

Review

A Review of Deep Learning Methods for Multimodal Medical Image Fusion

Yanan Shi and Maria Rigou * 

Department of Management Science and Technology, University of Patras, 26334 Patras, Greece;
yananshi857@gmail.com

* Correspondence: rigou@upatras.gr

Abstract

Multimodal medical image fusion (MMIF) aims to integrate complementary information from different imaging modalities into a single, more informative image to support clinical diagnosis. Since 2017, a growing number of deep learning-based approaches have been proposed for MMIF, including various network architectures designed to enhance visual quality. However, a comprehensive and up-to-date review of deep learning-based MMIF techniques is still lacking. To fill this gap, this paper provides a comprehensive survey of deep learning-based MMIF methods. First, we categorize MMIF approaches based on deep learning frameworks and conduct an in-depth analysis of loss functions, evaluation metrics, medical imaging modality pairs, and medical datasets. Then, we review the mainstream deep learning-based MMIF methods, including CNNs, Autoencoders, GANs, and Transformers. Subsequently, we introduce emerging deep learning methods, including diffusion models and Mamba-based methods. In addition, we summarize widely used datasets and evaluation metrics, conduct quantitative experiments on representative methods, and propose a unified set of evaluation metrics for standardized comparison. Finally, we identify key research challenges and outline promising future directions for deep learning-based MMIF. This survey aims to provide researchers with a clear understanding of recent progress in deep learning-based MMIF and to facilitate further studies in this area.

Keywords: deep learning; medical image fusion; multimodal fusion; review

1. Introduction

Multimodal medical image fusion (MMIF) has received significant attention in recent years. Owing to the complementary properties of different medical imaging sensors, MMIF has been widely used in disease diagnosis [1,2], surgical navigation [3], and treatment planning [4]. Specifically, computed tomography (CT) images clearly reflect anatomical structures but offer limited soft tissue contrast. Magnetic resonance imaging (MRI) images provide high-resolution soft tissue contrast but are less effective for visualizing calcified tissues. Positron emission tomography (PET) and single-photon emission computed tomography (SPECT) images provide metabolic and functional information, but their low spatial resolution limits anatomical detail. MMIF aims to fuse complementary information from different imaging modalities into a single image with more comprehensive information, thereby improving diagnostic efficiency and accuracy.

Before the emergence of deep learning-based methods, various traditional methods had been proposed for MMIF. Traditional MMIF methods can generally be divided into



Academic Editor: Adrian Barbu

Received: 2 May 2026

Revised: 21 June 2026

Accepted: 4 July 2026

Published: 22 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

three categories, including multiscale transform-based approaches [5–8], sparse representation techniques [9–11], and hybrid methods [12,13]. Although traditional image fusion methods exhibit stable fusion performance, activity level measurements and fusion rules need to be designed in a handcrafted manner, which limits their adaptability to complex fusion scenarios.

In recent years, deep learning has demonstrated remarkable performance in MMIF due to its superior feature representation ability. Liu et al. [14] were the first to use deep learning in MMIF, introducing a siamese convolutional neural network (CNN) to learn a mapping from source images to a weight map. Since then, research on deep learning-based MMIF methods has grown rapidly. Figure 1 presents a timeline of deep learning-based MMIF approaches, highlighting major milestones in the field. These approaches can be classified into CNN-based, GAN-based, Autoencoder-based, Transformer-based, diffusion model-based, and Mamba-based methods.

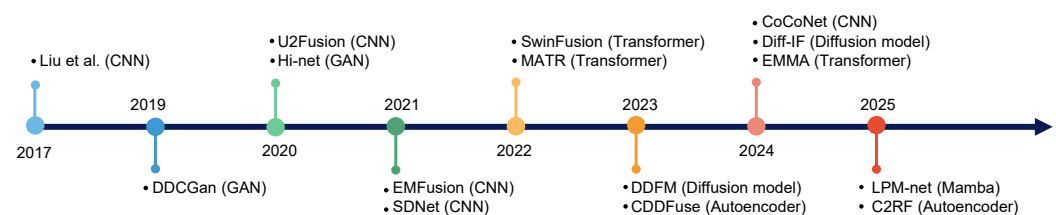


Figure 1. Milestones in deep learning-based MMIF methods from 2017 to 2025. Parentheses indicate the architecture each method is based on. CNN-based methods: Liu et al. [14], U2Fusion [15], EMFusion [4], SDNet [16], and CoCoNet [17]. GAN-based methods: DDcGAN [18] and Hi-net [19]. Transformer-based methods: SwinFusion [20], MATR [21], and EMMA [22]. AE-based methods: CDDFuse [23] and C2RF [24]. Diffusion model-based methods: DDFM [25] and Diff-IF [26]. Mamba-based method: LPM-net [27].

However, the current literature lacks a comprehensive review focusing on deep learning-based MMIF, which limits researchers' understanding of recent advances and emerging trends, particularly for those new to the field. Several review papers have been published in the field of MMIF. Hermessi et al. [28], Azam et al. [29], and Ullah Khan et al. [30] covered both traditional and deep learning-based MMIF methods, but provided only limited coverage of the latter. Zhou et al. [31] categorized deep learning-based MMIF methods, but their review did not cover studies published from 2022 to 2025.

To address the limitations of existing surveys, this review provides a comprehensive overview of deep learning-based multimodal medical image fusion methods published from 2017 to 2025. The main contributions are summarized as follows:

1. We establish a taxonomy covering CNNs, Autoencoders, GANs, Transformers, diffusion models, and Mamba models. Based on this taxonomy, we analyze 49 representative methods in terms of architectural designs, loss functions, evaluation metrics, modality combinations, and datasets.
2. We recommend a unified evaluation metric set comprising VIF, MI, SSIM, SCD, and SD. These metrics assess visual fidelity, information preservation, structural similarity, correlation between source and fused images, and image feature characteristics, respectively. We further compare nine representative methods from five architectural categories using the same dataset and evaluation protocol.
3. We identify the major limitations of current MMIF research and discuss future directions, including improved evaluation metrics, registration-aware fusion, Mamba-based methods, benchmark expansion, application-driven fusion, generalizable MMIF models, and clinical translation.

The overall framework of this paper is illustrated in Figure 2. Section 2 introduces the taxonomy of deep learning-based MMIF methods. Sections 3 and 4 review mainstream and emerging approaches, respectively, while Section 5 summarizes datasets and evaluation metrics, presents quantitative experiments on representative methods, and proposes a unified set of evaluation metrics. Section 6 discusses future research directions, and Section 7 concludes the paper.

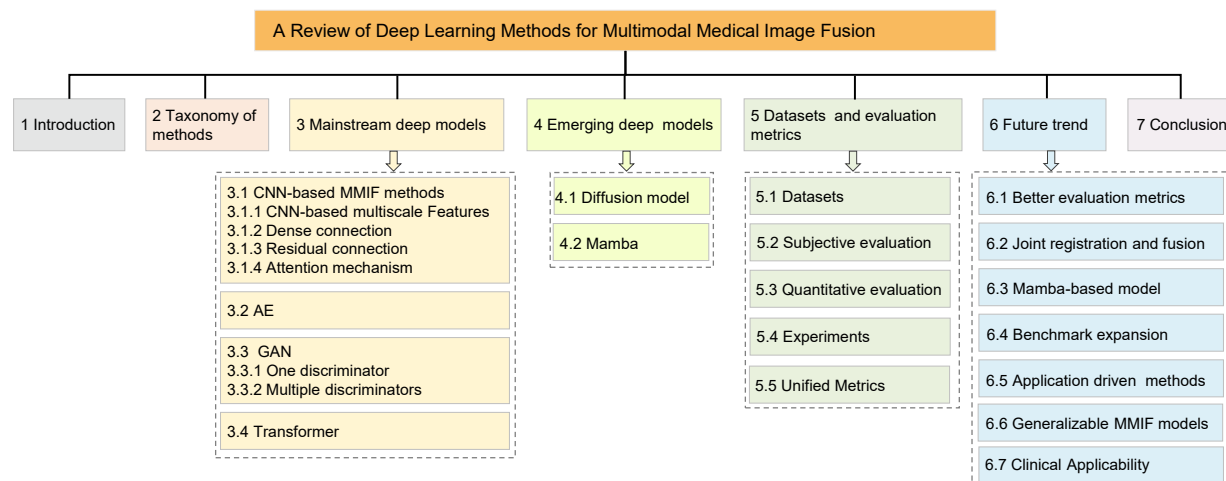


Figure 2. Overall framework of the review.

2. Deep Learning-Based MMIF Method Classification and Fusion Metric Analysis

A typical deep learning-based MMIF pipeline consists of three stages: feature extraction, feature fusion, and image reconstruction. In the first stage, features are extracted from the medical images. These features are then fused in the second stage. Finally, the fused image is reconstructed in the final stage.

In this review, deep learning-based MMIF methods are categorized into two major groups: mainstream and emerging methods. The mainstream approaches are further classified into four subcategories according to their backbone architectures: CNN-based, GAN-based, Autoencoder-based, and Transformer-based methods. The emerging deep learning-based MMIF methods include diffusion model-based and Mamba-based methods. A summary of representative MMIF methods is presented in Table 1.

We present the key trends identified from the review of 49 selected papers in the field of MMIF. Our goal is to provide a comprehensive overview of the current research landscape. We analyze the frequencies of loss functions, evaluation metrics, modality pairs, and datasets used in these studies, with each frequency calculated as the number of papers in which each item appears divided by the total number of papers (i.e., 49).

Loss functions: In deep learning-based image fusion, the choice of loss function is critical, as it directly influences the optimization of fusion performance. The six most commonly used loss functions are SSIM, intensity, MSE, gradient, content, and adversarial losses, as shown in Figure 3a.

Evaluation metrics: Choosing appropriate objective evaluation metrics is essential for the assessment of medical image fusion, supporting practical deployment and promoting scientific progress in the field. There is no standardized evaluation metric for MMIF, which leads to the use of different metrics across studies. Figure 3b shows the top ten most frequently used evaluation metrics, which are VIF, MI, SSIM, EN, $Q^{AB/F}$, SCD, SD, SF, PSNR, and CC.

Table 1. A summary of key deep learning-based MMIF methods.

