Simultaneous Tri-Modal Medical Image Fusion and Super-Resolution using Conditional Diffusion Model

Yushen Xu, Xiaosong Li(✉), Yuchan Jie, and Haishu Tan

Foshan University
lixiaosong@buaa.edu.cn

Abstract. In clinical practice, tri-modal medical image fusion, compared to the existing dual-modal technique, can provide a more comprehensive view of the lesions, aiding physicians in evaluating the disease's shape, location, and biological activity. However, due to the limitations of imaging equipment and considerations for patient safety, the quality of medical images is usually limited, leading to sub-optimal fusion performance, and affecting the depth of image analysis by the physician. Thus, there is an urgent need for a technology that can both enhance image resolution and integrate multi-modal information. Although current image processing methods can effectively address image fusion and super-resolution individually, solving both problems synchronously remains extremely challenging. In this paper, we propose TFS-Diff, a simultaneously realize tri-modal medical image fusion and super-resolution model. Specially, TFS-Diff is based on the diffusion model generation of a random iterative denoising process. We also develop a simple objective function and the proposed fusion super-resolution loss, effectively evaluates the uncertainty in the fusion and ensures the stability of the optimization process. And the channel attention module is proposed to effectively integrate key information from different modalities for clinical diagnosis, avoiding information loss caused by multiple image processing. Extensive experiments on public Harvard datasets show that TFS-Diff significantly surpass the existing state-of-the-art methods in both quantitative and visual evaluations. Code is available at https://github.com/XylonXu01 /TFS-Diff.

Keywords: Tri-Modal Medical Image Fusion · Super-Resolution · Conditional Diffusion Model.

1 Introduction

Multimodal medical images have become an indispensable tool in modern medical diagnosis and treatment planning. Computer Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and Single Photon Emission Computed Tomography (SPECT) each provide unique and complementary information [1], revealing the anatomical structure, physiological function, and molecular changes in the human body, respectively. However, due to the different imaging principles underlying these imaging technologies, images produced by various sensors exhibit significant differences in information content. Although the diversity of imaging enriches the sources of information for clinical diagnosis, it also poses additional challenges for physicians in integrating multi-source image information to make accurate diagnoses [10].

Multimodal image fusion holds promise for combining information from images of different modalities [14,3] to obtain a more comprehensive diagnostic view. Currently, image fusion is mainly divided into methods based on deep learning [15,34] and traditional methods [12,33]. Deep learning-based methods often use generative adversarial networks (GANs) to simulate the distribution of fused images to obtain high-quality fused images [17]. Although GAN-based methods can generate satisfactory fused images, they suffer from issues such as training instability, mode collapse and lack of interpretability. As an improvement, fusion methods based on Diffusion [20,7] have been proposed, which generate high-quality images by simulating the diffusion process of restoring images corrupted by noise to clean images, thereby mitigating common problems like training instability and mode collapse in GANs, and their generation process is interpretable. For example, Zhao et al. [35] proposed using DDPM for fusion tasks and employing a hierarchical Bayesian method to model the subproblems of maximum likelihood estimation. However, no deep learning fusion methods for tri-modal medical images have emerged, and only a few traditional methods have conducted preliminary research on this problem. For instance, Jie et al. [9] proposed a tri-modal medical image fusion based on an adaptive energy selection scheme and sparse representation, using sparse representation to fuse texture components, and adaptive energy selection scheme to fuse cartoon components. Jie et al.[8] proposed a tri-mode medical image fusion and denoising method based on BitonicX filtering. This method analyzes pixels in terms of gradient, energy, and sharpness to achieve medical image fusion and denoising.

