Bi-CMFM: A Multimodal Image Fusion Method Based on Bi-Path Residual and Cross-Modal Feature Merging

Yang Jingdong[1], Zhang Wendong[2(✉)], Liao Shengnan[1], Zhou Zhentao[1], Zhang Hongning[1], Teng Xiaoxi[1] and Liu Tao[1]

[1] College of Software, Xinjiang University, Urumqi 830091, China
[2] College of Computer Science and Technology, Xinjiang University, Urumqi 830046, China
zwd@xju.edu.cn

Abstract. Multimodal medical image fusion integrates complementary information from CT, MRI, and PET to support clinical diagnosis and lesion localization. Three fundamental challenges remain unresolved in existing methods: (1) global–local imbalance, where CNNs' local receptive fields limit long-range dependency modeling, causing blurred edges, while Transformers incur quadratic computational complexity; (2) insufficient inter-modal redundancy suppression, producing structural inconsistency and detail loss; and (3) the efficiency–quality trade-off, which limits clinical deployment of high-performing models. To address these three challenges, Bi-CMFM is proposed, comprising the Bi-Path Residual Fusion module (BPRF), which employs parallel standard and dilated convolutions to preserve local texture while expanding the receptive field, and the Cross-Modal Fusion Module (CMFM), which applies a Cross-modal Feature Enhancement component (CFEM) and a selective scanning mechanism to model long-range dependencies at linear complexity. Experimental results demonstrate that Bi-CMFM achieves EN = 5.1281 and AG = 7.0127 on CT-MRI fusion, PSNR of 12.7624/19.6562 on PET-MRI/SPECT-MRI tasks, and top-ranked EN, SF, AG, SCD, and CC on KAIST, outperforming six representative baselines including DATFuse, DRCM, and FusionMamba.

Keywords: multimodal image fusion, bi-path residual, cross-modal feature enhancement, selective scanning, state space model

1 Introduction

Multimodal medical image fusion generates comprehensive images by integrating skeletal structure from CT, soft-tissue detail from MRI, and metabolic activity from PET, thereby supporting clinical diagnosis[14], surgical planning, and lesion localization[1]. Driven by deep learning[15] and the adoption of CNNs, GANs, and Transformers[16], the field has shifted from handcrafted transform-domain strategies to data-driven paradigms. Early approaches such as the Laplacian Pyramid[3], guided filtering[4], and Empirical Mode Decomposition[5] offered interpretability but could not accommodate

cross-modal heterogeneity; their multi-stage pipelines remain sensitive to noise and registration errors, causing artifacts and fine lesion-boundary detail loss[8].

Three fundamental challenges targeted by this paper persist in existing methods. (1) Global–local imbalance: CNNs' local receptive fields limit long-range dependency modeling[2], causing blurred edges; Transformers capture global context but scale quadratically with resolution. (2) Redundancy suppression: most methods inadequately enhance modality-specific features or filter cross-modal redundancy, causing detail loss. (3) Efficiency–quality trade-off: high-performing models are computationally expensive, limiting their adoption in clinical deployment. CNN-based methods such as DRCM[5] and Res2Net[7] improve multi-scale representation but cannot model global spatial correlations. Transformer-based ECFusion[20] improves edge preservation via cross-scale attention[19] but suffers from quadratic complexity. FusionMamba[13] achieves linear complexity yet is limited by unidirectional scanning on dense medical structures.

To address these three issues, this paper proposes Bi-CMFM. BPRF directly tackles the global–local imbalance by reconciling local detail and global context via parallel multi-scale feature extraction and receptive field expansion. CMFM targets redundancy suppression and efficiency by employing CFEM to highlight inter-modal differences and lightweight attention[17] to suppress redundancy, while a selective scanning mechanism dynamically weights high-discrepancy regions, jointly improving fusion quality and efficiency.

Extensive experiments on the HMS Brain Atlas (CT-MRI, PET-MRI, SPECT-MRI) and KAIST multispectral pedestrian datasets demonstrate that Bi-CMFM consistently outperforms six representative baselines. Specifically, Bi-CMFM achieves the highest EN (5.1281) and AG (7.0127) on CT-MRI fusion, the best PSNR on SPECT-MRI fusion (19.6562), and top-ranked EN, SF, AG, SCD, and CC on KAIST, confirming the effectiveness of its global–local balance and redundancy suppression design.