Reference	Year	Backbone	Loss Function	Modalities	General /MMIF	Metrics	Medical Dataset
CoCoNet [17]	2024	CNN	SSIM loss, Intensity similarity loss, Self-adaptive loss, Contrastive loss	MRI-PET	General	EN, SD, VIF, SF, AG, SCD	AANLIB
MM-net [32]	2024	CNN	Spatial domain loss, Frequency domain loss	MRI-SPECT	MMIF	MI, NICE, Q_G , Q_P , Q_C , Q_Y , Q_{CB} , Q_{CV}	AANLIB
MRSCFusion [33]	2023	CNN	Content loss, Intensity loss	MRI-CT/PET/SPECT	MMIF	SCD, MS-SSIM, $Q^{AB/F}$, FMI, SF, VIF	AANLIB
SMAFusion [34]	2025	CNN	MSE loss, SSIM loss, Total variation loss	MRI-CT	MMIF	SD, SF, MI, SCD, VIF, SSIM, $Q^{AB/F}$	AANLIB
FDGNet [35]	2023	CNN	Weighted fidelity loss, Feature difference loss	MRI-CT/PET/SPECT	MMIF	NMI, $Q^{AB/F}$, Q_Y , VIFP	AANLIB
Fu et al. [36]	2021	CNN	MSE loss	MRI-CT/PET/SPECT	MMIF	SSIM, MI, PSNR, Q_{AC} , TMQI, FSIM	AANLIB
MdAFuse [37]	2024	CNN	SSIM loss, Pixel loss	MRI-CT/PET/SPECT	MMIF	EN, MI, SD, VIF, SCD, SSIM, CC, MSE, rSFE	AANLIB, ADNI
MsgFusion [38]	2024	CNN	SSIM loss, Pixel loss	MRI-CT/PET/SPECT	MMIF	EN, SD, MI, RSFE, SM, VIF, RQ	AANLIB, ADNI
U2Fusion [15]	2020	CNN	SSIM loss, MSE loss, EWC loss	MRI-PET	General	SCD, CC, SSIM, PSNR	AANLIB
SDNet [16]	2021	CNN	Gradient loss, Intensity loss	MRI-PET	General	EN, FMI, MG, PSNR	AANLIB
EMFusion [4]	2021	CNN	Surface-level loss, Deep-level loss, Similarity constraints loss	MRI-CT/PET/SPECT	MMIF	SSIM, PSNR, CC, VIF	AANLIB
MUFusion [39]	2023	CNN	Content loss, Memory loss	MRI-PET	General	EN, VIF, MI, EI, $Q^{AB/F}$	AANLIB
Liu et al. [40]	2021	CNN	Structure loss, Intensity loss	MRI-CT/PET/SPECT	General	VIF, AG, EN, SF, SCD, $Q^{AB/F}$	Brain Atlas
Liu et al. [41]	2022	CNN	Structure loss, Intensity loss	MRI-MRI	MMIF	LMI, Q_Y , VIF, $Q^{AB/F}$	AANLIB
MSAIF-net [42]	2023	CNN	Content loss, Texture loss	MRI-MRI	MMIF	CC, SSIM, Q_Y , Q_W , $Q^{AB/F}$, VIF, Q_P , Q_{CB}	BRATS, ISLES
GeSeNet [43]	2023	CNN	Edge loss, SSIM loss, Semantic loss	MRI-CT/PET/SPECT	MMIF	SSIM, SD, MI, VIF, SCD, $Q^{AB/F}$	AANLIB
ECINFusion [44]	2025	CNN	Structure loss, Semantic loss, Pixel loss	MRI-CT/PET/SPECT	MMIF	NCIE, MI, Q_G , LMI, MS-SSIM, VIF, NCC, PSNR, Cosin	AANLIB
AMMNet [45]	2024	CNN	Fusion loss, Decomposition loss	MRI-CT/PET	MMIF	AG, EN, SF, MI, SD, CC, $Q^{AB/F}$	AANLIB
IFSepR [46]	2023	AE	SSIM loss, MSE loss	MRI-MRI	General	EN, MG, VIF, N_{abf}	AANLIB
UNIFusion [47]	2021	AE	SSIM loss, MSE loss, Gradient loss	MRI-PET	General	MI, SD, FMI, NCIE, Q_P , VIF	AANLIB, MS-COCO
SGFusion [48]	2023	AE	SSIM loss, Mean absolute error loss	MRI-PET	General	EN, MI, MESSIM, SCD	AANLIB, ECSSD
CDDFuse [23]	2023	AE	Gradient loss, Intensity loss, Decomposition loss, SSIM loss	MRI-CT/SPECT	General	EN, SD, SF, MIF, SCD, VIF, $Q^{AB/F}$, SSIM	AANLIB
C2RF [24]	2025	AE	Fusion loss, Decomposition loss, Contrastive Learning-driven Registration loss	MRI-PET	General	MI, CC, SSIM, MSE, MEE, MAE	AANLIB

Table 1. Cont.

Reference	Year	Backbone	Loss Function	Modalities	General /MMIF	Metrics	Medical Dataset
HI-net [19]	2020	GAN	Reconstruction loss	MRI-MRI	MMIF	PSNR, SSIM, NRMSE	BRATS
BMGAN [49]	2022	GAN	Adversarial loss, KL-divergence constraint, Reconstruction loss, Perceptual loss	MRI-PET	MMIF	MAE, PSNR, MS-SSIM, FID	ADNI
MGM-GAN [50]	2022	GAN	Adversarial loss, Pixel loss, Gradient loss	MRI-MRI	MMIF	PSNR, SSIM, NRMSE	BRATS
PRRGAN [51]	2024	GAN	SSIM loss, Adversarial loss, Discriminator loss	CT-MRI-SPECT, MRI-MRI-PET	MMIF	EN, SSIM, SD, PSNR, Q_{CB} , Q_G	ADNI
DDCGAN [18]	2019	GAN	Adversarial loss, Content loss	MRI-PET	General	EN, MG, SF, SD, PSNR, CC, SSIM, VIF	AANLIB
Zhou et al. [52]	2022	GAN	Content loss, Adversarial loss, Intensity loss, Gradient loss	MRI-PET	Gneral	EN, CC, SD, SCD, Q_{CB}	AANLIB
TCGAN [53]	2023	GAN	Content loss, Adversarial loss, SSIM loss	MRI-PET	Gneral	SSIM, PSNR, CC, SF, VIF	AANLIB
Liu et al. [1]	2022	GAN	Content loss, Adversarial loss, Discriminator loss	MRI-MRI	MMIF	MI, Q_G , Q_W , VIF	BRATS
MHW-GAN [54]	2024	GAN	Adversarial loss, Content loss, Edge detail loss, Constraint loss, Similarity loss	MRI-PET/SPECT	Gneral	Q_W , Q_C , SSIM, N_{abf}	BRATS
MATR [21]	2022	Transformer	SSIM loss, Region-level loss	MRI-PET, GFP-PC	MMIF	MI, TE, NCIE, Q_G , Q_P , Q_{CV} , LMI, VIF, MS-SSIM	AANLIB
MACTFusion [55]	2025	Transformer	SSIM loss, Gradient loss, Intensity loss	MRI-CT/PET/SPECT	MMIF	SCD, MS-SSIM, $Q^{AB/F}$, SD, Q_W , VIF	AANLIB
FATFusion [56]	2024	Transformer	Pixel loss, Total variation loss	MRI-PET/SPECT	MMIF	MI, Q_G , Q_P , Q_Y , Q_{CB} , VIF, $M_{I_{abf}}$	AANLIB
Cheng et al. [57]	2025	Transformer	Public loss, Private loss	MRI-PET	General	EI, VIF, SCD, AG	AANLIB
EMMA [22]	2024	Transformer	L2 loss, Sensing loss, Equivariant loss	MRI-PET	General	EN, VIF, SF, AG, SD, SCD	AANLIB
SwinFusion [20]	2022	Transformer	SSIM loss, Texture loss, Intensity loss	MRI-CT/PET	General	FMI, $Q^{AB/F}$, SSIM, PSNR	AANLIB
Wang et al. [58]	2024	Transformer	MFF loss, MEF loss, Supervised loss, Unsupervised loss	MRI-CT/PET	General	EN, MI, MS-SSIM, $Q^{AB/F}$, FMI, NMI, Q_{CB} , N_{abf}	AANLIB
DDBFusion [59]	2025	Transformer	SSIM loss, Pixel loss, Decomposition loss	MRI-PET	General	EI, SCD, CC, SF, EN	AANLIB
LFDT-fusion [60]	2025	Transformer	Diffusion loss, Fusion loss	MRI-PET/SPECT	General	VIF, EN, $Q^{AB/F}$, SSIM, FMI	AANLIB
DDFM [25]	2023	Diffusion model	L1 loss	MRI-CT/PET/SPECT	General	EN, SD, MI, VIF, $Q^{AB/F}$, SSIM	AANLIB
Diff-IF [26]	2024	Diffusion model	Diffusion loss, SSIM loss, Gradient loss, Intensity loss	MRI-CT	General	MI, SCD, VIF, $Q^{AB/F}$, SSIM	AANLIB
FusionDiff [61]	2024	Diffusion model	Mean loss	MRI-CT	General	Q_{cy} , EN, SF, PSNR, MI, PC	AANLIB
UUD-Fusion [62]	2024	Diffusion model	SSIM loss, Smooth absolution loss	MRI-PET	General	EI, AVG, SD, SF, VIF, EN, MI, $Q^{AB/F}$	AANLIB

Table 1. Cont.

Reference	Year	Backbone	Loss Function	Modalities	General /MMIF	Metrics	Medical Dataset
LPM-net [27]	2025	Mamba	SSIM loss, MSE loss, Gradient loss, Reconstruction loss	CT-MRI/PET	MMIF	SSIM, MI, SCD, VIF, $Q^{AB/F}$	BRATS, IXI, RESECT, MMWHS, Prostate, HNSCC
Wu et al [63]	2025	Mamba	MSE loss, Decomposition loss, TV loss	MRI-CT/PET/SPECT	General	EN, SF, AG, SD, VIF, SCD, SSIM, MSE, $Q^{AB/F}$	AANLIB
ISCDFuse [64]	2025	Mamba	SSIM loss, Texture loss, Intensity loss, Decomposition loss, MSE loss	MRI-CT/PET/SPECT	General	SSIM, VIF, SCD, MI, SF, SD, EN, AG, Q_W , MS-SSIM, $Q^{AB/F}$	AANLIB
DWMFusion [65]	2025	Mamba	Wavelet loss, Contrastive loss, SSIM loss	MRI-CT/PET/SPECT	MMIF	EN, SD, MI, SF, SCD, VIF, SSIM, AG, $Q^{AB/F}$	AANLIB

Medical imaging modality pairs: The MRI-PET modality pair is dominant in imaging techniques, with 73% of the 49 papers utilizing this combination. The next most common pairs are MRI-CT, followed by MRI-SPECT and MRI-MRI. The “Other” category includes less frequent modality pairs, such as CT-PET and GFP-PC, as shown in Figure 3c.

Medical datasets: The most commonly used datasets are AANLIB (82%), BRATS (12%), and ADNI (8%). The “Other” category also includes less frequently used datasets, such as Brain Atlas, ISLES, and ECSSD, as illustrated in Figure 3d.

