BiMSRec: A Progressive Image Reconstruction Framework for Medical Image Fusion Guided by Multi-Scale Deformation Fields

Nuoer Long[1], Kaiwen Yang[1], Xinyu Xie[1], Zitong Yu[2], Tao Tan[1] and Yue Sun[1(✉)]

[1] Faculty of Applied Sciences, Macao Polytechnic University, Macao, China
yuesun@mpu.edu.mo
[2] School of Computing and Information Technology, Great Bay University, Dongguan, China

Abstract. Traditional multi-modal medical image fusion methods typically employ a hierarchical feature fusion strategy. However, due to inconsistencies among features at different scales, these approaches often introduce unanticipated deformations during the fusion process. Such deformations accumulate through successive registration steps and ultimately result in oscillatory distortions at the fine-detail level. To address this challenge, we propose a progressive image reconstruction framework that is guided by multi-scale deformation fields. Specifically, the input images are first mapped into feature spaces at multiple scales and a deformation field prediction strategy is employed to generate multiple deformation fields that capture both local and global transformation trends simultaneously. Notably, the deformation fields generated across all scales possess the intrinsic capability to directly perform image registration. This capability eliminates the need for sequential propagation of registration outcomes and effectively mitigates cumulative deformation issues. In the image reconstruction phase, we adopt a progressive coarse-to-fine strategy, leveraging multi-scale deformation fields to achieve accurate structure restoration and fusion. Extensive experimental results demonstrate that the proposed method significantly enhances image alignment accuracy and fusion quality across multiple datasets, offering an efficient and robust solution for multi-modal medical image processing.

Keywords: Medical Image Fusion · Multi-modal · Deformation Field

1 Introduction

Multi-modal medical image fusion plays a crucial role in clinical diagnosis, as complementary information from different imaging techniques helps to comprehensively characterize lesion features. Each modality's unique advantages pro-

N. Long and K. Yang — Equal Contribution.
✉ Corresponding author.

vide more comprehensive and accurate image data[23, 22]. In multi-modal fusion, key information needs to be retained according to the characteristics of each modality. For instance, Computed Tomography (CT) provides high-contrast bone information, while Magnetic Resonance Imaging (MRI) excels in soft tissue imaging, so CT should retain bone structures and MRI should preserve soft tissue details. In PET-MRI fusion, Positron Emision Tomograph (PET) provides metabolic information and MRI offers anatomical details, so metabolic areas from PET and anatomical features from MRI should be preserved. In SPECT-MRI fusion, (Single-Photon Emission Computed Tomography(SPECT) offers blood flow and functional information, while MRI provides structural details. The functional regions in SPECT and structural features in MRI should be retained. For other multi-modal image fusions, the information to retain must also be determined based on their specific characteristics to ensure the fused image's effectiveness and reliability in diagnosis[4, 3].

In actual clinical practice, due to different acquisition devices with varying imaging principles and parameters, as well as unavoidable subtle movements, breathing, and heartbeat of patients during the examination process, positional deviations are often introduced. Therefore, precise registration must be performed before image fusion. Medical image registration includes both traditional optimization-based methods and deep learning-based automatic registration schemes, aiming to achieve optimal spatial alignment through deformation or rigid transformation. For a long time, registration and fusion have been regarded as two relatively independent research directions. Researchers often focus on improving registration accuracy[1, 14] while neglecting the impact of registration errors on subsequent fusion quality. To address the above issue, some studies have proposed a two-stage strategy [20], where multi-modal images are first spatially aligned using a registration algorithm, followed by the fusion process. While this approach facilitates the utilization of state-of-the-art registration techniques and allows independent optimization of the fusion network, registration errors often propagate through subsequent processing stages. Additionally, the separate training objectives of the two stages may not be fully compatible, potentially leading to increased computational complexity and a decline in overall performance.

Meanwhile, multi-modal medical image fusion methods often rely on feature fusion strategies in RGB or YCrCb color spaces[8, 10], as these spaces can retain luminance and chrominance information, aiding in the alignment and enhancement of images from different modalities. However, these methods are typically limited to pixel intensity-based weighting, transformation, or filtering operations, making it difficult to capture complex local and global deformations effectively[6]. In particular, when non-linear deformations exist in different modalities, color space-based fusion strategies may lead to the accumulation of registration errors.