The main contributions of this paper are as follows:

1. We propose BPRF, a bi-path residual fusion module that balances local texture and global structure through parallel standard and dilated convolutions with a Local Detail Compensation branch, directly addressing the global–local imbalance challenge.
2. We propose CMFM, a hierarchical cross-modal fusion module that improves complementary information utilization via CFEM-driven dynamic feature enhancement and selective scanning, effectively suppressing inter-modal redundancy.
3. We introduce a feature-difference-driven selective scanning mechanism that enhances fusion quality at sub-Transformer complexity, enabling Bi-CMFM to achieve an effective efficiency–quality trade-off suitable for clinical deployment.

2 Related Work

Early methods relied on handcrafted transform-domain strategies. Burt and Adelson[3] proposed the Laplacian Pyramid (LP) for coarse-to-fine multi-scale decomposition via Gaussian filtering[4]. Huang et al.[5] proposed Empirical Mode Decomposition

(EMD), adaptively decomposing signals into intrinsic mode functions. Suryanarayana et al.[6] combined both approaches for CT-MRI integration. These methods share a critical limitation: handcrafted rules cannot accommodate cross-modal heterogeneity (e.g., PET vs. CT density), and multi-stage pipelines are sensitive to noise and registration errors, prone to artifacts and loss of fine lesion-boundary detail. Crucially, they fail to capture the deep semantic correlations between modalities that are essential for accurate lesion localization[8].

CNN-based encoder-decoder architectures became dominant with the rise of deep learning. DRCM disentangles shared and modality-specific features via cross-mutual-information loss and adversarial objectives, fusing them with coordinate and multi-modal attention[17]. Res2Net[10] enhances feature hierarchy through multi-scale residual blocks and spatial mean attention for dynamic weight adjustment. However, a fundamental limitation of CNN-based methods is that their local receptive fields inherently restrict long-range modeling[2], overlooking global correlations between distant lesions.

Transformer-based methods leverage self-attention for long-range correlations. ECFusion[20] extracts edge priors via Sobel operators (EAM), maps multi-scale features to a unified space (HCEL), and applies self-attention for cross-scale and cross-modal correlations[19], improving edge preservation in PET-MRI fusion. However, self-attention scales quadratically with resolution, and Transformers produce over-smoothed outputs due to weaker local texture modeling. In summary, existing methods fail to simultaneously address global–local imbalance, insufficient redundancy suppression, and the efficiency–quality trade-off. These three unresolved limitations directly motivate the design of Bi-CMFM.

3 Methodology

3.1 Overall Architecture

Bi-CMFM adopts a multi-scale feature pyramid architecture (Fig.1). The encoder progressively downsamples to extract hierarchical features; the decoder upsamples to reconstruct the fused output.

In Stage1, input image I is converted to a feature sequence by Patch Embedding and processed by BPRF, producing shallow feature maps F^1. CMFM fuses both modalities' F^1 to yield fused features X^1. In Stage2, X^1 is concatenated via Patch Merging and downsampled through BPRF to produce F^2 with a larger receptive field; CMFM then performs deeper cross-modal fusion. After two extract-fuse rounds, Patch Expanding upsamples the features, aligning deep semantics with shallow detail to produce a fused image with sharp edges and structural consistency.

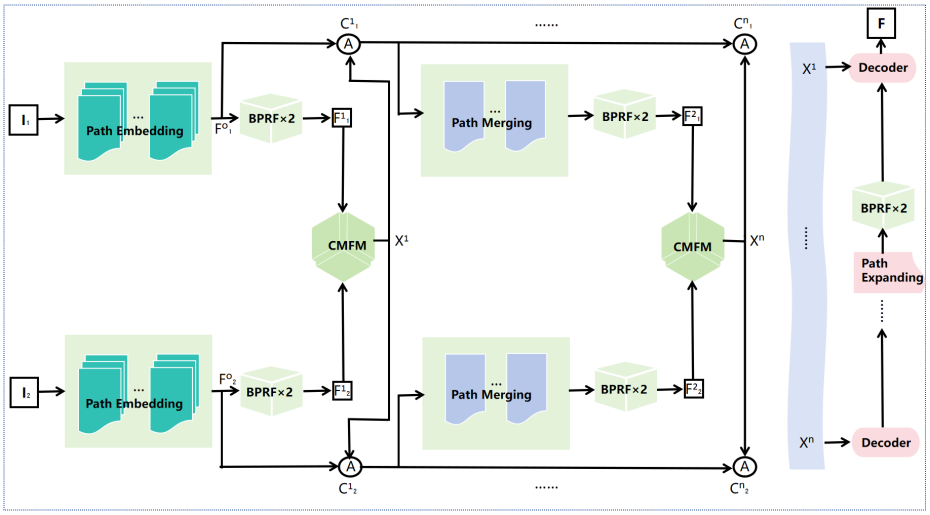

Fig. 1. Overall architecture of Bi-CMFM.