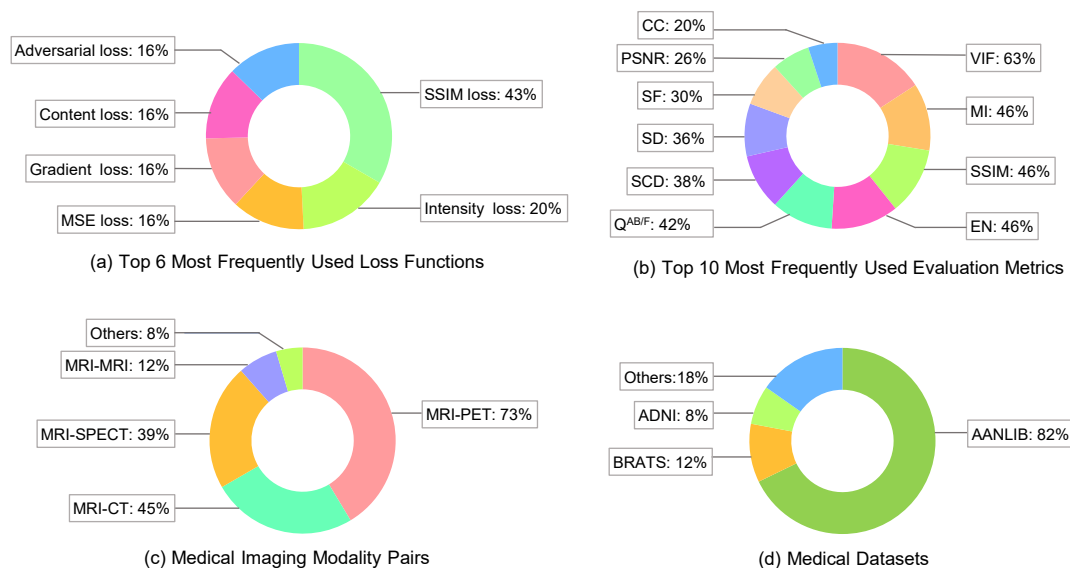


Figure 3. Distribution of loss functions, evaluation metrics, imaging modality pairs, and datasets in the reviewed papers listed in Table 1. The percentages denote the frequency of use of each item among these papers.

3. Mainstream Deep Learning-Based MMIF Methods

3.1. CNN-Based MMIF Methods

CNNs have emerged as one of the most widely adopted architectures for improving performance in MMIF. Liu et al. [14] first introduced a CNN-based MMIF method. Since then, a wide variety of CNN-based methods have emerged. These methods can generally be categorized into four groups: multiscale feature mechanisms, dense connection-based approaches, residual connection-based methods, and attention mechanisms.

3.1.1. CNN-Based Multiscale Feature Mechanisms

CNN-based multiscale feature mechanisms are widely applied in MMIF. These mechanisms employ convolutional operations with different receptive field sizes or spatial downsampling to extract features at multiple scales. Based on the implementation strategy, multiscale mechanisms for MMIF can be grouped into three types: downsampling-based mechanisms, convolutional kernel-based mechanisms, and hybrid mechanisms that combine both downsampling and multi-kernel convolutions.

Downsampling-based multiscale feature mechanisms: In MMIF methods, downsampling-based multiscale mechanisms typically involve the following steps: First, the source images are fed into a multi-branch feature extractor, where each branch performs downsampling through pooling layers or strided convolutions (with stride > 1) to extract features at multiple scales. Subsequently, the features at the same scale are fused. Finally, the fused features are transmitted to the corresponding scales in the image reconstruction layer via skip connections, where upsampling operations are performed to reconstruct the final image, as shown in Figure 4. The downsampling-based multiscale mechanism utilizes both

low-level fine details from shallow networks and high-level semantic features from deep networks, effectively enhancing the fusion quality [66].

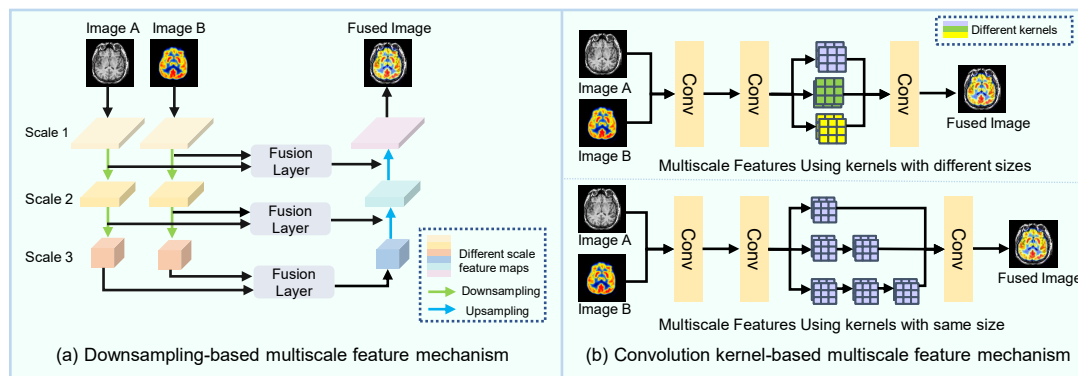


Figure 4. Convolution kernel-based multiscale feature mechanism for MMIF.

Several representative MMIF methods have adopted this downsampling-based multiscale strategy [17,32,33]. For instance, CoCoNet [17] proposes a contrastive learning network that incorporates multiscale features. Its feature extractor consists of three branches: one leverages max pooling to obtain multiscale features, while the other two employ VGG19 [67] as the backbone to extract features at different scales. MM-net [32] adopts a dual-branch multiscale feature extraction strategy based on MixFormer [68], in which each branch consists of five stages with different downsampling rates to extract local and global features across scales. Similarly, MRSCFusion [33] uses max-pooling-based downsampling to extract multiscale deep features from stacked convolutional blocks.

Convolution kernel-based multiscale feature mechanism: An effective method for obtaining multiscale features is to employ convolution kernels of varying sizes, as illustrated in Figure 4. For example, Tang et al. [24] present C2RF, an AE-based framework for multimodal image registration and fusion that employs three multiscale residual blocks to extract shallow features in the feature encoder. The multiscale residual block contains three branches, with convolution kernels of sizes 5×5 , 3×3 , and 1×1 . Similarly, Guo et al. [34] propose SMAFusion, which employs a multiscale feature encoder to capture both local and global features across different scales.

Another way of achieving multiscale features is to use same-size convolution kernels, because stacking two 3×3 convolutions approximates the receptive field of a 5×5 kernel, while stacking three approximates that of a 7×7 kernel [36], as illustrated in Figure 4. These approaches can achieve comparable performance with fewer parameters, making the network more efficient. For example, MATR [21] introduces a three-branch multiscale feature extractor based on adaptive convolutions and transformers. The first branch employs a single 3×3 adaptive convolution, while the second and third branches stack two and three such modules, respectively, to approximate the receptive fields of 5×5 and 7×7 kernels. Recently, FDGNet [35] introduces a feature difference guided network that consists of two parallel branches of stacked 3×3 convolutions, in which feature differences are computed at each scale and transmitted to deeper layers through skip connections.

Hybrid multiscale features mechanism: Some methods combine downsampling with convolutional kernels to effectively capture multiscale features [36–38], as shown in Figure 5. For instance, Fu et al. [36] propose a multiscale residual pyramid attention network, which performs downsampling at multiple spatial resolutions and applies parallel branches of stacked 3×3 convolutions to approximate 5×5 and 7×7 receptive fields. MdAFuse [37] captures multiscale representations by combining downsampling operations with convolutional layers of varying kernel sizes, including 1×1 , 3×3 , 5×5 , and 7×7 .

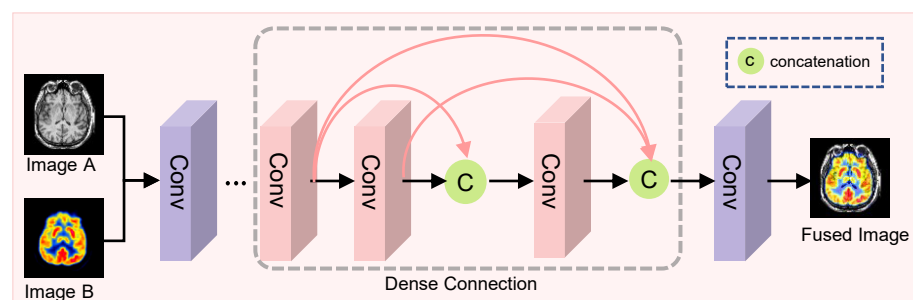


Figure 5. Image Fusion Methods Based on Dense Connection. All arrows indicate the direction of information flow.

3.1.2. Dense Connection-Based MMIF Methods

Dense connection-based MMIF methods refer to architectures in which feature maps from each convolutional layer are passed to subsequent layers. Each layer receives concatenated inputs from preceding layers, as illustrated in Figure 5. By fully exploiting intermediate-layer features, such methods enhance feature propagation, promote feature reuse, and mitigate gradient vanishing [4,69].

Owing to these advantages, many MMIF methods based on dense connections have been proposed. For instance, Xu et al. [15] develop U2Fusion, which employs a DenseNet module to generate the fused image. The architecture of the DenseNet comprises multiple densely connected blocks to improve fusion performance. Zhang et al. [16] propose SDNet, which consists of a squeeze network and a decomposition network. Dense connections within the squeeze network are employed to reduce information loss. EMFusion [4] applies dense connections in the FusionNet module to create short connections among the first four convolutional layers. MUFusion [39] uses two densely connected feature extraction blocks to capture deep features from the input images.

3.1.3. Residual Connection-Based MMIF Methods

Residual connections are highly effective in deep learning-based MMIF methods. They utilize skip connections to directly add the input features to the output of subsequent convolutional layers as the final feature output, as illustrated in Figure 6. Residual connections in image fusion facilitate the utilization of multi-level information, enhance feature representation, and mitigate gradient vanishing.

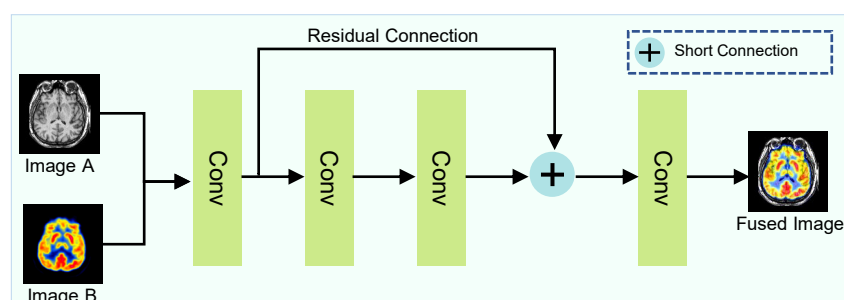


Figure 6. Image Fusion Methods Based on Residual Connection.

Residual connections can be employed in different stages of the fusion process. For example, Liu et al. [41] employ a residual block to process the fused features in the image reconstruction stage. Liu et al. [40] propose an adaptive fusion strategy based on adaptive integration weight maps, utilizing residual blocks to construct the base weight map integration. Fu et al. [36] introduce a multiscale residual pyramid attention module for feature extraction, which uses a residual attention mechanism to enhance training efficiency and enhance feature representation quality.