Simultaneously, in medical imaging, due to various factors such as the resolution limits of imaging equipment, time constraints on image acquisition, and the radiation doses patients can tolerate, the resulting medical images often have limited resolution. Despite this, there is still an urgent demand for high-resolution (HR) medical images in clinical practice [13]. Currently, deep learning methods for super-resolution can capture fine details and accurately preserve the original structure of images [31,25,23]. For example, Mao et al. proposed a decoupled conditional diffusion model and extended it to multi-contrast MRI super-resolution, effectively estimating the uncertainty of the restoration and ensuring a stable optimization process [16,18]. However, performing image super-resolution and image fusion in separate steps can propagate and amplify artifacts generated in the first step, thereby degrading the quality of the image. To address this issue, research has proposed end-to-end fusion and super-resolution methods for low-resolution images [29,11,28]. For instance, Xiao et al. [28] introduced a heterogeneous knowledge distillation network that embeds multi-layer attention to emphasize the texture details of visible light images and the prominent targets of infrared images to achieve both infrared and visible light image fusion and super-resolution simultaneously. However, this approach is only effective for the fusion of infrared and visible light images and lacks generalizability to medical images, and it cannot achieve fusion and super-resolution for three modalities.

Overall, although existing methods [9,8,31] have achieved significant accomplishments in improving image quality, enhancing feature extraction, and improving single-modality medical image processing, they still face challenges in several key aspects: (1) Most research focuses on processing dual-modal source images, and there is a relative lack of in-depth studies on image processing problems that cover three modalities. (2) Current methods for medical image processing mostly focus on executing single tasks, lacking strategies for jointly optimizing fusion and super-resolution tasks. (3) Existing technologies that achieve both image fusion and super-resolution have limited generalization capabilities for tri-modal medical images.

To address these challenges, we propose an innovative Tri-modal Conditional Denoising Fusion-Super Resolution Diffusion model (TFS-Diff). To the best of our knowledge, this is the first study to achieve tri-modal medical image fusion and super-resolution tasks synchronously in an end-to-end manner. Our work's main contributions are threefold:

1. The TFS-Diff model synchronously implements end-to-end tri-modal image fusion and super-resolution processing, eliminating the need for manually designing complex fusion and super-resolution network architectures. This significantly simplifies the model design process.
2. We propose a feature fusion module based on a channel attention mechanism that can learn and extract shared features and modality-specific features from different modal medical images.
3. A new fusion and super-resolution loss function is proposed to retain the sharpness, texture and contrast information of the medical images into the fused result. Meanwhile, it guarantees the stability of the model training process and the high quality of the fused results.

2 Method

2.1 Overall Architecture

For multimodal image fusion and super-resolution tasks, we propose TFS-Diff, a method based on a conditional denoising diffusion probability model. This method aims to generate high-resolution fused images that contain rich multimodal information and are highly consistent with the source images. As shown in Fig. 1, TFS-Diff achieves precise tri-modal medical image fusion and its super-resolution through a forward and reverse Markov chain process.

Taking the fusion of MR-T1, MR-T2, and SPECT as an example, let the low-resolution of the three images be denoted as $x \in R^{HW}$, $y \in R^{HW}$, and $s \in R^{HW}$ respectively, and the high-resolution fusion result be represented as $\boldsymbol{I}_0 \in R^{3HW}$. The three modal images x, y, s are input into the model simultaneously, first upsampled to the target resolution through bicubic interpolation sampling of x,

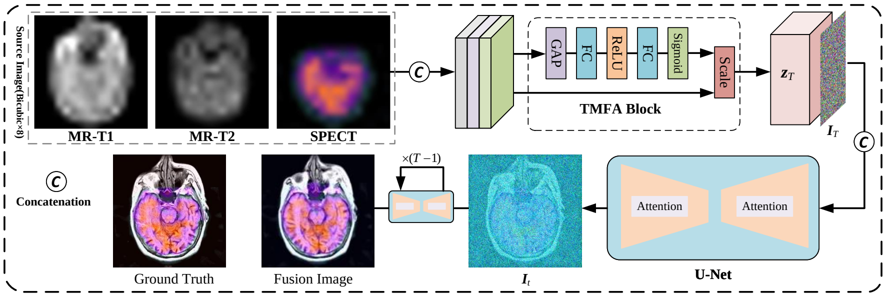

Fig. 1. Implementing super-resolution and fusion model structures synchronously. Taking MR-T1/MR-T2/SPECT fusion as an example, the original images need to be sampled to the specified resolution through bicubic interpolation as the input for the TFS-Diff model, and the output of TFS-Diff is compared with the Ground Truth to calculate the loss.