Currently, most fusion methods fail to fully exploit the inherent similarity between multi-modal images in feature representation. The structural or texture features originally hidden in the images could have provided strong guidance for spatial deformation. Based on this, this study designed a multi-modal and multi-scale deformation field registration network (M2FReg), which uses the implicit similar structure and texture information between images to directly predict the deformation field independently at each scale, ensuring that each scale can complete the registration task independently, effectively avoiding the cumulative error caused by dependence on previous results. Finally, through the Progressive Multi-Scale Flow-Guided Reconstruction Network (PFRecon) specially designed for multi-scale deformation field. This network adopts a top-down reconstruction strategy, first capturing the basic outline of the overall structure of the image on a global scale, and then gradually introducing medium-scale and local detailed features. Through the fusion of multi-scale deformation fields, the method achieves detailed structural restoration and high-quality fusion of multi-modal images.

In summary, the main contributions of this research are: 1. Traditional multi-modal medical image fusion methods rely on RGB or YCrCb color spaces for feature fusion, but they have limitations in capturing local and global deformations. This paper introduces a deformation field-based guidance mechanism, providing richer motion and deformation features to enhance the structural reference for image registration. 2. Progressive feature fusion methods often rely on continuously generated deformation fields, which can lead to the gradual accumulation of errors. Our method employs a direct multi-scale deformation field estimation strategy, where the deformation field at each scale independently performs registration. This approach prevents error accumulation from coarse-to-fine propagation and substantially enhances registration accuracy and robustness. 3. To process the multi-scale deformation field inputs, we designed a progressive global-to-local reconstruction strategy that effectively integrates deformation information at each scale, thereby enhancing both image alignment accuracy and fusion quality. 4. Our method exhibits excellent performance under different registration states, proving its wide applicability and practical value in multi-modal medical image processing.

2 Methodology

2.1 General framework

As shown in Fig 1, the framework includes three main components: a feature extraction network, M2FReg network, and PFRecon network. First, the feature extraction network extracts rich features to enhance the model's ability to understand multi-modal data. The M2FReg network estimates bidirectional multi-scale deformation fields, operating in both global-to-local and local-to-global directions, enabling independent image registration at each scale. Finally, the PFRecon network performs progressive image reconstruction from coarse to fine by integrating all multi-scale deformation fields.

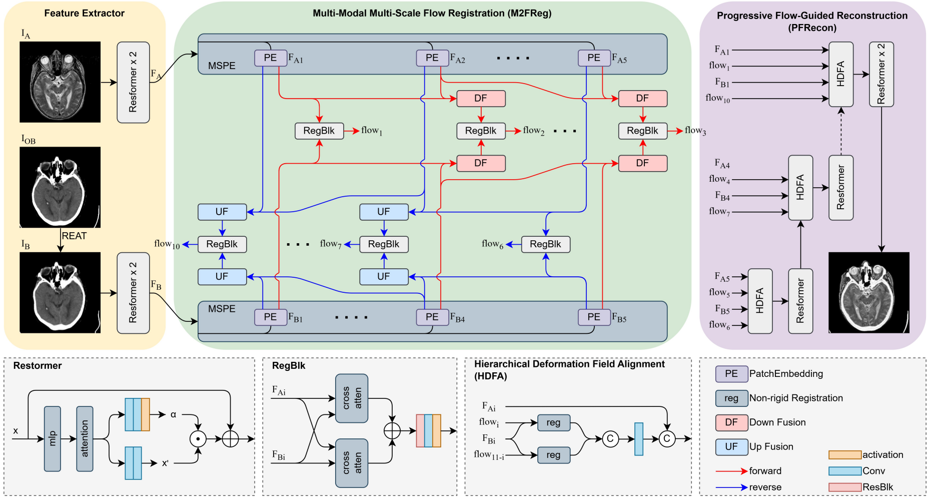

Fig. 1: Taking CT-MRI data as an example, the figure shows the overall framework of the proposed method.

2.2 Feature Extraction Network

Restormer[26] has shown excellent performance in image feature extraction tasks, so in the feature extraction network, we use the Restormer module to extract deep semantic features from unaligned input images. Since BiMSRec does not fully rely on the characteristics of color space, we only obtain part of the color space information. Specifically, during data preprocessing, the input grayscale or RGB image is transformed into the YCrCb color space, and only the luminance channel (Y) is retained to obtain single-channel data. Given the input images I_A and $I_B(I \in \mathbf{R}^{1\times H\times W})$, the feature extraction network processes them to generate feature maps F_A and $F_B(F \in \mathbf{R}^{C\times H\times W})$, where the number of feature channels C is set to 16.