3.1.4. Attention Mechanism-Based MMIF Methods

Attention mechanisms have been widely used in CNN-based MMIF methods. The principle of attention mechanisms is to dynamically assign varying weights to different elements of the feature maps, enabling the network to focus on the most informative elements and thus improve performance in MMIF. Attention mechanisms in CNN-based MMIF are generally categorized into spatial attention [41–43] and channel attention [23,44,45], as shown in Figure 7.

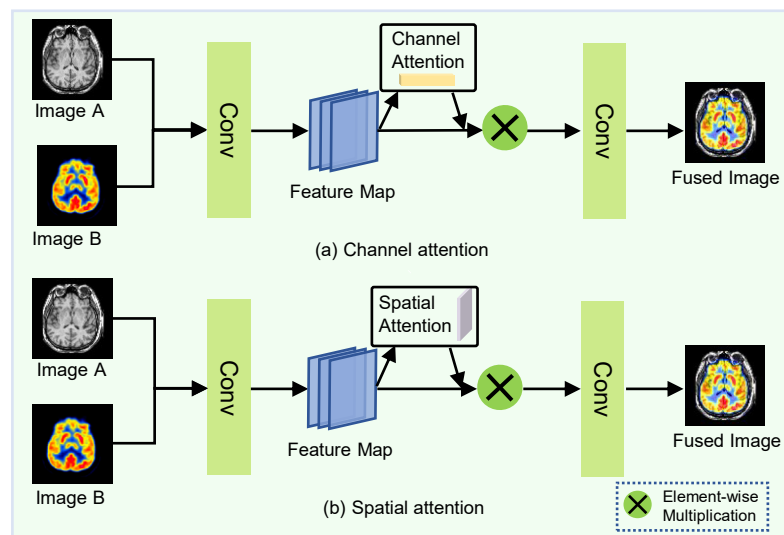


Figure 7. Attention Mechanisms in CNN-based MMIF.

Attention mechanisms can be employed at different stages of MMIF approaches. In some methods, attention mechanisms are used for feature extraction. For instance, MSAIF-net [42] introduces a multistage spatial attention mechanism that captures long-range contextual information from each image and adaptively assigns attention weights at different stages within the feature extraction module. In other approaches, attention mechanisms are used for feature fusion. For instance, Liu et al. [41] propose an attention-based feature fusion module to enhance feature learning, where parallel spatial and channel attention branches jointly contribute to the fusion process. Di et al. [45] also propose an AE-based MMIF method that incorporates spatial and channel attention mechanisms to guide the feature fusion process more effectively.

CNN-based models exhibit strong local feature extraction capability due to inherent inductive biases such as locality and translational invariance. In terms of computational complexity, convolution operations are generally efficient and highly parallelizable, and their computational cost increases mainly with image resolution, kernel size, and channel number, making CNN-based models relatively practical for image fusion tasks. Nevertheless, the limited receptive field of convolution operations restricts long-range dependency modeling and global context aggregation, which may reduce their effectiveness in capturing complex structural relationships for image fusion.

3.2. AE-Based MMIF Methods

AE-based MMIF methods typically involve two stages. In the first stage, an autoencoder is trained on a medical image dataset such as the Harvard Whole Brain Atlas or on a natural image dataset such as the Microsoft Common Objects in Context (MS COCO). In the second stage, the encoder and decoder previously trained are employed for feature extraction and image reconstruction, respectively. The feature fusion between the encoder and

decoder is achieved either through hand-crafted fusion strategies or a secondary training step. An overview of these stages is illustrated in Figure 8.

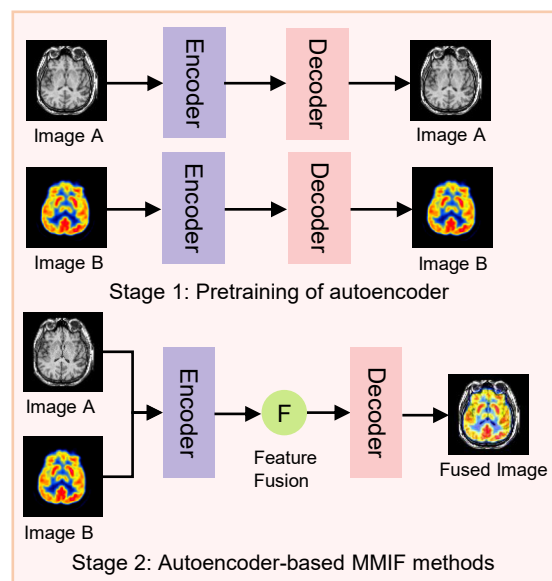


Figure 8. Image Fusion Methods Based on Autoencoder.

IFSepR [46] proposes a general three-stream AE-based method to disentangle common and private features. A feature-space adaptive fusion rule is utilized for feature fusion to preserve complementary information. UNIFusion [47] introduces a fusion strategy that integrates features using averaging, maximum selection, spatial attention, and gradient perception. SGFusion [48] utilizes an addition strategy to integrate the features extracted during the encoding of the source images.

However, the above-mentioned methods adopt handcrafted fusion strategies, which may be limited in performance. To overcome these limitations, some methods propose an additional learning stage specifically designed for feature fusion [23,24]. For example, CDDFuse [23] designs a fusion layer that combines transformer-based global components with invertible neural network (INN)-based local components. In C2RF [24], the extracted common features are fused through convolutional feature concatenation, while the unique features are integrated using a pixel-wise attention mechanism.

Although AE-based MMIF methods have shown promising results, their performance is often constrained by the limited availability of medical image datasets. Some studies address this issue by pretraining autoencoders on large-scale natural image datasets and then fine-tuning them on medical datasets. Nevertheless, the structural and textural differences between natural and medical images may weaken cross-domain transferability. Therefore, constructing large-scale, well-aligned multimodal medical image datasets is essential for improving the generalization and reliability of AE-based MMIF methods.

3.3. GAN-Based MMIF Methods

GANs have been widely adopted in computer vision tasks since their introduction in 2014 [70]. GAN-based MMIF methods aim to learn the data distribution of medical images through adversarial learning. Typically, a GAN-based MMIF method consists of two modules: the generator aims to produce fused images to deceive the discriminator, while the discriminator is responsible for distinguishing real from generated data. Both modules are continuously optimized through adversarial training, enabling the model to progressively improve the quality of the fused medical images. Existing GAN-based MMIF

methods can be divided into two categories: one generator and one discriminator, and one generator and multiple discriminators, as illustrated in Figure 9.

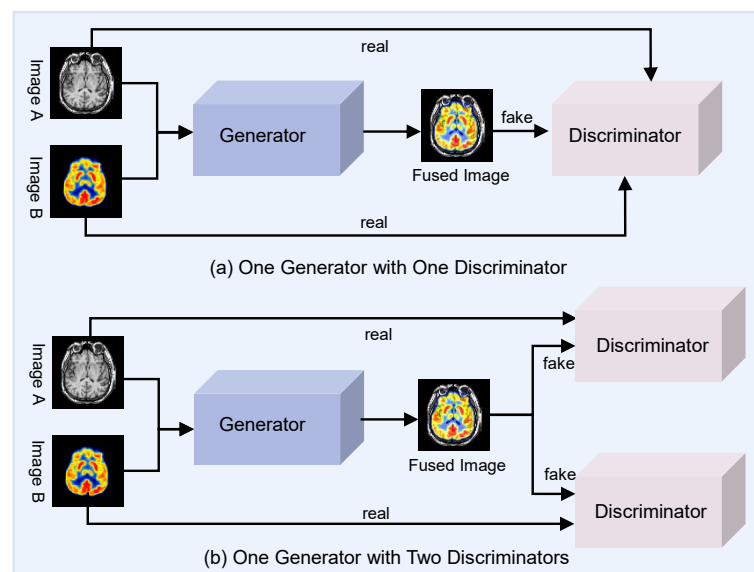


Figure 9. Architecture of GAN-Based Image Fusion Methods.

3.3.1. One Generator and One Discriminator

Zhou et al. [19] propose Hi-Net, which consists of three components: a modality-specific network, a fusion network, and a synthesis network. The synthesis network contains a generator responsible for generating fused outputs and a discriminator that distinguishes generated images from real images. BMGAN [49] introduces a 3D end-to-end network, where the generator incorporates a bi-directional mapping mechanism and a 3D Dense-UNet to enhance the visual fidelity of synthetic images. This design effectively mitigates the limitations of conventional GANs in encoding meaningful latent representations and maintaining structural integrity in the fused output. MGM-GAN [50] constructs its generator based on multiscale convolutional modules and a gating mechanism to integrate more information into the fused image. Recently, PRRGAN [51] introduces a tri-modal generative adversarial network for MMIF, in which the generator is composed of multiscale residual modules and inference attention blocks.

3.3.2. One Generator and Multiple Discriminators

A major limitation of one-discriminator GANs is their tendency to generate fused images biased toward one modality, which may cause the loss of complementary information from the other modality and lead to modality imbalance. To address this problem, some researchers have proposed GANs based on two discriminators [1,18,52,53]. Ma et al. [18] introduce DDcGAN, which adopts two discriminators to discriminate the structural differences between the fused image and the source images, and integrates a content loss to guide optimization. Zhou et al. [52] propose a generator with an encoder-decoder architecture and a dual-discriminator to evaluate gradient and intensity consistency in the fused image. Gradient-based and intensity-based loss functions are also introduced to further guide the training process. TCGAN [53] embeds a wavelet transform-based pre-fusion module into the GAN framework to introduce domain-specific priors. The model employs a multiscale CNN–Transformer hybrid as the generator and utilizes dual Transformer-based discriminators to capture complementary local and global features. Departing from conventional approaches that adopt classification networks as discriminators, Liu et al. [1] employ a

semantic segmentation network to distinguish between fused and source images, enabling more fine-grained evaluation.

In addition to the aforementioned dual-discriminator approaches, MHW-GAN [54] adopts a GAN architecture comprising one generator and three discriminators for general image fusion tasks. Besides the two discriminators corresponding to the two source images, a third discriminator is introduced to evaluate the consistency of the fused edge image with the joint edge representation.

GANs with multiple discriminators can better preserve the distinctive features of each source image and improve fusion performance. However, the use of multiple discriminators increases the training cost by introducing additional parameters, adversarial losses, and optimization steps, although the inference cost mainly depends on the generator. In addition, training such models remains challenging due to instability and mode collapse.

3.4. Transformer-Based MMIF Methods

Transformers leverage the self-attention mechanism to effectively capture long-range dependencies and global semantic information, demonstrating remarkable success in natural language processing and computer vision tasks. Transformer models were first introduced to the field of image fusion in 2021. Since then, many Transformer-based MMIF methods have been proposed. Transformers are usually integrated with CNNs in image fusion to leverage their complementary advantages: CNNs are effective in extracting local spatial features, while transformers specialize in modeling long-range dependencies. An overview of a typical hybrid CNN-Transformer architecture is illustrated in Figure 10. Transformers are often applied to different stages of MMIF methods, including feature extraction, fusion, and reconstruction.