y, s, and then feature extraction is performed through the TMFA Block (see Section 2.2) to obtain $z_t = \varepsilon(x, y, s)$, which is concatenated on the channel dimension with $\boldsymbol{I}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The objective function optimized by TFS-Diff is:

$$L_{\text{TFS}} := E_{\varepsilon(x,y,s),\epsilon\sim\mathcal{N}(0,I),t}\left[\|\epsilon - \epsilon_\theta(z_t, I_t, \gamma_t)\|_2^2\right] \tag{1}$$

where, γ_t represents the noise variance, and ϵ_θ is the UNet [21] used for noise prediction. z_t is obtained through the TMFA Block.

The backbone of TFS-Diff adopts the U-Net structure from SR3 [22], with z and t as inputs to ϵ_θ. The backbone consists of a contracting path, an expansive path, and a diffusion head. Unlike the U-Net in DDPM [7], TFS-Diff uses residual blocks from BigGAN [2]as connections and incorporates a self-attention mechanism. Both the contracting and expansive paths are comprised of 4 convolutional layers. The diffusion head consists of a single convolutional layer, used for generating predicted noise[30]. Parameters are initialized using the Kaiming initialization method[6].

2.2 Tri-modal Fusion Attention (TMFA) Block

During the process of tri-modal medical image fusion, different modalities provide unique perspectives on anatomical structures, physiological functions, and molecular levels. Existing fusiong methods overlook the complementarity between modalities and the differences in feature importance across different channels. The main purpose of the TMFA Block is to extract deep feature of the concatenated multimodal images before entering the diffusion phase of the fusion super-resolution network. It utilizes a channel attention mechanism to learn the importance weights of each channel, enhancing useful features and suppressing irrelevant information. This approach highlights the most critical information for clinical diagnosis across different modalities.

As shown in Fig.1, the structure of the TMFA Block is based on the classic SE (Squeeze-and-Excitation) block, which has been customized to adapt to trimodal image. Initially, a Global Average Pooling (GAP) layer compresses the features of the input images $C = \text{concatenate}(x, y, s)$ to capture global context information. Subsequently, a bottleneck structure composed of two Fully Connected (FC) layers is introduced to learn the nonlinear relationships between channels, where the ReLU activation function is applied to the first FC layer, and the Sigmoid activation function is applied to the second FC layer, to output the attention weights of the channels. Finally, these attention weights are utilized to adjust the importance of each channel in the original feature image, thus accomplishing feature recalibration to obtain the feature Z_c.

2.3 Fusion super-resolution joint loss function

To make TFS-Diff training converge more stably, a new joint loss design is implementedL_{PSF}, combining Mean Squared Error (MSE) loss and Structural Similarity Index (SSIM) loss. L_{PSF} leverages the advantages of MSE loss in terms of pixel-level reconstruction accuracy, as well as the effectiveness of SSIM loss in maintaining image structural information and enhancing visual quality to optimize both the pixel accuracy and visual quality of the image simultaneously.

$$L_{PSF} = \lambda_1 L_{MSE} + \lambda_2 L_{ssim} \tag{2}$$

$$L_{MSE} = \frac{1}{N} \sum_{i=1}^{N} (P_i - T_i)^2 \tag{3}$$

$$L_{SSIM} = \frac{(2\mu_P \mu_T + c_1)(2\sigma_{PT} + c_2)}{(\mu_P^2 + \mu_T^2 + c_1)(\sigma_P^2 + \sigma_T^2 + c_2)} \tag{4}$$

where $\lambda_1, \lambda_2 \in (0, 1]$ represent the weights of the two losses, respectively.