3.2 Bi-Path Residual Fusion Module (BPRF)

BPRF captures local texture and global context within a single pipeline (Fig.2), formulated as:

$$F^{Dn+1} = BPRB(\,ECA(\,LN(\,ESSM(\,LN(\,F^{Dn}\,)\,)\,)\,)\,) + LDC(\,F^{Dn}\,) \quad (1)$$

where F^{Dn} is the input at layer n. LN stabilizes training; ESSM models long-range spatial dependencies at linear complexity; ECA recalibrates channels to suppress redundancy[17]. The local path applies standard convolution to preserve high-frequency texture; the global path uses dilated convolution to extend the receptive field. Their outputs are summed. In parallel, a Local Detail Compensation (LDC) branch applies dynamic convolution to F^{Dn} to recover structural details lost during sequential processing, added residually to produce F^{Dn+1}. This "global modeling + local compensation" design balances detail sharpness with structural consistency at minimal cost.

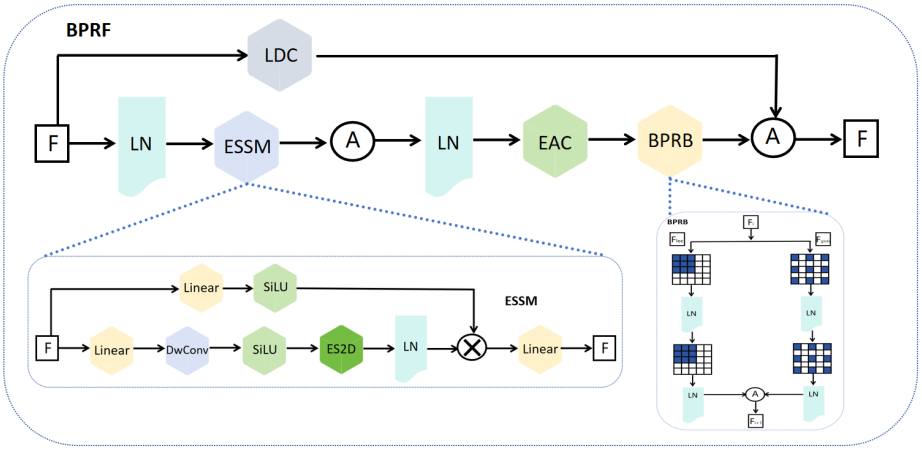

Fig. 2. Architecture of the BPRF module.

3.3 Cross-Modal Fusion Module (CMFM)

CMFM achieves hierarchical cross-modal fusion from local enhancement to global integration (Fig.3). The Cross-modal Feature Enhancement Module (CFEM) is first applied in parallel to each modality: it computes feature difference maps to highlight inter-modal discriminability and adaptively enhances key regions, ensuring modality-specific information is fully preserved. The enhanced features then enter the Cross-modal Fusion Module (CFM) for global integration via deep feature mixing and spatial dependency modeling, followed by channel attention to suppress redundant information flow. A feature alignment mechanism further prioritizes high-discrepancy regions. Through CFEM's parallel local analysis and CFM's sequential global fusion, CMFM forms a complete "enhance–interact–select" pipeline that resolves inter-modal misalignment and optimizes complementarity via adaptive weighting.