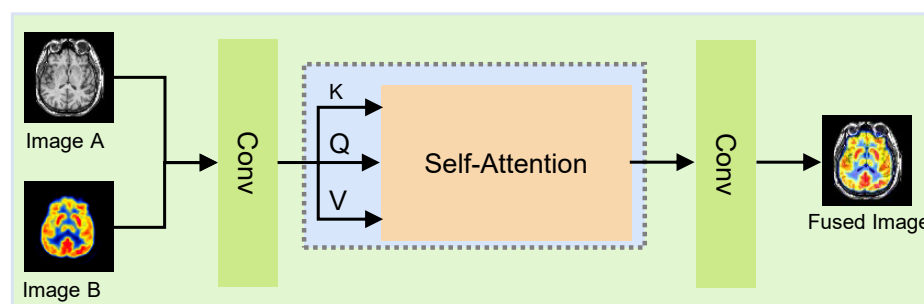


Figure 10. Architecture of Transformer-based image fusion methods. K, Q, and V denote Key, Query, and Value, respectively.

Transformers for feature extraction: In some MMIF methods, Transformers are applied to feature extraction [21,23,32,59]. For example, CDDFuse [23] proposes a dual-branch Transformer-CNN architecture in the feature extraction module. Specifically, both Restormer [71] and Lite transformer [72] are used to extract shared features and private features. MATR [21] integrates adaptive convolutions and adaptive transformers to enhance semantic extraction capability. MM-Net [32] proposes a multiscale fusion method based on MixFormer [68], which effectively captures both local features and global contextual information across multiple scales, thereby improving feature representation quality.

Transformers for feature fusion: In addition to feature extraction, Transformers have also been employed in the fusion stage of MMIF frameworks to effectively integrate information from different modalities [33,55–57]. For example, MRSCFusion [33] employs a dual-branch fusion strategy combining Transformer and CNN modules to integrate multiscale features. The global branch incorporates Swin Transformer blocks and residual connections to model high-level contextual dependencies, while the local branch utilizes densely

connected convolutional layers to preserve fine-grained spatial details. FATFusion [56] employs a Transformer-based guidance module to promote effective feature transmission and aggregation, thereby improving the integration of global contextual information.

Transformers across multiple stages: Transformers have also been applied to multiple stages of MMIF methods [20,22,58]. For instance, EMMA [22] employs four stacked Transformer–CNN blocks across both the feature extraction and image reconstruction stages. SwinFusion [20] is a representative full-stage fusion model built upon the Swin Transformer architecture. It integrates self-attention and cross-attention to capture intra-modal and inter-modal dependencies. It represents one of the earliest studies to explicitly visualize global information flow in image fusion.

While most transformer-based MMIF methods adopt hybrid CNN–Transformer architectures, a few recent studies have begun exploring the integration of transformers with MLPs [59] or diffusion models [60]. These efforts suggest a growing interest in developing alternative architectures beyond CNNs. Despite their advantages in capturing long-range dependencies and extracting global contextual information, Transformer-based methods usually introduce high computational complexity and memory consumption due to the attention mechanism, which presents challenges for deployment in environments with limited resources.

4. Emerging Deep Learning-Based MMIF Methods

4.1. Diffusion Model-Based MMIF Methods

The diffusion model, first proposed in 2014 as a family of generative models, has recently surpassed GAN-based approaches in tasks such as image generation [73] and image restoration [74], thereby challenging the long-standing dominance of GANs in this field. DDFM [25] was the first method to introduce diffusion models into MMIF in 2023. Since then, an increasing number of diffusion model-based MMIF methods have been proposed, showing promising performance.

To illustrate the progress in this domain, we briefly review three notable diffusion model-based MMIF frameworks. UUD-Fusion [62] introduces a two-stage general image fusion method based on diffusion model. In the first stage, the CNN generates a rough fusion result. In the second stage, two diffusion-based sampling algorithms are employed to explore the solution space and refine the fusion quality. Diff-IF [26] leverages the distribution information provided by the fusion knowledge prior and targeted search to construct the forward diffusion process, while the reverse diffusion process generates fused images that conform to the fusion knowledge prior distribution. FusionDiff [61] addresses the data dependency issue in existing unified image fusion approaches by pre-fusing source images through a diffusion process. It introduces a noise prediction network, a Spatially-Adaptive Constraint layer, and a skip-sampling mechanism to improve image quality while reducing the computational cost of traditional diffusion models.

Overall, diffusion models offer notable advantages, including high-quality generation, stable training, and a mathematically interpretable modeling process. However, both training and inference usually require substantial computational resources and time.

4.2. Mamba-Based MMIF Methods

Mamba [75], a recently introduced selective state space model (SSM), has received increasing attention for its ability to capture long-range dependencies with linear computational complexity.

Given these advantages, many Mamba-based MMIF methods have been proposed [27,63–65]. For example, LPM-Net [27] employs three processing branches, each integrating a dense CNN with Mamba to enhance feature extraction. Zhang et al. [64]

propose a VSSM-based cross-domain image fusion framework with a dual-branch extractor and a VMamba-based decoder for image fusion and generation. DWMFusion [65] is designed as a lightweight MMIF framework that integrates discrete wavelet transform and convolution for feature extraction, and adopts a Mamba-based multilevel module to efficiently fuse multiscale complementary features.

5. Datasets and Evaluation Metrics

5.1. Datasets for MMIF

This section reviews the most commonly used public medical datasets, as shown in Figure 3d. These datasets encompass various imaging modalities, such as CT, MRI, and PET. They provide high-quality samples that facilitate the development of image fusion algorithms. The datasets are summarized in Table 2.

The Whole Brain Atlas (AANLIB), developed by Harvard Medical School, includes approximately 13,000 brain images from both healthy subjects and patients with neurological disorders such as cerebrovascular disease, brain tumors, and Alzheimer's disease. It includes four imaging modalities: MRI, CT, PET, and SPECT, with all images offered in GIF format.

The Brain Tumor Segmentation (BRATS) dataset, part of the MICCAI BRATS Challenge series, consists of multimodal MRI scans from glioma patients, including four clinically relevant sequences: T1-weighted, T1 with contrast enhancement (T1Gd), T2-weighted, and FLAIR. The data are provided in the standardized NIFTI format. As of BRATS 2021, the dataset includes over 1250 cases, covering both low-grade and high-grade gliomas.

Table 2. MMIF dataset.

Dataset	Year	Organ	Modality	Disease	Download Address	Format
AANLIB	1995	Brain	CT, MRI, PET, SPECT	Normal brain, Brain tumor, Alzheimer's disease	http://www.med.harvard.edu/AANLIB/ (accessed on 3 July 2026)	GIF
BRATS	2012	Brain	MRI	Glioma	https://www.med.upenn.edu/cbica/brats2021/ (accessed on 3 July 2026)	NIFTI
ADNI	2003	Brain	MRI, PET	Alzheimer's disease	https://adni.loni.usc.edu/data-samples/adni-data (accessed on 3 July 2026)	-
OASIS	2010	Brain	MRI, PET	Alzheimer's disease	https://www.oasis-brains.org/ (accessed on 3 July 2026)	NIFTI
TCIA	2014	Brain, Lungs	CT, MRI, PET, Ultrasound	Gliomas, Lung cancer	https://www.cancerimagingarchive.net/ (accessed on 3 July 2026)	DICOM

The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset was developed to support research into the early diagnosis and progression of Alzheimer's disease. It comprises multimodal imaging data from patients with Alzheimer's disease, individuals with mild cognitive impairment, and cognitively normal controls. The dataset includes three imaging modalities: MRI, PET, and CT, which are used to capture various aspects of brain structure and function.

The Open Access Series of Imaging Studies (OASIS) is a publicly available brain imaging dataset developed by Washington University, focusing on Alzheimer's disease and its prodromal stage, mild cognitive impairment. The dataset consists of four versions: OASIS-1, OASIS-2, OASIS-3, and OASIS-4. Among these, OASIS-3, which was released in 2019, represents the most comprehensive version and includes longitudinal structural MRI and PET scans acquired from more than 1000 participants.

The Cancer Imaging Archive (TCIA) is a publicly accessible medical imaging repository maintained by the U.S. National Cancer Institute (NCI). TCIA focuses on the collection and dissemination of multimodal cancer-related imaging data. TCIA offers datasets cov-

ering various cancer types from different anatomical sites, such as gliomas, lung cancer, breast cancer, liver cancer, and prostate cancer. The imaging modalities available include CT, MRI, PET, and ultrasound. All data are stored in the standardized DICOM format for consistency and interoperability.

5.2. Subjective Evaluation Criteria

Subjective qualitative evaluation assesses the quality of fused images through human visual perception, considering factors such as salient features, sharpness, and spatial detail. It captures the perceived visual quality and provides intuitive and reliable assessments. Consequently, it has been widely used in the evaluation of MMIF methods. However, subjective evaluation relies on individual judgment, and differences in evaluation criteria among observers may compromise the consistency [76] and reproducibility of the results [77]. To address these challenges, a variety of objective metrics have been proposed by researchers. These metrics enable quantitative and automated assessments of fusion quality, improving both the efficiency and reliability of the evaluation process.

5.3. Quantitative Evaluation Metrics

Objective quantitative evaluation does not rely on subjective human perception but assesses image quality through mathematical computation. It is efficient, consistent, and repeatable, and is less affected by visual bias. This section introduces several representative evaluation metrics commonly used in the surveyed papers and their corresponding categories. These metrics are listed in Table 3. We use A and B to denote the two source images and F to denote the fused image. X represents either source image A or B.

Table 3. Objective quantitative evaluation metrics sorted by frequency of use in the surveyed papers.

Metric	Full Name	Description	Formula	Category
VIF [78]	Visual information fidelity	It measures the preservation of visual information in the fused image.	$VIF(I_V, I_R, I_F) = \sum_k p_k \frac{\sum_b FVID_{s,b}(I_V, I_R, I_F)}{\sum_b FVIND_{s,b}(I_V, I_R, I_F)}$	Human visual quality metrics
MI [79]	Mutual information	It measures the degree of information shared between the fused image and its source images.	$MI = MI_{A,F} + MI_{B,F}$ $MI_{X,F} = \sum_{x,f} p_{X,F}(x,f) \log \frac{p_{X,F}(x,f)}{p_X(x)p_F(f)}$	Information-based metrics
SSIM [80]	Structural similarity index measure	It measures the similarity between the source and fused images by comparing their structural information.	$SSIM = SSIM_{A,F} + SSIM_{B,F}$ $SSIM_{X,F} = \frac{2\mu_x\mu_f + C_1}{\mu_x^2 + \mu_f^2 + C_1} \cdot \frac{2\sigma_x\sigma_f + C_2}{\sigma_x^2 + \sigma_f^2 + C_2} \cdot \frac{\sigma_{xf} + C_3}{\sigma_x\sigma_f + C_3}$	Structural preservation metrics
EN [81]	Entropy	It measures the information content of the fused image, reflecting the complexity and diversity of its pixel distribution.	$EN = -\sum_{l=0}^{L-1} p_l \log_2 p_l$	Information-based metrics
$Q^{AB/F}$ [82]	Gradient-based fusion performance	It measures how much edge information from the source images is retained in the fused image.	$Q^{AB/F} = \frac{\sum_{i=1}^N \sum_{j=1}^M (Q^{AF}(i,j)w^A(i,j) + Q^{BF}(i,j)w^B(i,j))}{\sum_{i=1}^N \sum_{j=1}^M (w^A(i,j) + w^B(i,j))}$	Information-based metrics
SCD	Sum of the correlations of differences	It evaluates the quality of a fused image by measuring the correlation between each source image and its corresponding difference image with the fused result.	$SCD = r(I_V, D_{V,F}) + r(I_R, D_{R,F})$	Correlation-based metrics
SD [83]	Standard deviation	The standard deviation metric measures the distribution and contrast of the fused image.	$SD = \sqrt{\sum_{i=1}^M \sum_{j=1}^N (F(i,j) - \mu)^2}$	Image Feature-based metrics

Table 3. Cont.