3 Experiments

3.1 Experimental Detail

The dataset covers five different types of registered medical images, including MR-T2/MR-Gad/PET, CT/MR-T2/SPECT, MR-T1/MR-T2/PET, MR-T2/MR-Gad/SPECT, and MR-T1/MR-T2/SPECT. All source images are from the whole brain atlas database of Harvard Medical School[24], We randomly divided the data into 84, 10 and 25 groups as training set, validation set and test set respectively.The resolution of the training and testing images is 256x256, which was downsampled using bicubic interpolation to construct super-resolution datasets with different magnification levels (8x, 4x, 2x).The Ground Truth is fused by the BitonicX [8].

Five state-of-the-art (sota) methods were used for comparison, including three dual-modal fusion methods: CDDFuse [34], TGFuse [19], DDFM [35], and two tri-modal fusion methods: BitonicX Filtering [8], CTSR [9]. Additionally, the SR3 model [22] was used as the baseline for super-resolution.

The model was optimized using the Adam optimizer with a fixed learning rate of 1e-4 and a diffusion step count T of 4000. The model was trained for 800,000 steps on a computer equipped with four NVIDIA GeForce RTX 3090 GPUs, with a batch size set to 32.

3.2 Objective evaluation metric

In this study, we evaluated the proposed model's performance using several quantitative metrics: Average Gradient (AG) [4], Mean Squared Error (MSE) [27], Visual Information Fidelity (VIF) [5], Structural Similarity Index (SSIM) [26], Peak Signal-to-Noise Ratio (PSNR)[26], Perceptual Image Quality Loss (LPIPS) [32], Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These metrics assess the model's performance in various aspects: AG measures image edge and texture clarity; MSE, MAE, and RMSE gauge pixel-level accuracy; VIF evaluates visual quality; SSIM and PSNR judge structural similarity and noise ratio, indicating image visual effects and fidelity; LPIPS assesses perceptual quality from a deep learning perspective.

3.3 Comparison with SOTA methods

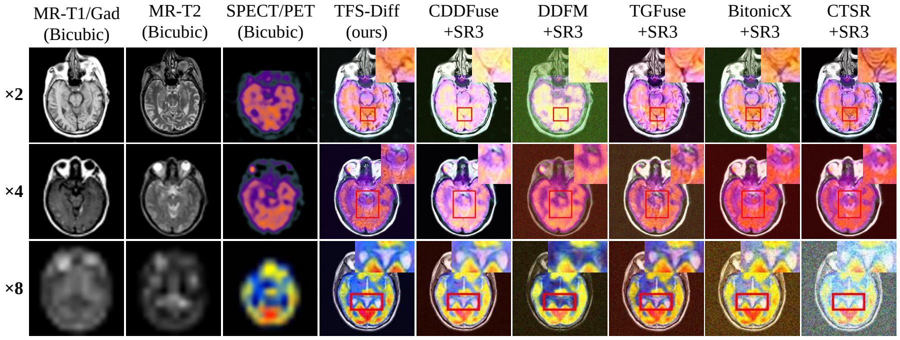

Fig. 2. Tri-modal fusion on the Harvard dataset showed super-resolution results at amplification factors 2, 4, and 8, using MR-T1/MR-T2/SPECT, MR-Gad/MR-T2/SPECT, and MR-T1/MR-T2/PET configurations, respectively.

Table 1 and Fig.2 respectively show the objective evaluation metrics and fusion results of TFS-Diff, CDDFuse, DDFM, TGFuse, BitonicX and CTSR methods. From Table 1, it can be observed that TFS-Diff ranks first in all evaluation metrics, demonstrating its outstanding performance. Specifically, under the conditions of resolutions enlarged by ×8, ×4, and ×2, the methods of TGFuse, DDFM and CDDFuse rank second to TFS-Diff in overall objective evaluation metrics, respectively. Additionally, Fig.2 clearly reveals the problem that the SR3 model, when applied to the DDFM method for super-resolution, fails to completely eliminate noise at three different magnification rates. As the magnification rate increases, the quality of the fused images obtained by the CTSR, TGFuse, and BitonicX methods declines. In contrast, TFS-Diff shows its effectiveness in maintaining the texture and color information of the source images under the three types of magnifications.Due to space constraints, more detailed renderings will be shown in the supplemental documentation

