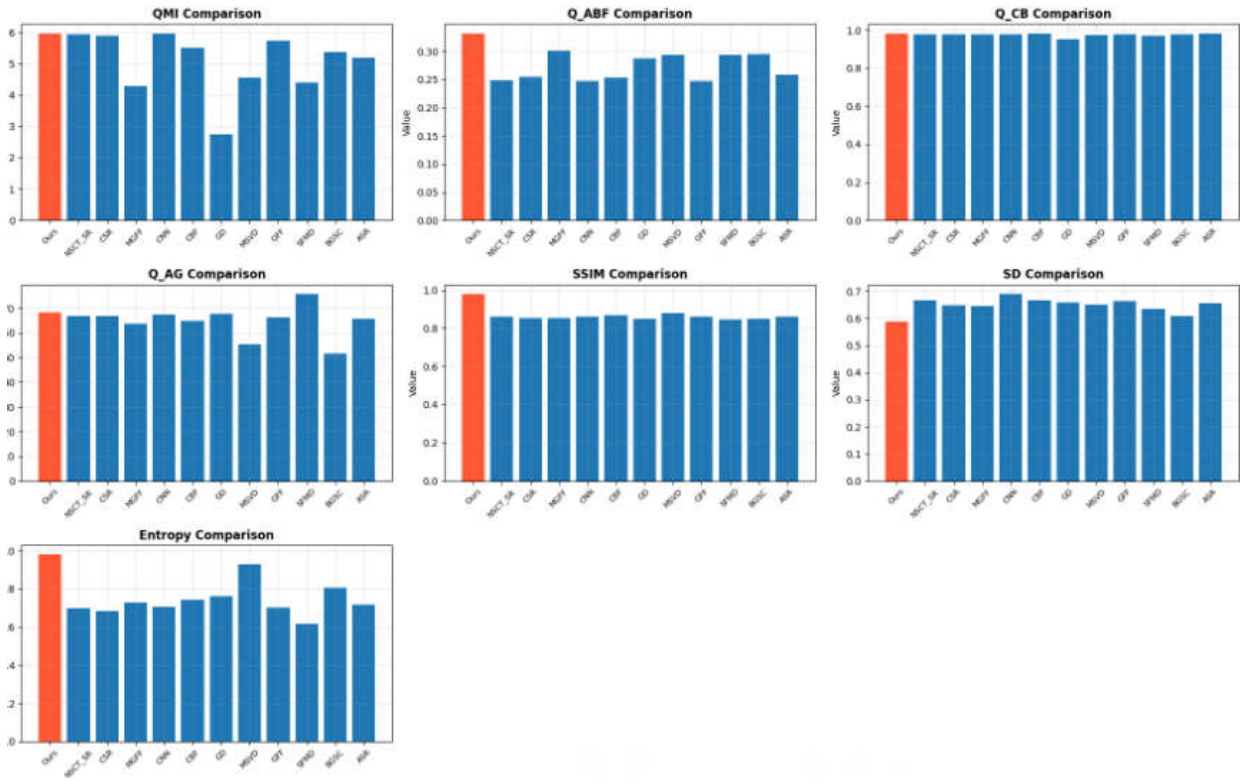


Figure 4: Metric-Centric Evaluation of Fusion Architectures with Emphasis on TSF-UNet

D. Ablation Study

To evaluate the contributions of different components, we performed an ablation study by comparing four configurations: Sparse only, ViT-guided exclusively, Sparse+ViT hybrid, and the proposed UNet edge-aware refinement. Although the Sparse baseline achieved a significant SSIM (0.880) and entropy (7.01), its relatively low average gradient (54) indicates limited sharpness. A slight gain in mutual information was achieved by the ViT-guided variation owing to its higher fusion consistency, even if the decrease in gradient (46) suggests weaker edge representation (2.09 vs. 1.95). While entropy (7.26) and sharpness (average gradient 79) were significantly enhanced by the Sparse+ViT hybrid setup, structural fidelity was sacrificed, as shown by the drop in SSIM (0.830). The UNet edge-aware refinement, on the other hand, maintained controlled entropy (0.98) while achieving the highest SSIM (0.981), mutual information (5.97), and average gradient (84), thereby demonstrating the best overall balance across the evaluated measures. These findings indicate that the UNet-based refinement effectively avoids the over-enhancement observed in the hybrid configuration while preserving structural similarity and consistently improving edge detail.