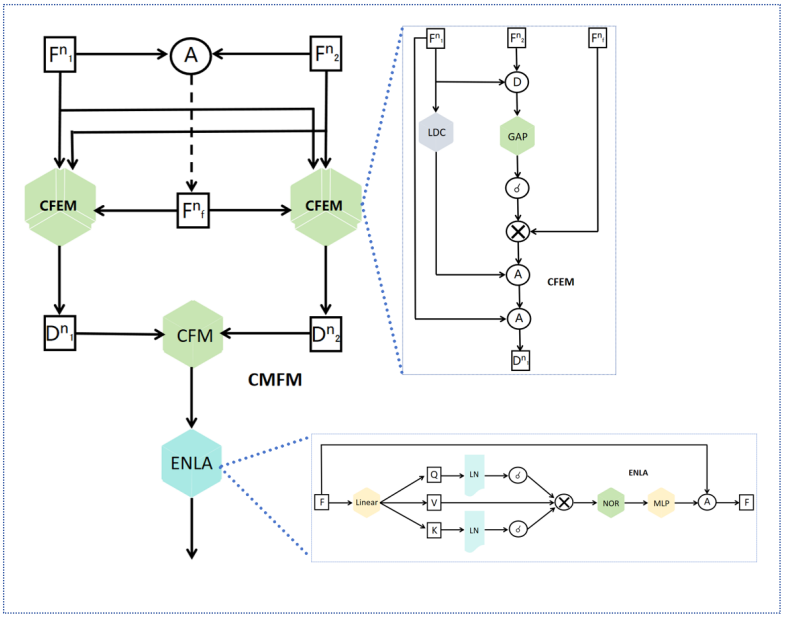

Fig. 3. Architecture of the CMFM module.

4 Experiments

4.1 Datasets

Medical experiments used the Harvard Medical School (HMS) Brain Atlas dataset: 166 CT-MRI, 329 PET-MRI, and 539 SPECT-MRI pairs of rigorously co-registered multi-modal brain images covering normal anatomy and pathologies (tumors, stroke). CT provides skeletal structure; MRI captures soft-tissue detail; PET and SPECT reflect metabolic activity and cerebral blood perfusion. All images were resampled to 256×256 via cubic spline interpolation and augmented with 90°/180°/270° rotations, expanding each modality to 30,000 pairs.

Generalization used the KAIST[9] multispectral pedestrian dataset: 70,000 spatio-temporally co-registered IR–VIS pairs covering day/night, rain/fog, and complex occlusions. Infrared images highlight heat-emitting targets[18]; visible images provide texture and color detail. Images were converted to grayscale and resampled to 256×256 via bilinear interpolation. The fundamental imaging mechanism difference between IR and VIS provides a rigorous test of cross-domain adaptability.

4.2 Training Configuration and Evaluation Metrics

All models are trained with a batch size of 4 using the Adam optimizer with a learning rate of 1×10^{-4}. The loss function combines structural, texture, and intensity terms

weighted by $\alpha_1 = 1$, $\alpha_2 = 1$, and $\alpha_3 = 10$, respectively. All experiments are conducted on an NVIDIA A30 GPU using the PyTorch framework.

Six evaluation metrics[10][11]: Entropy (EN, information richness↑), Average Gradient (AG, edge sharpness↑), Spatial Frequency (SF↑), Correlation Coefficient (CC↑), Sum of Correlations of Differences (SCD↑), Peak Signal-to-Noise Ratio (PSNR↑).

Comparison Methods. Six representative baselines spanning conventional, CNN, GAN, Mamba, and Transformer paradigms: DATFuse[12] (dual-attention Transformer), DRCM[5] (GAN-based disentangled representation), FusionMamba[13] (Mamba-based dynamic fusion), Res2Net[7] (CNN multi-scale residual), LP&EMD[6] (conventional LP+EMD), ECFusion[20] (cross-scale Transformer with edge enhancement). All use public code with original hyperparameters.

4.3 Comparative Experiments

We evaluate three medical fusion tasks: CT-MRI, PET-MRI, and SPECT-MRI. Qualitative and quantitative results are shown in Fig.4 and Table 1 – 3.

Table 1. Comparison of Methods for CT-MRI Datasets.

Method	EN	SF	AG	CC	SCD	PSNR
DATFuse	3.3689	32.7665	6.8772	0.7896	1.4643	12.7951
DRCM	3.5357	32.8182	6.8754	0.7909	1.5699	12.8262
FusionMamba	3.5542	32.8227	6.8901	0.7927	1.5888	12.8337
Res2Net	4.2046	32.9227	6.9034	0.7487	1.6479	12.1277
LP&EMD	4.2158	32.8129	6.9038	0.7895	1.6676	12.1276
ECFusion	4.2758	32.9129	6.0255	0.7959	1.6127	13.9857
BiPathCMFM	5.1281	33.2385	7.0127	0.8114	1.7021	13.0838

Table 2. Comparison of Methods in PET-MRI Datasets.