2.3 M2FReg Network

In the MFReg network, we employ a bidirectional deformation field estimation framework to jointly perform registration and fusion of multi-scale features, producing high-quality dense deformation fields. The network maps F_A and F_B to F_{A_i} and F_{B_i} through Multi-Scale Patch Embedding (MSPE), where $F_i \in \mathbf{R}^{C_i\times H_i\times W_i}(i = 1, ..., 5)$, with the number of channels C_i and resolution (H_i, W_i) decreasing progressively. This hierarchical design allows high-level features to capture richer semantic information while preserving fine-grained details in the low-level features, enhancing the accuracy and robustness of the registration process.

The core component of M2FReg is the feature registration block (RegBlk). This structure adopts a differential cross-attention mechanism to provide strong support for image registration by mining the intrinsic similarities in multi-modal feature images. The bidirectional cross-attention calculation is as follows:

$$F'_A = Softmax(Q_A K_B^T/\sqrt{d})V_B, F'_B = Softmax(Q_B K_A^T/\sqrt{d})V_A. \tag{1}$$

Among them, Q, K, and V are obtained by linear transformation of F_{A_i} and F_{B_i}. Then, the cross-attention features F'_{A_i} and F'_{B_i} are fused to construct rich matching information. Finally, the prediction network $flow_i$ is obtained through the residual block and activation function.

The overall network employs a bidirectional computation mechanism. As illustrated in Fig 1, the top-down pathway estimates a forward deformation field. Local feature layers are progressively integrated into global feature layers through Down Fusion (DF), ensuring robust estimation of global deformation fields while maintaining consistency under large transformations. Conversely, the bottom-up pathway estimates a backward deformation field. Global feature layers are progressively integrated into local feature layers through Up Fusion (UF) to enhance fine-grained detail restoration, effectively recovering high-frequency information lost during large-patch processing.

2.4 PFRecon Network

Building upon the multi-modal multi-scale deformation field registration network, we propose a progressive multi-scale flow-guided reconstruction network (PFRecon) that performs coarse-to-fine image reconstruction by fully utilizing the bidirectionally predicted deformation fields.

The network input consists of five sets of multi-scale deformation fields estimated by M2FReg. Given that M2FReg employs a bidirectional information propagation mechanism, we define the scale index i to range from 1 to 5. The deformation field predictions at each scale are denoted as $flow_i$ and $flow_{11-i}$, where $flow \in \mathbf{R}^{2\times H_i \times W_i}$, with the resolution (H_i, W_i) progressively decreasing. The PFRecon network initiates reconstruction using the coarsest-scale deformation field and progressively incorporates finer-scale deformation information, hierarchically refining the output until reaching the highest resolution. Within the Hierarchical Deformation Field Alignment (HDFA) module, bidirectionally estimated deformation fields are employed to align images and fuse their feature representations, ensuring robust and consistent reconstruction:

$$F_{B_i}^{re} = concat(reg(F_{B_i}, flow_i), reg(F_{B_i}, flow_{11-i})). \tag{2}$$

Among them, $F_{B_i}^{re} \in \mathbf{R}^{2C\times H_i \times W_i}$ represents the registration feature at the current scale. reg stands for Non-rigid Registration, which can form an image for deformation field guidance registration. Next, we fuse this feature with the image feature F_{A_i}. The specific calculation is as follows:

$$F_i^{fuse} = concat(F_{B_i}^{re}, F_{A_i}, F_i^{pre}), \tag{3}$$

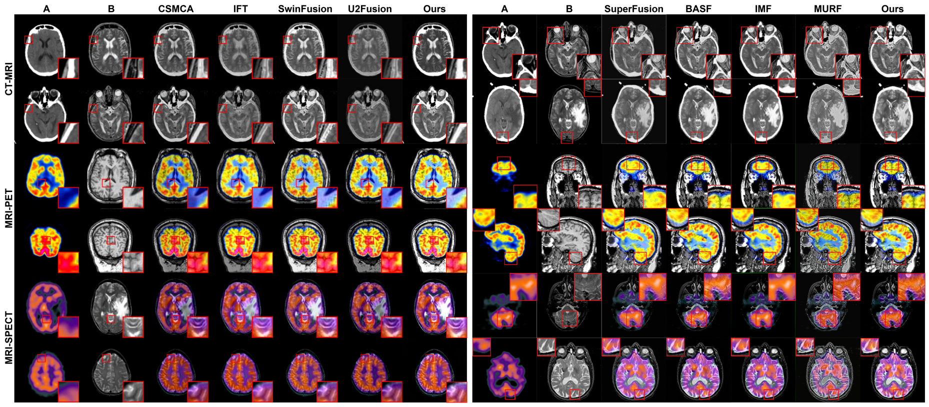

Fig. 2: Visual Comparison of Fusion Results: The left side shows the experimental results of registration-based fusion, and the right side shows the experimental results of joint registration and fusion.