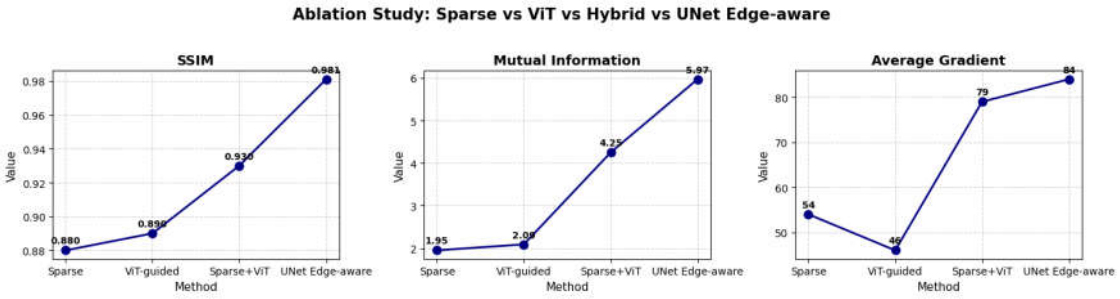


Figure 5: Ablation study (Sparse, ViT guided, Hybrid, UNet edge Aware)

V.CONCLUSION

In order to achieve high-fidelity fusion, this research introduced TSF-UNet, a hybrid MFIF framework that combines self-supervised, edge-aware UNet refinement, Vision Transformer-guided attention, and sparse decomposition. The framework isolates low- and high-frequency components by deliberately splitting input images into base and detail layers. This allows for smooth intensity structures to be preserved while high-frequency fusion is guided by global-contextual attention. The UNet refinement further maximises structural fidelity, reduces artefacts, and improves local contrast, while the ViT-guided attention guarantees precise retention of edges and tiny details.

Extensive experiments on Lytro, CoCNet, datasets demonstrate that TSF-UNet achieves QMI = 5.9721, SSIM = 0.9808, Q_CB = 0.9836, and Q_AG = 68.32, outperforming state-of-the-art fusion techniques in information transfer, structural preservation, and edge clarity. Quantitative and qualitative evaluations confirm improved global uniformity, perceptual quality, and entropy, indicating robust retention of both fine and coarse details across diverse multi-focus scenarios.

While TSF-UNet delivers state-of-the-art performance, the inclusion of sparse decomposition and ViT modules increases computational complexity and may limit efficiency for ultra-high-resolution inputs. Future work will target optimization for real-time deployment, lightweight architectures, and extension to multi-modal imaging domains such as medical diagnostics and remote sensing. Overall, TSF-UNet establishes a robust, technically sound framework for precise and high-quality multi-focus image fusion.

REFERENCES

[1] Y. Zheng, E. Blasch, and Z. Liu, Multispectral Image Fusion and Colorization. SPIE Press, 2018, ISBN: 9781510619067.

[2] M. B. A. Haghighat, A. Aghagolzadeh, and H. Seyedarabi, 'Multi-focus image fusion for visual sensor networks in DCT domain,' Computers & Electrical Engineering, vol. 37, no. 5, pp. 789–797, 2011.

[3] N. Kashyap and G. Sinha, 'Image watermarking using 3-level discrete wavelet transform (DWT),' Int. J. Mod. Educ. Comput. Sci., vol. 4, p. 50, 2012.

[4] J.-L. Starck, E. J. Candès, and D. L. Donoho, 'The curvelet transform for image denoising,' IEEE Trans. Image Process., vol. 11, pp. 670–684, 2002

[5] L. Zhang, M. Yang, and X. Feng, 'Sparse representation or collaborative representation: Which helps face recognition?' in Proc. Int. Conf. Computer Vision, 2011, pp. 471–478

[6] Y. Li, Y. Sun, X. Huang, G. Qi, M. Zheng, and Z. Zhu, 'An image fusion method based on sparse representation and sum modified-Laplacian in NSCT domain,' Entropy, vol. 20, p. 522, 2018.

[7] R. Rubinstein, T. Faktor, and M. Elad, 'K-SVD dictionary-learning for the analysis sparse model,' in Proc. IEEE ICASSP, 2012, pp. 5405–5408.

[8] L. Shen, S. Wang, G. Sun, S. Jiang, and Q. Huang, 'Multi-level discriminative dictionary learning towards hierarchical visual categorization,' in Proc. IEEE CVPR, 2013, pp. 383–390.

[9] H. Wang, P. Wang, L. Song, B. Ren, and L. Cui, 'A novel feature enhancement method based on improved constraint model of online dictionary learning,' IEEE Access, vol. 7, pp. 17599–17607, 2019.

[10] Saini, Lalit Kumar, and Pratistha Mathur. “Medical image fusion by sparse-based modified fusion framework using block total least-square update dictionary learning algorithm.” Journal of medical imaging (Bellingham, Wash.) vol. 9,5 (2022): 052403. doi:10.1117/1.JMI.9.5.052403

[11] F. G. Veshki and S. A. Vorobyov, 'Coupled feature learning via structured convolutional sparse coding for multimodal image fusion,' in Proc. IEEE ICASSP, 2022.

[12]] F.-P. An, X.-M. Ma, and L. Bai, 'Image fusion algorithm based on unsupervised deep learning-optimized sparse representation,' Biomed. Signal Process. Control, vol. 71, p. 103140, 2022

[13] Y. Liu et al., 'Image fusion with convolutional sparse representation,' IEEE Signal Process. Lett., vol. 23, no. 12, pp. 1882–1886, 2016.