Metric	Full Name	Description	Formula	Category
SF [84]	Spatial frequency	It is an index that indicates the variation in grayscale values within an image. It reflects the level of detail and texture present.	$SF = \sqrt{H^2 + V^2}$ $H = \sqrt{\frac{1}{MN} \sum_{i=1}^M \sum_{j=2}^N I_F(i, j) - I_F(i, j-1) ^2}$ $V = \sqrt{\frac{1}{MN} \sum_{i=2}^M \sum_{j=1}^N I_F(i, j) - I_F(i-1, j) ^2}$	Image feature-based metrics
PSNR [85]	Peak signal-to-noise ratio	It reflects the level of distortion in the fused image by calculating the ratio between peak signal power and noise power.	$PSNR = 10 \log_{10} \frac{r^2}{MSE}$	Information-based metrics
CC [86]	Correlation coefficient	It evaluates the strength of the linear relationship between the fused result and the original inputs.	$CC = \frac{r_{AF} + r_{BF}}{2}$	Correlation-based metrics

5.4. Experiments

To quantitatively compare representative deep learning-based MMIF methods, this study selects nine methods from the 49 reviewed methods, covering five major architectural categories: CNN-based methods [4,15], AE-based methods [23,24], a GAN-based method [18], Transformer-based methods [20,22], and diffusion model-based methods [25,26]. Experiments are conducted on 40 randomly selected PET–MRI image pairs from the public Harvard Whole Brain Atlas dataset. The 10 fusion metrics introduced in Section 5.3 are used to evaluate fusion performance. For all selected deep learning-based methods, the pretrained models released by the original authors are used, and no additional retraining is performed in this study.

Table 4 presents the quantitative results of the nine selected deep learning-based MMIF methods. The results show that different methods exhibit advantages under different metrics. Diff-IF [26] achieves the best VIF and SSIM, indicating better visual information preservation and structural consistency. Diff-IF [26] and DDFM [25] obtain the highest MI, suggesting effective preservation of complementary source information. DDFM [25] achieves the best PSNR and CC, reflecting lower reconstruction distortion and stronger source-image correlation. CDDFuse [23] obtains the best SD, SCD, and SF, indicating advantages in contrast representation, complementary information preservation, and spatial detail representation. EMMA [22] achieves the highest EN, while SwinFusion [20] obtains the best $Q^{AB/F}$, reflecting their respective strengths in information richness and edge-related feature preservation. Overall, no single method consistently outperforms the others across all metrics, indicating the necessity of multi-metric evaluation combined with qualitative analysis.

Table 4. Quantitative comparison results of different MMIF methods.

Method	VIF	MI	SSIM	SD	SCD	EN	$Q^{AB/F}$	SF	PSNR	CC
DDCGAN	0.461	1.788	0.120	48.230	0.492	5.454	0.275	18.110	12.230	0.786
U2Fusion	0.468	1.796	0.542	59.466	0.531	5.303	0.402	20.961	14.993	0.860
EMFusion	0.655	2.113	1.270	80.298	0.893	5.271	0.714	33.365	13.602	0.798
C2RF	0.730	2.307	1.266	90.569	1.409	5.298	0.729	36.714	12.576	0.798
CDDFuse	0.715	2.337	1.287	93.250	1.591	5.287	0.724	37.953	12.415	0.806
EMMA	0.516	1.776	0.683	90.375	1.502	5.979	0.571	28.261	12.957	0.833
SwinFusion	0.747	2.234	0.657	88.406	1.384	5.782	0.753	37.135	12.212	0.806
Diff-IF	0.772	2.611	1.297	92.001	1.438	5.055	0.716	37.356	12.528	0.806
DDFM	0.643	2.611	1.128	75.346	1.234	5.043	0.509	22.492	15.267	0.869

5.5. Unified Evaluation Metrics

Numerous evaluation metrics have been proposed for image fusion. Based on [87,88], these metrics can be broadly classified into five categories: human visual quality metrics, information-based metrics, structural preservation metrics, image feature-based metrics, and correlation-based metrics. Despite the large number of proposed metrics, no single metric can comprehensively characterize all aspects of fusion performance. As shown in Table 1, existing studies typically employ only a subset of evaluation metrics, which may limit objective comparisons of performance.

To address these issues, this study recommends a unified set of evaluation metrics. The selection is guided by metric classification, the use of evaluation metrics in existing studies summarized in Table 1, and the frequency analysis of commonly used metrics shown in Figure 3b. This design ensures coverage of complementary metric categories while maintaining a balance between practical feasibility and evaluation comprehensiveness. The selected metrics and their corresponding categories are VIF (human visual quality), MI (information-based), SSIM (structural preservation), SCD (correlation-based), and SD (image feature-based). These unified metrics facilitate systematic performance comparisons across different fusion methods. Moreover, these unified metrics may provide a quantitative basis for future studies linking objective fusion quality with downstream clinical task performance, thereby supporting more clinically meaningful assessment.

6. Future Research Directions

6.1. Better Evaluation Metrics

Due to the lack of ground truth in MMIF, current evaluation methods primarily rely on subjective visual assessments and objective metrics for quantitative analysis. However, as illustrated in Table 1, different studies use inconsistent evaluation criteria, making it difficult to conduct fair and reproducible comparisons across fusion methods. This inconsistency further impedes the establishment of authoritative benchmarking systems, thereby slowing down systematic progress and methodological evolution in the field. Moreover, some metrics, such as EN, SD, and CC, typically capture only limited low-level aspects of fused images, such as contrast or statistical distribution, and often fail to reflect their overall perceptual quality [89,90].

To address these limitations, it is essential to establish unified and standardized evaluation metrics to ensure methodological fairness in comparison. Furthermore, selecting a diverse and appropriate set of metrics that are aligned with human perception and capable of capturing multiple dimensions of fusion quality is critical for the comprehensive assessment of MMIF methods.

6.2. Registration-Aware Robust Medical Image Fusion

Robustness to spatial misalignment is essential for the reliable clinical application of medical image fusion models. In real-world MMIF scenarios, images from different modalities are often difficult to align precisely due to differences in imaging mechanisms, acquisition protocols, and patient motion. However, most existing methods still assume well-registered inputs [17,91], while the literature on the fusion of misaligned medical images remains limited [24,92]. Future research should develop registration-aware fusion networks to improve the robustness of MMIF models to spatially inconsistent inputs.

6.3. Mamba-Based Architectures for Future MMIF Research

Although research on applying Mamba to MMIF is still in its early stages, the model's ability to capture long-range dependencies with linear computational complexity positions it as a highly promising foundation for future developments. As interest in efficient

sequence modeling grows, it is expected that an increasing number of Mamba-based MMIF frameworks will emerge. Notably, the design of fully Mamba-based MMIF architectures remains largely unexplored and deserves further investigation.

6.4. Benchmark Expansion in MMIF

The quantity and diversity of available datasets are important factors affecting MMIF model performance and generalizability in clinical applications. At present, publicly available MMIF datasets remain limited and are primarily concentrated on brain imaging and common modality combinations, such as MRI-CT, MRI-PET, and MRI-SPECT. AANLIB is the most frequently used benchmark; however, its predominant focus on brain imaging limits its applicability to broader clinical scenarios. Meanwhile, datasets covering other anatomical regions and disease types remain scarce, which constrains the development of more generalizable MMIF models.

Existing datasets also exhibit variations in image format, spatial resolution, and registration accuracy, while standardized dataset construction and evaluation protocols remain insufficiently established. Therefore, developing large-scale, standardized, and clinically diverse MMIF datasets covering multiple anatomical regions, disease types, and imaging modalities is essential for further advancing this field.

6.5. Application-Driven MMIF Methods

One important application of MMIF is to support downstream applications such as medical image segmentation or classification. Recently, some researchers [1] have explored deep learning-based MMIF methods designed for specific tasks. However, this direction remains relatively underexplored. We expect that task-driven MMIF methods will become a key research direction in the future.

6.6. Generalizable MMIF Models

Developing generalizable MMIF models remains an important future direction for adapting fusion methods to diverse clinical scenarios and fusion tasks. Early unified frameworks such as U2Fusion [15] used continual learning to handle multiple fusion tasks and alleviate subtask conflicts, but they did not promote task interaction during training. Later studies have explored different architectures to improve fusion performance and generalizability, including CNN-based methods [16,39], AE-based methods [24], GAN-based methods [18], Transformer-based methods [20,22], diffusion model-based methods [93], and Mamba-based methods [27].

However, these methods still depend on large-scale data from multiple fusion tasks and usually require task-specific retraining or multiple dedicated models, which limits their ability to achieve genuine cross-task generalization. GIFNet [57] addresses this limitation through a cross-fusion gating mechanism that introduces pixel-level supervision from digital photography fusion, promoting task-shared feature learning and enabling generalizable multi-task fusion within a single model. Future research should therefore move toward genuinely generalizable MMIF models that reduce dependence on task-specific retraining and dedicated models while enabling multiple fusion tasks within a single unified framework.

6.7. Clinical Applicability and Translation of MMIF

Although deep learning-based MMIF methods have shown promising performance, their clinical translation remains limited by several practical challenges. For PET/SPECT fusion, preserving quantitative information is essential because these modalities provide functional information related to radiotracer distribution, metabolic activity, and tissue uptake. Therefore, fusion methods should enhance visual quality without distorting

clinically meaningful quantitative features. In addition, external validation across different institutions, scanners, imaging protocols, and patient populations remains insufficient. Practical deployment also requires compatibility with clinical infrastructure and workflows, including PACS/RIS systems, DICOM standards, radiological reading, and reporting processes. Future studies should emphasize quantitative consistency, external validation, workflow compatibility, and clinical usability to support reliable clinical translation.