where F_i^{pre} represents the fused features of the previous layer. This hierarchical feature propagation method can complete the initial alignment at the low-resolution stage and gradually supplement the detailed information at the high-resolution stage. It is worth noting that each predicted deformation field has independent registration capability. The PFRecon network progressively optimizes these fields, producing smoother and more accurate results.

2.5 Loss Settings

In this study, the loss function design mainly includes registration loss and fusion loss. Registration loss evaluates the quality of the deformation field and the accuracy of image alignment, ensuring correct registration of multi-modal images. Fusion loss assesses the overall quality of the fused image, ensuring that the output retains critical information from the original images while minimizing detail loss and avoiding the introduction of artifacts.

Inspired by traditional image fusion loss functions, we use SSIM loss[27], L1 loss[15], and gradient loss[13] to ensure the fused image maintains structural integrity, fine details, and smooth transitions.

For the registration task, we adopt color-space-based SSIM loss and L1 loss to regulate image alignment, ensuring accurate registration across different modalities. Additionally, given that our network relies on deformation field guidance, we design an additional set of deformation field losses, which are computed as follows:

$$L_{flow_i} = loss_{ssim}(flow_i, flow_{GT}) + loss_{L1}(flow_i, flow_{GT}), \tag{4}$$

where $flow_{GT}$ is the deformation field of the registered image I_{OB} after random elastic affine transformation $(REAT)$[2]. Compared with the traditional image

Table 1: Qualitative Analysis of Registration-Based Fusion Methods.

	CT-MRI				**PET-MRI**				**SPECT-MRI**			
Model	**SD**	**$Q^{AB/F}$**	**VIF**	**SSIM**	**SD**	**$Q^{AB/F}$**	**VIF**	**SSIM**	**SD**	**$Q^{AB/F}$**	**VIF**	**SSIM**
CSMCA[9]	97.53	0.59	0.31	0.94	95.68	0.50	0.53	0.93	94.23	**0.69**	0.44	0.93
IFT[19]	89.78	0.45	0.37	0.89	102.76	0.52	0.49	0.91	88.47	0.64	0.51	0.92
SwinFusion[11]	98.76	0.63	0.49	**0.95**	100.31	0.74	**0.55**	**0.92**	98.77	0.68	0.64	**0.96**
U2Fusion[23]	90.22	0.28	0.31	0.83	92.29	0.50	0.48	0.89	92.77	0.53	0.54	0.94
BiMSRec	**103.94**	**0.68**	**0.57**	0.92	**106.58**	**0.78**	0.53	0.90	**105.92**	0.68	**0.80**	0.92

Table 2: Qualitative Analysis of Joint Registration and Fusion Methods.

	CT-MRI				**PET-MRI**				**SPECT-MRI**			
Model	**SD**	**$Q^{AB/F}$**	**VIF**	**SSIM**	**SD**	**$Q^{AB/F}$**	**VIF**	**SSIM**	**SD**	**$Q^{AB/F}$**	**VIF**	**SSIM**
SuperFusion[16]	87.83	0.27	0.24	0.46	92.55	0.49	0.26	0.67	84.61	0.51	0.30	0.47
BASFusion[7]	90.24	0.42	0.27	0.66	97.63	**0.78**	0.42	0.52	84.61	**0.66**	0.42	0.56
IMFusion[21]	**97.27**	0.28	0.19	0.62	90.27	0.62	0.30	**0.55**	**98.17**	0.52	0.31	0.54
MURF[24]	82.60	0.31	0.20	0.61	92.24	0.51	0.24	0.43	89.23	0.49	0.26	0.42
BiMSRec	93.71	**0.49**	**0.29**	**0.70**	**100.76**	0.60	**0.46**	0.51	97.46	0.64	**0.47**	**0.59**

fusion task based on RGB or YCrCb color space, our method can additionally compare the predicted deformation field with the ground truth deformation field.