Method	EN	SF	AG	CC	SCD	PSNR
DATFuse	4.7384	33.6551	11.3979	0.7207	1.2978	11.5523
DRCM	4.7443	33.6853	11.3882	0.7194	1.2379	11.5887
FusionMamba	4.9296	33.8087	11.4194	0.7239	1.3906	11.6668
Res2Net	6.1594	36.0952	11.5353	0.7559	1.7388	12.7057
LP&EMD	6.1023	36.1189	11.5294	0.7508	1.6997	12.6015
ECFusion	6.0554	36.2548	11.5320	0.7552	1.9488	12.6945
BiPathCMFM	6.2406	36.2875	11.5409	0.7573	1.8733	12.7624

Table 3. Comparison of Methods in SPECT-MRI Datasets.

Method	EN	SF	AG	CC	SCD	PSNR
DATFuse	3.7629	15.6018	4.9587	0.8519	1.0892	19.1893
DRCM	3.7551	15.5871	4.9463	0.8502	1.0538	19.2885
FusionMamba	3.9403	15.7195	4.9678	0.8533	1.2106	19.3574

Res2Net	4.5455	17.9573	5.1559	0.8675	1.6729	19.4778
LP&EMD	4.5729	18.9851	5.1593	0.8692	1.6018	19.5036
ECFusion	4.5553	17.9683	5.1466	0.8671	1.6667	19.4627
BiPathCMFM	4.7378	18.1029	5.1684	0.8703	1.7621	19.6562

Qualitative Analysis. In CT-MRI fusion, DATFuse[12] produces artifacts at cranial bone edges. DRCM[5] loses brain tissue texture due to insufficient global interaction. FusionMamba[13] is limited by unidirectional scanning on dense CT structures. Res2Net[7] and LP&EMD[6] preserve acceptable soft-tissue texture but lack edge sharpness. ECFusion[20] improves structural consistency but amplifies noise. Bi-CMFM's standard path preserves MRI sulcal texture while the dilated path captures CT bone density distribution; CMFM's dynamic weighting further suppresses redundancy, achieving a balance between cranial clarity and soft-tissue detail.

In PET-MRI fusion, DATFuse produces low contrast with poor metabolic region visibility; LP&EMD and ECFusion deviate significantly from PET color distribution; DRCM and Res2Net retain metabolic features but lose anatomical detail. Bi-CMFM's CFEM computes differential maps between PET activity and MRI anatomy; CFM's selective scanning prioritizes high-discrepancy regions, preserving both metabolic hotspots and anatomical structures. SPECT-MRI follows a similar pattern.

Quantitative Analysis. On CT-MRI (Table2), Bi-CMFM achieves EN=5.1281, AG=7.0127, CC=0.8114[10][11], ranking first across all six metrics. These gains are attributable to BPRF's dual-path design: the dilated-convolution path extends the receptive field to capture long-range bone boundaries (improving AG), while the LDC branch recovers fine sulcal texture suppressed during deep feature extraction (improving EN and CC). On PET-MRI, Bi-CMFM leads on EN and AG; the slightly lower SCD compared to ECFusion (1.8733 vs. 1.9488) reflects a deliberate trade-off between metabolic contrast amplification and structural fidelity, where Bi-CMFM preserves more balanced anatomical detail. On SPECT-MRI, Bi-CMFM achieves the best PSNR (19.6562), confirming that the LDC branch effectively recovers fine cerebral perfusion detail with high signal fidelity. Notably, ECFusion slightly exceeds Bi-CMFM on CT-MRI PSNR (13.9857 vs. 13.0838): this is because ECFusion applies aggressive edge sharpening via Sobel-based edge priors, which boosts pixel-level signal similarity at the cost of over-sharpened artifacts—a trade-off reflected in its lower EN. Bi-CMFM achieves a more balanced and clinically meaningful fusion profile overall.