Table 1. The objective results for Harvard dataset (Bold: the best; comparison methods all use SR3 for super-resolution)

Scale	×2							
Metrics	MSE ↓	VIF ↑	SSIM ↑	PSNR ↑	LPIPS ↓	MAE ↓	RMSE ↓	AG ↑
BitonicX	2634.907	0.471	0.579	14.42	0.419	53.95	110.82	8.913
CDDFuse	3519.701	0.501	0.759	12.74	0.339	54.92	120.62	8.237
CTSR	2733.143	0.452	0.611	14.11	0.435	55.74	114.17	9.436
DDFM	2920.907	0.480	0.549	13.50	0.496	56.96	107.37	8.237
TGFuse	3013.091	0.479	0.600	14.13	0.424	54.58	110.61	9.945
Ours	**2021.23**	**0.577**	**0.818**	**15.27**	**0.319**	**44.80**	**103.64**	**9.959**
Scale	×4							
Metrics	MSE ↓	VIF ↑	SSIM ↑	PSNR ↑	LPIPS ↓	MAE ↓	RMSE ↓	AG ↑
BitonicX	2010.889	0.424	0.743	15.26	0.432	49.16	106.78	7.475
CDDFuse	3390.033	0.430	0.671	13.08	0.415	56.96	119.82	7.732
CTSR	2159.267	0.426	0.672	14.99	0.439	49.74	107.16	7.577
DDFM	2511.116	0.458	0.694	13.90	0.489	49.04	99.45	7.082
TGFuse	2213.12	0.448	0.715	15.32	0.415	47.86	102.75	8.109
Ours	**1740.448**	**0.560**	**0.788**	**15.78**	**0.340**	**46.11**	**97.98**	**8.152**
Scale	×8							
Metrics	MSE ↓	VIF ↑	SSIM ↑	PSNR ↑	LPIPS ↓	MAE ↓	RMSE ↓	AG ↑
BitonicX	4413.324	0.394	0.532	12.14	0.521	64.89	124.32	13.558
CDDFuse	5310.433	0.378	0.511	11.21	0.535	70.20	132.68	12.671
CTSR	5480.484	0.369	0.568	11.56	0.549	72.41	135.18	13.612
DDFM	4998.928	0.307	0.466	11.50	0.622	68.10	125.33	12.337
TGFuse	4142.136	0.401	0.566	12.69	0.508	62.77	121.60	13.193
Ours	**1559.803**	**0.579**	**0.980**	**16.50**	**0.299**	**40.68**	**96.42**	**13.77**

3.4 Ablation Study

This section aims to verify the effectiveness and contribution of the TMFA Block and PSF Loss in our TFS-Diff.

1. **w/o TMFA Block**: We removed the TMFA Block to assess its significance in feature extraction and information fusion processes.
2. **w/o PSF Loss**: We replaced PSF Loss with MSE loss to evaluate the importance of PSF Loss in balancing pixel accuracy and visual quality.

Table 2. Ablation Study on the Harvard dataset(Bold: the best).

Harvard dataset(×2)								
Metrics	VIF↑	SSIM↑	PSNR↑	AG↑	MSE↓	LPIPS↓	MAE↓	RMSE↓
w/o TMFA Block	0.468	0.761	14.63	9.529	2777.63	0.362	60.60	120.18
w/o PSF Loss	0.452	0.802	14.76	9.021	2171.51	0.351	55.07	120.81
TFS-Diff	**0.577**	**0.8180**	**15.27**	**9.959**	**2021.23**	**0.319**	**44.80**	**103.64**
Harvard dataset(×4)								
Metrics	VIF↑	SSIM↑	PSNR↑	AG↑	MSE↓	LPIPS↓	MAE↓	RMSE↓
w/o TMFA Block	0.558	0.746	15.32	7.13	1907.22	0.404	56.42	114.96
w/o PSF Loss	0.545	0.728	15.67	7.801	1761.69	0.365	51.42	112.71
TFS-Diff	**0.560**	**0.788**	**15.78**	**8.152**	**1740.44**	**0.340**	**46.10**	**97.98**
Harvard dataset(×8)								
Metrics	VIF↑	SSIM↑	PSNR↑	AG↑	MSE↓	LPIPS↓	MAE↓	RMSE↓
w/o TMFA Block	0.539	0.792	15.40	13.60	1871.20	0.347	45.81	105.15
w/o PSF Loss	0.550	0.569	15.63	12.67	1712.41	0.354	50.82	108.77
TFS-Diff	**0.579**	**0.980**	**16.50**	**13.77**	**1559.80**	**0.299**	**40.68**	**96.42**