3 Experiments

3.1 Experimental Setup

This study utilizes multi-modal medical imaging data from the Harvard Medical Dataset, which comprises three sub-datasets: CT-MRI, PET-MRI, and SPECT-MRI. To ensure a consistent data distribution during training and testing, the experimental data division follows established methodologies, such as KPSFusion[17]. In this study, MRI is designated as the reference modality, while the other modalities (CT/PET/SPECT) undergo random elastic and affine transformations to simulate real-world deformations.

The experimental evaluation adopts two methods. On the one hand, multiple indicators are calculated to measure the quality of the fused image, and on the other hand, the performance of different methods is directly compared on three datasets. The indicators we choose include standard deviation (SD)[25], gradient-based quality index ($Q_{AB/F}$), fidelity to visual information (VIF)[5], and structural similarity index ($SSIM$)[12].

The experiment was implemented in PyTorch. The Adam[18] optimizer was used in the training process. The initial learning rate was set to 5e-5, the batch size was 8, and the number of training rounds was 100. All experiments were run on the NVIDIA A40 GPU.

http://www.med.harvard.edu/aanlib/

3.2 Comparison With the State-of-the-art Methods

In this section, we conduct experiments on three representative multi-modal medical image fusion tasks and compare our method with state-of-the-art approaches. Qualitative results are shown in Fig 2. Our method excels in preserving multi-modal structural information, enhancing detail clarity, and improving contrast. Unlike conventional methods relying on RGB and YCrCb color spaces, our approach estimates the registered color distribution by computing offsets based on predicted deformation fields, enabling BiMSRec to achieve robust performance in RGB modality fusion with accurate registration and high-quality outcomes.

Table 1 and Table 2 show the quantitative comparison results across three datasets based on four evaluation metrics. Our method achieves the highest scores on multiple metrics, demonstrating its effectiveness. Notably, it excels in SD, indicating superior contrast preservation and edge sharpness in multi-modal fusion. The result of visual comparison aligns with the qualitative analysis, further validating BiMSRec's robustness in enhancing fusion quality.

3.3 Ablation experiment

To validate M2FReg, we designed forward and reverse deformation field registration networks, each retaining a single direction of flow. Additionally, to assess PFRecon's contribution, we designed a reconstruction network guided by single-scale deformation fields. The ablation study includes both ablation results in Table 3 and visual ablation in Fig 3.

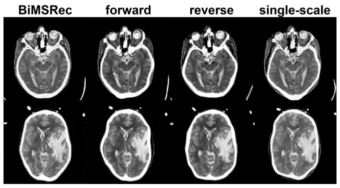

Fig. 3: Visual ablation

Table 3: Ablation results

	SD	$Q^{AB/F}$	VIF	SSIM
BiMSRec	103.94	0.68	0.57	0.92
forward only	82.41	0.47	0.32	0.76
reverse only	98.26	0.52	0.49	0.88
single-scale	82.35	0.41	0.34	0.72

In comparison, the unidirectional deformation field registration network results in degraded evaluation metrics, demonstrating that bidirectional registration preserves critical details like texture and contrast. Furthermore, single-scale deformation field reconstruction significantly compromises registration performance, underscoring PFRecon's ability to effectively integrate multi-scale deformation fields - a cornerstone of the BiMSRec framework.

4 Conclusion

This study presents a novel framework for multi-modal, multi-scale deformation field-based registration and fusion. Leveraging the multi-scale deformation fields estimated by M2FReg, the framework achieves precise cross-modal feature alignment. Meanwhile, PFRecon employs a progressive reconstruction strategy to hierarchically integrate multi-scale deformation features, refining structural details and enabling high-fidelity image fusion. The proposed approach demonstrates strong adaptability across various registration scenarios, underscoring its effectiveness in multi-modal medical image fusion and its potential applicability in clinical settings that demand precise registration and seamless multi-modal data integration.