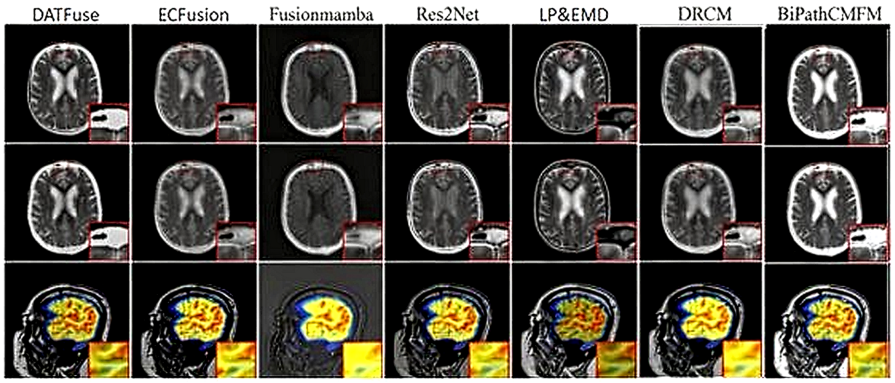

Fig. 4. Visual comparison on the HMS dataset.

4.4 Generalization Experiments

Generalization is evaluated on the KAIST IR-VIS task (Fig.5, Table4).

Table 4. Generalization results on the KAIST dataset.

Method	EN	SF	AG	CC	SCD	PSNR
DATFuse	6.5344	11.2526	3.6267	0.2172	1.2774	10.8783
DRCM	6.5429	11.2037	3.6189	0.2157	1.3089	10.8352
FusionMamba	6.7276	11.3511	3.6425	0.2203	1.4403	11.0239
Res2Net	6.6735	13.9986	3.9389	0.2907	1.5829	12.4501
LP&EMD	6.6489	14.1213	3.9474	0.2845	1.6067	12.4188
ECFusion	6.6587	14.0419	3.9321	0.2962	1.6503	12.3952
BiPathCMFM	6.8412	14.1731	3.9556	0.2950	1.7394	12.5833

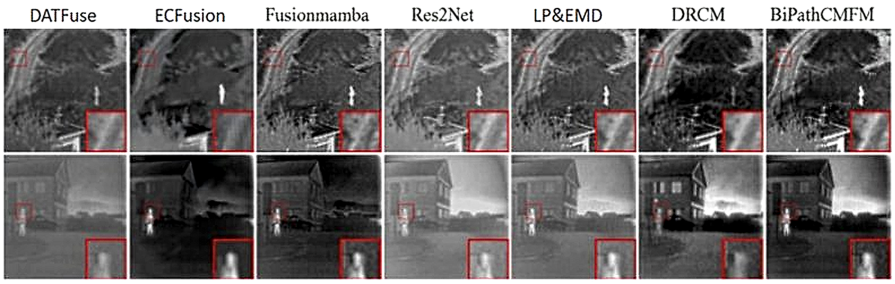

Fig. 5. Visual generalization results on the KAIST dataset.

DATFuse's handcrafted rules cause severe visible-light texture loss under low illumination; DRCM blurs infrared edges; FusionMamba degrades thermal clarity in fog; Res2Net cannot jointly preserve visible-light detail and infrared targets. By contrast, Bi-CMFM's BPRF parallel paths preserve fine-grained texture and capture infrared

contour structure simultaneously. CFEM drives selective scanning toward high-temperature-differential regions, while lightweight attention adaptively weights visible-light background detail. Bi-CMFM achieves top scores on EN, SF, AG, and SCD (Table4), confirming strong unified modeling capacity across heterogeneous modalities.

4.5 Ablation Study

Five ablation experiments are summarized in Table5. Replacing BPRF with Dynamic Visual State Space (DVSS) degrades all metrics; replacing it with Transformer causes larger drops in EN, AG, and SCD. Removing CFEM (w/o CFEM) drops EN from 6.8412 to 5.9489 and SCD from 1.7394 to 1.2258, confirming CFEM's indispensable role. Replacing all of CMFM with BPRF yields similar degradation, substantiating CFEM and CMFM's joint contribution. Removing BPRF (w/o BPRF) also degrades all metrics, validating the bi-path residual design.

Table 5. Ablation study results.