7. Conclusions

In this review, a comprehensive overview of deep learning-based methods for MMIF is presented. It categorizes existing methods into mainstream methods and emerging methods, highlights representative models in each category, and provides a statistical analysis of loss functions, evaluation metrics, modality combinations, and dataset usage. It also reviews commonly used MMIF datasets and evaluation metrics. A significant limitation in current MMIF research lies in the inconsistency of evaluation metrics, which hampers direct and fair comparisons across methods. To address this issue, this review recommends a unified metric set to support more standardized and comprehensive assessment of MMIF methods. Moreover, this review also outlines future trends and key challenges in the field. Promising research directions include registration-aware robust medical image fusion, Mamba-based architectures, benchmark expansion in MMIF, and application-driven MMIF methods. This review serves as a useful reference for researchers in medical image fusion, enabling efficient access to prior work and guiding the design of improved MMIF methodologies.

Author Contributions: Conceptualization, Y.S. and M.R.; methodology, Y.S. and M.R.; investigation, Y.S.; data curation, Y.S.; formal analysis, Y.S.; visualization, Y.S.; writing—original draft preparation, Y.S.; writing—review and editing, M.R.; validation, Y.S. and M.R.; supervision, M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The publication fees of this manuscript have been financed by the Research Council of the University of Patras.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new original data were generated.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MMIF	Multimodal medical image fusion
MRI	Magnetic resonance imaging
CT	Computed tomography
PET	Positron emission tomography
SPECT	Single-photon emission computed tomography
DL	Deep learning
CNN	Convolutional neural network
AE	Autoencoder
GAN	Generative adversarial network
INN	Invertible neural network
SSM	Structured state space model
DICOM	Digital Imaging and Communications in Medicine
PACS	Picture archiving and communication system

RIS	Radiology information system
VIF	Visual information fidelity
MI	Mutual information
SSIM	Structural similarity index measure
SCD	Sum of correlations of differences
SD	Standard deviation
EN	Entropy
SF	Spatial frequency
AG	Average gradient
CC	Correlation coefficient
PSNR	Peak signal-to-noise ratio
MSE	Mean squared error

References

1. Liu, Y.; Shi, Y.; Mu, F.; Cheng, J.; Chen, X. Glioma segmentation-oriented multi-modal MR image fusion with adversarial learning. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 1528–1531. [\[CrossRef\]](#)
2. Zhang, Y.; Liu, Y.; Sun, P.; Yan, H.; Zhao, X.; Zhang, L. IFCNN: A general image fusion framework based on convolutional neural network. *Inf. Fusion* **2020**, *54*, 99–118. [\[CrossRef\]](#)
3. Yin, M.; Liu, X.; Liu, Y.; Chen, X. Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampling shearlet transform domain. *IEEE Trans. Instrum. Meas.* **2018**, *68*, 49–64. [\[CrossRef\]](#)
4. Xu, H.; Ma, J. EMFusion: An unsupervised enhanced medical image fusion network. *Inf. Fusion* **2021**, *76*, 177–186. [\[CrossRef\]](#)
5. Du, J.; Li, W.; Xiao, B. Anatomical-functional image fusion by information of interest in local Laplacian filtering domain. *IEEE Trans. Image Process.* **2017**, *26*, 5855–5866. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Vishwakarma, A.; Bhuyan, M.K. Image fusion using adjustable non-subsampling shearlet transform. *IEEE Trans. Instrum. Meas.* **2018**, *68*, 3367–3378.
7. Hill, P.; Al-Mualla, M.E.; Bull, D. Perceptual image fusion using wavelets. *IEEE Trans. Image Process.* **2016**, *26*, 1076–1088. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Li, X.; Guo, X.; Han, P.; Wang, X.; Li, H.; Luo, T. Laplacian redecomposition for multimodal medical image fusion. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 6880–6890. [\[CrossRef\]](#)
9. Yang, B.; Li, S. Pixel-level image fusion with simultaneous orthogonal matching pursuit. *Inf. Fusion* **2012**, *13*, 10–19. [\[CrossRef\]](#)
10. Jiang, Y.; Wang, M. Image fusion with morphological component analysis. *Inf. Fusion* **2014**, *18*, 107–118. [\[CrossRef\]](#)
11. Liu, Y.; Chen, X.; Ward, R.K.; Wang, Z.J. Medical image fusion via convolutional sparsity based morphological component analysis. *IEEE Signal Process. Lett.* **2019**, *26*, 485–489. [\[CrossRef\]](#)
12. Liu, Y.; Liu, S.; Wang, Z. A general framework for image fusion based on multi-scale transform and sparse representation. *Inf. Fusion* **2015**, *24*, 147–164. [\[CrossRef\]](#)
13. Singh, S.; Anand, R.S. Multimodal medical image sensor fusion model using sparse K-SVD dictionary learning in nonsubsampling shearlet domain. *IEEE Trans. Instrum. Meas.* **2019**, *69*, 593–607. [\[CrossRef\]](#)
14. Liu, Y.; Chen, X.; Cheng, J.; Peng, H. A medical image fusion method based on convolutional neural networks. In *Proceedings of the 2017 20th International Conference on Information Fusion (Fusion)*; IEEE: New York, NY, USA, 2017; pp. 1–7.
15. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 502–518. [\[CrossRef\]](#)
16. Zhang, H.; Ma, J. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. *Int. J. Comput. Vis.* **2021**, *129*, 2761–2785. [\[CrossRef\]](#)
17. Liu, J.; Lin, R.; Wu, G.; Liu, R.; Luo, Z.; Fan, X. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *Int. J. Comput. Vis.* **2024**, *132*, 1748–1775.
18. 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 Trans. Image Process.* **2020**, *29*, 4980–4995. [\[CrossRef\]](#)
19. Zhou, T.; Fu, H.; Chen, G.; Shen, J.; Shao, L. Hi-net: Hybrid-fusion network for multi-modal MR image synthesis. *IEEE Trans. Med. Imaging* **2020**, *39*, 2772–2781. [\[CrossRef\]](#) [\[PubMed\]](#)
20. 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 J. Autom. Sin.* **2022**, *9*, 1200–1217. [\[CrossRef\]](#)
21. Tang, W.; He, F.; Liu, Y.; Duan, Y. MATR: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Trans. Image Process.* **2022**, *31*, 5134–5149. [\[CrossRef\]](#) [\[PubMed\]](#)

22. Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Zhang, K.; Xu, S.; Chen, D.; Timofte, R.; Van Gool, L. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 25912–25921.
23. 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*; IEEE: New York, NY, USA, 2023; pp. 5906–5916.
24. Tang, L.; Yan, Q.; Xiang, X.; Fang, L.; Ma, J. C2RF: Bridging multi-modal image registration and fusion via commonality mining and contrastive learning. *Int. J. Comput. Vis.* **2025**, *133*, 5262–5280. [[CrossRef](#)]
25. 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*; IEEE: New York, NY, USA, 2023; pp. 8082–8093.
26. Yi, X.; Tang, L.; Zhang, H.; Xu, H.; Ma, J. Diff-IF: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Inf. Fusion* **2024**, *110*, 102450. [[CrossRef](#)]
27. Wen, M.; Zhang, X. LPM-Net: Lightweight pixel-level modeling network based on CNN and Mamba for 3D medical image fusion. *Inf. Fusion* **2025**, *123*, 103306. [[CrossRef](#)]
28. Hermessi, H.; Mourali, O.; Zagrouba, E. Multimodal medical image fusion review: Theoretical background and recent advances. *Signal Process.* **2021**, *183*, 108036. [[CrossRef](#)]
29. Azam, M.A.; Khan, K.B.; Salahuddin, S.; Rehman, E.; Khan, S.A.; Khan, M.A.; Kadry, S.; Gandomi, A.H. A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Comput. Biol. Med.* **2022**, *144*, 105253. [[CrossRef](#)] [[PubMed](#)]
30. Khan, S.U.; Khan, M.A.; Azhar, M.; Khan, F.; Lee, Y.; Javed, M. Multimodal medical image fusion towards future research: A review. *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 101733. [[CrossRef](#)]
31. Zhou, T.; Cheng, Q.; Lu, H.; Li, Q.; Zhang, X.; Qiu, S. Deep learning methods for medical image fusion: A review. *Comput. Biol. Med.* **2023**, *160*, 106959. [[CrossRef](#)] [[PubMed](#)]
32. Liu, Y.; Yu, C.; Cheng, J.; Wang, Z.J.; Chen, X. MM-Net: A mixformer-based multi-scale network for anatomical and functional image fusion. *IEEE Trans. Image Process.* **2024**, *33*, 2197–2212. [[CrossRef](#)] [[PubMed](#)]
33. Xie, X.; Zhang, X.; Ye, S.; Xiong, D.; Ouyang, L.; Yang, B.; Zhou, H.; Wan, Y. Mrscfusion: Joint residual swin transformer and multiscale CNN for unsupervised multimodal medical image fusion. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–17. [[CrossRef](#)]
34. Guo, W.; Wang, L.; Zeng, J.; Han, Q.; Jin, K.; Wang, X. SMAFusion: Multimodal medical image fusion based on spatial registration and local-global multi-scale feature adaptive fusion. *Neurocomputing* **2025**, *652*, 131039. [[CrossRef](#)]
35. Zhang, G.; Nie, R.; Cao, J.; Chen, L.; Zhu, Y. FDGNet: A pair feature difference guided network for multimodal medical image fusion. *Biomed. Signal Process. Control* **2023**, *81*, 104545. [[CrossRef](#)]
36. Fu, J.; Li, W.; Du, J.; Huang, Y. A multiscale residual pyramid attention network for medical image fusion. *Biomed. Signal Process. Control* **2021**, *66*, 102488. [[CrossRef](#)]
37. Wen, J.; Khan, A.; Chen, A.; Peng, W.; Fang, M.; Chen, C.L.P.; Li, P. High-quality fusion and visualization for MR-PET brain tumor images via multi-dimensional features. *IEEE Trans. Image Process.* **2024**, *33*, 3550–3563. [[CrossRef](#)] [[PubMed](#)]
38. Wen, J.; Qin, F.; Du, J.; Fang, M.; Wei, X.; Chen, C.L.P.; Li, P. MsgFusion: Medical semantic guided two-branch network for multimodal brain image fusion. *IEEE Trans. Multimed.* **2023**, *26*, 944–957.
39. Cheng, C.; Xu, T.; Wu, X.-J. MUFusion: A general unsupervised image fusion network based on memory unit. *Inf. Fusion* **2023**, *92*, 80–92. [[CrossRef](#)]
40. Liu, R.; Liu, J.; Jiang, Z.; Fan, X.; Luo, Z. A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion. *IEEE Trans. Image Process.* **2020**, *30*, 1261–1274. [[PubMed](#)]
41. Liu, Y.; Shi, Y.; Mu, F.; Cheng, J.; Li, C.; Chen, X. Multimodal MRI volumetric data fusion with convolutional neural networks. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–15. [[CrossRef](#)]
42. Zhang, X.; Liu, A.; Jiang, P.; Qian, R.; Wei, W.; Chen, X. MSAIF-net: A multistage spatial attention-based invertible fusion network for MR images. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–13. [[CrossRef](#)]
43. Li, J.; Liu, J.; Zhou, S.; Zhang, Q.; Kasabov, N.K. Gesenet: A general semantic-guided network with couple mask ensemble for medical image fusion. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 16248–16261.
44. Wei, X.; Qiu, Y.; Xu, X.; Xu, J.; Mei, J.; Zhang, J. ECINFusion: A novel explicit channel-wise interaction network for unified multi-modal medical image fusion. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *35*, 4011–4025. [[CrossRef](#)]
45. Di, J.; Guo, W.; Liu, J.; Ren, L.; Lian, J. AMMNet: A multimodal medical image fusion method based on an attention mechanism and MobileNetV3. *Biomed. Signal Process. Control* **2024**, *96*, 106561. [[CrossRef](#)]
46. Luo, X.; Gao, Y.; Wang, A.; Zhang, Z.; Wu, X.-J. IFSepR: A general framework for image fusion based on separate representation learning. *IEEE Trans. Multimed.* **2021**, *25*, 608–623. [[CrossRef](#)]