Acknowledgments. This work is supported by Science and Technology Development Fund of Macao (0041/2023/RIB2), Science and Technology Development Fund of Macao (0004/2024/E1B1), Macao Polytechnic University Grant (RP/FCA-10/2023) and Macao Polytechnic University Grant (RP/FCA-08/2024).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article

References

1. Chen, Z., Wei, J., Li, R.: Unsupervised multi-modal medical image registration via discriminator-free image-to-image translation. arXiv preprint arXiv:2204.13656 (2022)
2. Dalca Adrian, V., Guha, B., John, G., Sabuncu Mert, R.: Unsupervised learning for fast probabilistic diffeomorphic registration, 729–738 (2018)
3. Duan, Y., Pang, P.C.I., He, P., Wang, R., Sun, Y., Liu, C., Zhang, X., Yuan, X., Song, P., Lam, C.T., et al.: 3mt-net: A multi-modal multi-task model for breast cancer and pathological subtype classification based on a multicenter study. IEEE Journal of Biomedical and Health Informatics (2024)
4. Florkow, M.C., Willemsen, K., Mascarenhas, V.V., Oei, E.H., van Stralen, M., Seevinck, P.R.: Magnetic resonance imaging versus computed tomography for three-dimensional bone imaging of musculoskeletal pathologies: a review. Journal of Magnetic Resonance Imaging **56**(1), 11–34 (2022)
5. Han, Y., Cai, Y., Cao, Y., Xu, X.: A new image fusion performance metric based on visual information fidelity. Information fusion **14**(2), 127–135 (2013)
6. Hong, W., Zhang, H., Ma, J.: Ofpf-mef: An optical flow guided dynamic multi-exposure image fusion network with progressive frequencies learning. IEEE Transactions on Multimedia (2024)
7. Li, H., Su, D., Cai, Q., Zhang, Y.: Bsafusion: A bidirectional stepwise feature alignment network for unaligned medical image fusion. arXiv preprint arXiv:2412.08050 (2024)
8. Li, H., Wu, X.J.: Infrared and visible image fusion using latent low-rank representation. arXiv preprint arXiv:1804.08992 (2018)
9. Liu, Y., Chen, X., Ward, R.K., Wang, Z.J.: Medical image fusion via convolutional sparsity based morphological component analysis. IEEE Signal Processing Letters **26**(3), 485–489 (2019)
10. Ma, J., Ma, Y., Li, C.: Infrared and visible image fusion methods and applications: A survey. Information fusion **45**, 153–178 (2019)

11. Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., Ma, Y.: Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. IEEE/CAA Journal of Automatica Sinica **9**(7), 1200–1217 (2022)
12. Ma, K., Zeng, K., Wang, Z.: Perceptual quality assessment for multi-exposure image fusion. IEEE Transactions on Image Processing **24**(11), 3345–3356 (2015)
13. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440 (2015)
14. Mok, T.C., Li, Z., Bai, Y., Zhang, J., Liu, W., Zhou, Y.J., Yan, K., Jin, D., Shi, Y., Yin, X., et al.: Modality-agnostic structural image representation learning for deformable multi-modality medical image registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11215–11225 (2024)
15. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016)
16. Tang, L., Deng, Y., Ma, Y., Huang, J., Ma, J.: Superfusion: A versatile image registration and fusion network with semantic awareness. IEEE/CAA Journal of Automatica Sinica **9**(12), 2121–2137 (2022)
17. Tang, L., Zhang, H., Xu, H., Ma, J.: Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. Information Fusion **99**, 101870 (2023)
18. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: Maxvit: Multi-axis vision transformer. In: European conference on computer vision. pp. 459–479. Springer (2022)
19. Vs, V., Valanarasu, J.M.J., Oza, P., Patel, V.M.: Image fusion transformer. In: 2022 IEEE International conference on image processing (ICIP). pp. 3566–3570. IEEE (2022)
20. Wang, D., Liu, J., Fan, X., Liu, R.: Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. arXiv preprint arXiv:2205.11876 (2022)
21. Wang, D., Liu, J., Ma, L., Liu, R., Fan, X.: Improving misaligned multi-modality image fusion with one-stage progressive dense registration. IEEE Transactions on Circuits and Systems for Video Technology (2024)
22. Xie, X., Cui, Y., Tan, T., Zheng, X., Yu, Z.: Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. Visual Intelligence **2**(1), 37 (2024)
23. Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2fusion: A unified unsupervised image fusion network. IEEE transactions on pattern analysis and machine intelligence **44**(1), 502–518 (2020)
24. Xu, H., Yuan, J., Ma, J.: Murf: Mutually reinforcing multi-modal image registration and fusion. IEEE transactions on pattern analysis and machine intelligence **45**(10), 12148–12166 (2023)
25. Xydeas, C.S., Petrovic, V.: Objective image fusion performance measure. Electronics letters **36**(4), 308–309 (2000)
26. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022)
27. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. IEEE Transactions on computational imaging **3**(1), 47–57 (2016)