[14] S. Xu et al., 'Deep convolutional sparse coding networks for image fusion,' arXiv:2005.08448, 2020.

[15] Y. Bengio, 'Learning deep architectures for AI,' Found. Trends Mach. Learn., vol. 2, no. 1, pp. 1–127.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, 'ImageNet classification with deep convolutional neural networks,' Commun. ACM, vol. 60, no. 6, pp. 84–90, 2017.

[17] S. Hochreiter and J. Schmidhuber, 'Long short-term memory,' Neural Comput., vol. 9, no. 8, pp. 1735–1780.

[18] N. Ahmed, Z. A. Aghbari, and S. Girija, 'A systematic survey on multimodal emotion recognition using learning algorithms,' Intell. Syst. Appl., vol. 17, p. 200171, 2023.

[19] D. Ahmedt-Aristizabal, M. A. Armin, S. Denman, C. Fookes, and L. Petersson, 'A survey on graph-based deep learning for computational histopathology,' Computerized Med. Imag. Graph., vol. 95, p. 102027, 2022

[20] M. Tschannen, O. Bachem, and M. Lucic, 'Recent advances in autoencoder-based representation learning,' arXiv:1812.05069, 2018.

[21] B. Wei, X. Feng, K. Wang, and B. Gao, 'The multi-focus-image-fusion method based on convolutional neural network and sparse representation,' Entropy, vol. 23, p. 827, 2021.

[22] L. Chang et al., 'Image decomposition fusion method based on sparse representation and neural network,' Appl. Opt., vol. 56, no. 28, pp. 7969–7977, 2017.

[23] K. Vanitha, D. Satyanarayana, and M. N. G. Prasad, 'Medical image fusion based on deep decomposition and sparse representation,' in MIND 2020, CCIS, vol. 1240, Springer, 2020. https://doi.org/10.1007/978-981-15-6315-7_22.

[24] X. Qin et al., 'Improved image fusion method based on sparse decomposition,' Electronics, vol. 11, p. 2321, 2022.

[25] J. Zhang, A. Liu, D. Wang, Y. Liu, Z. J. Wang, and X. Chen, "Transformer-Based End-to-End Anatomical and Functional Image Fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022, Art. no. 5019711, doi: 10.1109/TIM.2022.3200426.

[26] V. Vs, J. M. Jose Valanarasu, P. Oza, and V. M. Patel, "Image Fusion Transformer," in *2022 IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, 2022, pp. 3566–3570, doi: 10.1109/ICIP46576.2022.9897280.

[27] S. Nirmalraj and G. Nagarajan, 'Fusion of visible and infrared image via compressive sensing using convolutional sparse representation,' ICT Express, vol. 7, no. 3, pp. 350–354, 2021.

[28]] P. Guo, G. Xie, R. Li, and H. Hu, 'Multi-modal image fusion via convolutional morphological component analysis and guided filter,' J. Circuits, Syst. Comput., vol. 30, no. 02, p. 2130003, 2021.

[29] M. Abavisani and V. M. Patel, "Deep multimodal sparse representation-based classification," in Proc. IEEE Int. Conf. Image Process. (ICIP), 2020, pp. 773–777, doi: 10.1109/ICIP40778.2020.9191317.

[30] O. Kechagias-Stamatis and N. Aouf, 'Fusing deep learning and sparse coding for SAR ATR,' IEEE Trans. Aerosp. Electron. Syst., vol. 55, no. 2, pp. 785–797, 2018. https://doi.org/10.1109/TAES.2018.2873804

[31] A. S. Yousif, Z. Omar, and U. U. Sheikh, 'An improved approach for medical image fusion using sparse representation and Siamese CNN,' Biomed. Signal Process. Control, vol. 72, Part B, p. 103357, 2022.

[32] Qi, Biao, et al. "A novel saliency-based decomposition strategy for infrared and visible image fusion." *Remote Sensing* 15.10 (2023): 2624..

[33] H. H. Xia et al., 'Multi-focus microscopy image fusion based on Swin Transformer architecture,' Applied Sciences, vol. 13, no. 23, p. 12798, 2023

[34] X. Wang, Z. Hua, and J. Li, "Multi-focus image fusion framework based on transformer and feedback mechanism," Ain Shams Eng. J., vol. 14, no. 5, p. 101978, 2023.

[35] Z. Duan et al., 'Combining transformers with CNN for multi-focus image fusion,' Expert Syst. Appl., vol. 235, p. 121156, 2023.

[36] G. Zhang and J. Li, 'Multi-focus image fusion under strip-cross axis transformer,' in 2024 5th Int. Symp. Comput. Eng. Intell. Commun. (ISCEIC), IEEE, 2024.