As shown in the Table 2, the ablation study results confirmed the significant contribution of the TMFA Block and PSF Loss to enhance the performance of the tri-modal medical image fusion super-resolution model. The combination of these components not only optimized pixel-level reconstruction but also significantly improved the visual quality of the images, thus providing an effective solution for complex medical image processing tasks.

4 Conclusion

This study proposed a conditional diffusion model, TFS-Diff, for tri-modal medical image fusion super-resolution, introducing two key innovations: the TMFA Block and PSF Loss, which ensure the generation accuracy of the diffusion model. Through comprehensive experimental validation, our approach has achieved significant improvements in detail restoration and visual quality compared to existing technologies.

The TMFA block optimized the model's capability to fuse features from different modal medical images, enhancing the efficiency and quality of information integration. Simultaneously, the design of PSF Loss successfully balanced pixel-level accuracy and structural similarity, further enhancing the model's performance in image reconstruction. Ablation results confirmed the significant contribution of these two components to model performance improvement, reflecting their application value in complex medical image processing tasks. Moreover, if TFD-Diff is applied to tri-modal medical image fusion medical instruments, it will enhance the diagnostic efficiency of doctors and reduce the time and money spent by patients.

Our future work will integrate large language models to assist in modeling the diffusion process of TFS-Diff, which is expected to enhance the model's generalizability across various types of medical images.

Acknowledgments. This research was supported in part by the National Natural Science Foundation of China under Grant 62201149, in part by the Basic and Applied Basic Research of Guangdong Province under Grant 2023A1515140077, in part by the Natural Science Foundation of Guangdong Province under Grant 2024A1515011880, in part by the Guangdong Higher Education Innovation and Strengthening of Universities Project under Grant 2023KTSCX127, and in part by the Foshan Key Areas of Scientific and Technological Research Project under Grant 2120001008558, China.

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

References

1. Bhutto, J.A., Tian, L., Du, Q., Sun, Z., Yu, L., Tahir, M.F.: Ct and mri medical image fusion using noise-removal and contrast enhancement scheme with convolutional neural network. Entropy **24**(3), 393 (2022)
2. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
3. Chen, J., Li, X., Luo, L., Ma, J.: Multi-focus image fusion based on multi-scale gradients and image matting. IEEE Transactions on Multimedia **24**, 655–667 (2021)
4. Cui, G., Feng, H., Xu, Z., Li, Q., Chen, Y.: Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. Optics Communications **341**, 199–209 (2015)
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. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
8. Jie, Y., Li, X., Zhou, F., Ye, T.: Tri-modal medical image fusion and denoising based on bitonicx filtering. IEEE Transactions on Instrumentation and Measurement **72**, 1–15 (2023)
9. Jie, Y., Zhou, F., Tan, H., Wang, G., Cheng, X., Li, X.: Tri-modal medical image fusion based on adaptive energy choosing scheme and sparse representation. Measurement **204**, 112038 (2022)
10. Karim, S., Tong, G., Li, J., Qadir, A., Farooq, U., Yu, Y.: Current advances and future perspectives of image fusion: A comprehensive review. Information Fusion **90**, 185–217 (2023)