47. Cheng, C.; Wu, X.-J.; Xu, T.; Chen, G. Unifusion: A lightweight unified image fusion network. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–14. [[CrossRef](#)]
48. Liu, J.; Dian, R.; Li, S.; Liu, H. SGFusion: A saliency guided deep-learning framework for pixel-level image fusion. *Inf. Fusion* **2023**, *91*, 205–214. [[CrossRef](#)]
49. Hu, S.; Lei, B.; Wang, S.; Wang, Y.; Feng, Z.; Shen, Y. Bidirectional mapping generative adversarial networks for brain MR to PET synthesis. *IEEE Trans. Med Imaging* **2021**, *41*, 145–157. [[CrossRef](#)] [[PubMed](#)]
50. Zhan, B.; Li, D.; Wu, X.; Zhou, J.; Wang, Y. Multi-modal MRI image synthesis via GAN with multi-scale gate mergence. *IEEE J. Biomed. Health Inform.* **2021**, *26*, 17–26.
51. Huang, J.; Li, X.; Tan, H.; Cheng, X. Generative adversarial network for trimodal medical image fusion using primitive relationship reasoning. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 5729–5741. [[CrossRef](#)]
52. Zhou, H.; Hou, J.; Zhang, Y.; Ma, J.; Ling, H. Unified gradient-and intensity-discriminator generative adversarial network for image fusion. *Inf. Fusion* **2022**, *88*, 184–201. [[CrossRef](#)]
53. Zhang, J.; Jiao, L.; Ma, W.; Liu, F.; Liu, X.; Li, L.; Chen, P.; Yang, S. Transformer based conditional GAN for multimodal image fusion. *IEEE Trans. Multimed.* **2023**, *25*, 8988–9001. [[CrossRef](#)]
54. Zhao, C.; Yang, P.; Zhou, F.; Yue, G.; Wang, S.; Wu, H.; Chen, G.; Wang, T.; Lei, B. MHW-GAN: Multidiscriminator hierarchical wavelet generative adversarial network for multimodal image fusion. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 13713–13727. [[CrossRef](#)]
55. Xie, X.; Zhang, X.; Tang, X.; Zhao, J.; Xiong, D.; Ouyang, L.; Yang, B.; Zhou, H.; Ling, B.W.-K.; Teo, K.L. MACTFusion: Lightweight cross transformer for adaptive multimodal medical image fusion. *IEEE J. Biomed. Health Inform.* **2024**, *29*, 3317–3328. [[CrossRef](#)]
56. Tang, W.; He, F. FATFusion: A functional-anatomical transformer for medical image fusion. *Inf. Process. Manag.* **2024**, *61*, 103687. [[CrossRef](#)]
57. Cheng, C.; Xu, T.; Feng, Z.; Wu, X.; Tang, Z.; Li, H.; Zhang, Z.; Atito, S.; Awais, M.; Kittler, J. One model for ALL: Low-level task interaction is a key to task-agnostic image fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*; IEEE: New York, NY, USA, 2025; pp. 28102–28112.
58. Wang, W.; Deng, L.-J.; Vivone, G. A general image fusion framework using multi-task semi-supervised learning. *Inf. Fusion* **2024**, *108*, 102414. [[CrossRef](#)]
59. Zhang, Z.; Li, H.; Xu, T.; Wu, X.-J.; Kittler, J. DDBFusion: An unified image decomposition and fusion framework based on dual decomposition and Bézier curves. *Inf. Fusion* **2025**, *114*, 102655. [[CrossRef](#)]
60. Yang, B.; Jiang, Z.; Pan, D.; Yu, H.; Gui, G.; Gui, W. LFDT-Fusion: A latent feature-guided diffusion transformer model for general image fusion. *Inf. Fusion* **2025**, *113*, 102639. [[CrossRef](#)]
61. Huang, Z.; Yang, S.; Wu, J.; Zhu, L.; Liu, J. FusionDiff: A unified image fusion network based on diffusion probabilistic models. *Comput. Vis. Image Underst.* **2024**, *244*, 104011. [[CrossRef](#)]
62. Wang, X.; Fang, L.; Zhao, J.; Pan, Z.; Li, H.; Li, Y. UUD-Fusion: An unsupervised universal image fusion approach via generative diffusion model. *Comput. Vis. Image Underst.* **2024**, *249*, 104218. [[CrossRef](#)]
63. Wu, P.; Tang, J. Mamba-based hybrid network for multi-modality feature fusion. *Eng. Appl. Artif. Intell.* **2025**, *156*, 111064. [[CrossRef](#)]
64. Zhang, L.; Wang, L.; Wu, Y.; Chen, M.; Zheng, D.; Cai, Y. ISCDFuse: Interval sampling correlation driven visual state space models for multimodal image fusion. *Neurocomputing* **2025**, *640*, 130329. [[CrossRef](#)]
65. Fan, C.; Yun, M.; Chen, L.; Zhao, H.; Peng, B. DWMFusion: A lightweight medical image fusion network based on discrete wavelet convolution and selective state space models. *Expert Syst. Appl.* **2025**, *286*, 128040. [[CrossRef](#)]
66. Li, H.; Wu, X.-J.; Kittler, J. RFN-Nest: An end-to-end residual fusion network for infrared and visible images. *Inf. Fusion* **2021**, *73*, 72–86. [[CrossRef](#)]
67. Allapakam, V.; Karuna, Y. An ensemble deep learning model for medical image fusion with Siamese neural networks and VGG-19. *PLoS ONE* **2024**, *19*, e0309651. [[CrossRef](#)] [[PubMed](#)]
68. Chen, Q.; Wu, Q.; Wang, J.; Hu, Q.; Hu, T.; Ding, E.; Cheng, J.; Wang, J. Mixformer: Mixing features across windows and dimensions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 5249–5259.
69. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 4700–4708.
70. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
71. 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*; IEEE: New York, NY, USA, 2022; pp. 5728–5739.
72. Wu, Z.; Liu, Z.; Lin, J.; Lin, Y.; Han, S. Lite transformer with long-short range attention. *arXiv* **2020**, arXiv:2004.11886.

73. Graikos, A.; Yellapragada, S.; Le, M.-Q.; Kapse, S.; Prasanna, P.; Saltz, J.; Samaras, D. Learned representation-guided diffusion models for large-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 8532–8542.
74. Zhu, Y.; Zhang, K.; Liang, J.; Cao, J.; Wen, B.; Timofte, R.; Van Gool, L. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2023; pp. 1219–1229.
75. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the First Conference on Language Modeling*, Philadelphia, PA, USA, 7–9 October 2024.
76. Cao, P.; Wang, Z.; Ma, K. Debiased subjective assessment of real-world image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2021; pp. 711–721.
77. Chen, Y.; Blum, R.S. A new automated quality assessment algorithm for image fusion. *Image Vis. Comput.* **2009**, *27*, 1421–1432. [[CrossRef](#)]
78. Han, Y.; Cai, Y.; Cao, Y.; Xu, X. A new image fusion performance metric based on visual information fidelity. *Inf. Fusion* **2013**, *14*, 127–135. [[CrossRef](#)]
79. Qu, G.; Zhang, D.; Yan, P. Information measure for performance of image fusion. *Electron. Lett.* **2002**, *38*, 313–315. [[CrossRef](#)]
80. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)] [[PubMed](#)]
81. Roberts, J.W.; Van Aardt, J.A.; Ahmed, F.B. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *J. Appl. Remote Sens.* **2008**, *2*, 023522. [[CrossRef](#)]
82. Petrović, V.; Xydeas, C. Objective evaluation of signal-level image fusion performance. *Opt. Eng.* **2005**, *44*, 087003. [[CrossRef](#)]
83. Aslantas, V.; Bendes, E. A new image quality metric for image fusion: The sum of the correlations of differences. *AEU-Int. J. Electron. Commun.* **2015**, *69*, 1890–1896. [[CrossRef](#)]
84. Eskicioglu, A.M.; Fisher, P.S. Image quality measures and their performance. *IEEE Trans. Commun.* **2002**, *43*, 2959–2965.
85. Raju, P.V.R.; Rani, D.S.; Challaram, G. Comparison of medical image fusion methods using image quality metrics. In *Proceedings of the 2018 International Conference on Communication, Computing and Internet of Things (IC3IoT)*; IEEE: New York, NY, USA, 2018; pp. 449–454.
86. Han, Z.; Tang, X.; Gao, X.; Hu, F. Image fusion and image quality assessment of fused images. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2013**, *40*, 33–36. [[CrossRef](#)]
87. Liu, Z.; Blasch, E.; Xue, Z.; Zhao, J.; Laganieri, R.; Wu, W. Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: A comparative study. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 94–109. [[CrossRef](#)] [[PubMed](#)]
88. Huang, Y.; Li, W. A review of medical image fusion methods. *J. Image Graph.* **2023**, *28*, 118–143. [[CrossRef](#)]
89. Zhang, X.; Demiris, Y. Visible and infrared image fusion using deep learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10535–10554. [[CrossRef](#)] [[PubMed](#)]
90. Liu, J.; Wu, G.; Liu, Z.; Wang, D.; Jiang, Z.; Ma, L.; Zhong, W.; Fan, X. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 2349–2369. [[CrossRef](#)]
91. Li, J.; Han, D.; Wang, X.; Yi, P.; Yan, L.; Li, X. Multi-sensor medical-image fusion technique based on embedding bilateral filter in least squares and salient detection. *Sensors* **2023**, *23*, 3490. [[CrossRef](#)] [[PubMed](#)]
92. Xu, H.; Yuan, J.; Ma, J. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12148–12166. [[CrossRef](#)] [[PubMed](#)]
93. Shi, Y.; Liu, Y.; Cheng, J.; Wang, Z.J.; Chen, X. VDMUFusion: A versatile diffusion model-based unsupervised framework for image fusion. *IEEE Trans. Image Process.* **2024**, *34*, 441–454. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.