Method	EN	SF	AG	CC	SCD	PSNR
BPRF→DVSS	6.6489	14.0475	3.9238	0.2908	1.6497	12.4068
BPRF→Transforme r	5.9198	13.8207	3.8013	0.2689	1.1572	11.9005
CMFM→BPRF	5.8557	13.8119	3.8278	0.2907	1.2389	12.4501
w/o CFEM	5.9489	13.8475	3.8367	0.2735	1.2258	11.9406
w/o BPRF	6.3589	13.9164	3.8664	0.2825	1.5409	12.2092
BiPathCMFM	6.8412	14.1731	3.9556	0.2950	1.7394	12.5833

5 Conclusion

Bi-CMFM is proposed to address two core limitations in multimodal medical image fusion: insufficient global context capture and improper inter-modal redundancy handling. BPRF integrates standard and dilated convolutions in parallel to compensate for the long-range dependency limitation of CNNs. CMFM employs hierarchical processing via CFEM and CFM, using selective scanning and lightweight attention to dynamically enhance inter-modal differential features and suppress redundancy at sub-Transformer complexity, making it suitable for clinical settings. On HMS, Bi-CMFM outperforms DATFuse, DRCM, ECFusion, and others on EN, AG, and CC across three fusion tasks. KAIST generalization confirms cross-modal modeling capacity. Ablation confirms both modules are necessary; Transformer replacement causes clear degradation, validating the bi-path residual design.

Limitations: Bi-CMFM does not surpass ECFusion on PSNR, indicating a brightness–noise trade-off for further optimization. The framework currently processes 2D slices only; 3D extension remains to be explored. Future work may incorporate refined cross-modal alignment and acquisition-pipeline priors to further improve clinical utility.

References

[1] Kavitha, C.T., et al.: A comprehensive review of techniques, algorithms, advancements, challenges, and clinical applications of multi-modal medical image fusion for improved diagnosis. Comput. Methods Programs Biomed. Update 7, 100189 (2025)

[2] Zhou, T., Cheng, Q.R., Lu, H.L., et al.: Deep learning methods for medical image fusion: A review. Comput. Biol. Med. 160, 106959 (2023)

[3] Burt, P.J., Adelson, E.H.: The Laplacian pyramid as a compact image code. In: Readings in Computer Vision, pp. 671–679. Morgan Kaufmann (1987)

[4] Li, S., Kang, X., Hu, J.: Image fusion with guided filtering. IEEE Trans. Image Process. 22(7), 2864–2875 (2013)

[5] Huang, W., Zhang, H., Cheng, Y., et al.: DRCM: A disentangled representation network based on coordinate and multimodal attention for medical image fusion. Front. Physiol. 14, 1241370 (2023)

[6] Suryanarayana, G., Nimmagadda, S.M., Nageswara Rao, S., et al.: Enhanced MRI-PET fusion using Laplacian pyramid and empirical mode decomposition for improved oncology imaging. PLoS ONE 20(5), e0322443 (2025)

[7] Chen, W., Li, Q., Zhang, H., et al.: MR–CT image fusion method of intracranial tumors based on Res2Net. BMC Med. Imaging 24(1), 169 (2024)

[8] Sorin, V., et al.: Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. Life 14(8), 1013 (2024)

[9] Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I.S.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1037–1045 (2015)

[10] Liu, Z., et al.: A review on infrared and visible image fusion algorithms based on neural networks. J. Vis. Commun. Image Represent. 103, 104269 (2024)

[11] Cheng, C., et al.: EvaNet: Towards more efficient and consistent infrared and visible image fusion assessment. IEEE Trans. Image Process. (2025)

[12] Tang, W., He, F., Liu, Y., Duan, Y., Si, T.: DATFuse: Infrared and visible image fusion via dual attention transformer. IEEE Trans. Circuits Syst. Video Technol. 33(7), 3159–3172 (2023)

[13] Xie, X., Cui, Y., Tan, T., Zheng, X., Yu, Z.: FusionMamba: Dynamic feature enhancement for multimodal image fusion with Mamba. Vis. Intell. 2(1), 37 (2024)

[14] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer, Cham (2015)

[15] Chen, J., Lu, Y., Yu, Q., et al.: TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

[16] Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. 30 (2017)

[17] Wang, X., Girshick, R., Gupta, A., et al.: Non-local neural networks. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)

[18] Ma, J., Tang, L., Fan, F., et al.: SwinFusion: Cross-domain long-range learning for general image fusion via Swin Transformer. IEEE/CAA J. Autom. Sin. 9(7), 1200–1217 (2022)
[19] Niu, Z., Zhong, G., Yu, H.: A review on the attention mechanism of deep learning. Neurocomputing 452, 48–62 (2021)
[20] Luo, F., Wu, D., Pino, L.R., et al.: A novel multimodel medical image fusion framework with edge enhancement and cross-scale transformer. Sci. Rep. 15(1), 1165025 (2024)