11. Li, H., Yang, M., Yu, Z.: Joint image fusion and super-resolution for enhanced visualization via semi-coupled discriminative dictionary learning and advantage embedding. Neurocomputing **422**, 62–84 (2021)
12. Li, X., Zhou, F., Tan, H.: Joint image fusion and denoising via three-layer decomposition and sparse representation. Knowledge-Based Systems **224**, 107087 (2021)
13. Li, Y., Sixou, B., Peyrin, F.: A review of the deep learning methods for medical images super resolution problems. Irbm **42**(2), 120–133 (2021)
14. Ma, J., Chen, C., Li, C., Huang, J.: Infrared and visible image fusion via gradient transfer and total variation minimization. Information Fusion **31**, 100–109 (2016)
15. 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)
16. Ma, J., Xu, H., Jiang, J., Mei, X., Zhang, X.P.: Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. IEEE Transactions on Image Processing **29**, 4980–4995 (2020)
17. Ma, J., Yu, W., Liang, P., Li, C., Jiang, J.: Fusiongan: A generative adversarial network for infrared and visible image fusion. Information fusion **48**, 11–26 (2019)
18. Mao, Y., Jiang, L., Chen, X., Li, C.: Disc-diff: Disentangled conditional diffusion model for multi-contrast mri super-resolution. arXiv preprint arXiv:2303.13933 (2023)
19. Rao, D., Xu, T., Wu, X.J.: Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network. IEEE Transactions on Image Processing (2023)
20. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
22. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4713–4726 (2022)
23. Stimpel, B., Syben, C., Schirrmacher, F., Hoelter, P., Dörfler, A., Maier, A.: Multimodal super-resolution with deep guided filtering. In: Bildverarbeitung für die Medizin 2019: Algorithmen–Systeme–Anwendungen. Proceedings des Workshops vom 17. bis 19. März 2019 in Lübeck. pp. 110–115. Springer (2019)
24. Summers, D.: Harvard whole brain atlas: www. med. harvard. edu/aanlib/home. html. Journal of Neurology, Neurosurgery & Psychiatry **74**(3), 288–288 (2003)
25. Tsiligianni, E., Zerva, M., Marivani, I., Deligiannis, N., Kondi, L.: Interpretable deep learning for multimodal super-resolution of medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 421–429. Springer (2021)
26. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
27. Xiang, T., Yan, L., Gao, R.: A fusion algorithm for infrared and visible images based on adaptive dual-channel unit-linking pcnn in nsct domain. Infrared Physics & Technology **69**, 53–61 (2015)

28. Xiao, W., Zhang, Y., Wang, H., Li, F., Jin, H.: Heterogeneous knowledge distillation for simultaneous infrared-visible image fusion and super-resolution. IEEE Transactions on Instrumentation and Measurement **71**, 1–15 (2022)
29. Yin, H., Li, S., Fang, L.: Simultaneous image fusion and super-resolution using sparse representation. Information Fusion **14**(3), 229–240 (2013)
30. Yue, J., Fang, L., Xia, S., Deng, Y., Ma, J.: Dif-fusion: Towards high color fidelity in infrared and visible image fusion with diffusion models. arXiv preprint arXiv:2301.08072 (2023)
31. Zeng, K., Zheng, H., Cai, C., Yang, Y., Zhang, K., Chen, Z.: Simultaneous single- and multi-contrast super-resolution for brain mri images based on a convolutional neural network. Computers in biology and medicine **99**, 133–141 (2018)
32. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
33. Zhang, Y., Wang, M., Xia, X., Sun, D., Zhou, X., Wang, Y., Dai, Q., Jin, M., Liu, L., Huang, G.: Medical image fusion based on quasi-cross bilateral filtering. Biomedical Signal Processing and Control **80**, 104259 (2023)
34. Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., Timofte, R., Van Gool, L.: Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5906–5916 (2023)
35. Zhao, Z., Bai, H., Zhu, Y., Zhang, J., Xu, S., Zhang, Y., Zhang, K., Meng, D., Timofte, R., Van Gool, L.: Ddfm: Denoising diffusion model for multi-modality image fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8082–8093 (October 2023)