Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 8, No. 3, July 2026, pp: 948-965 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS
(Correction)

Comparative Analysis of Attention Mechanisms in Pix2Pix for Multimodal MRI Fusion

Ali-Abdelatif Betouil[1], Abdelmadjid Benmachiche[1], Khadija Rais[2], Amel Sahki[1], and Imene Soualmia[1]

[1] Laboratory of Computer Science and Applied Mathematics, Dept. of Computer Science, Faculty of Science and Technology, Chadli Bendjedid, University, El-Tarf, Algeria.
[2] Laboratory of mathematics, informatics and systems (LAMIS), Echahid Cheikh Larbi Tebessi University, Tebessa, Algeria.

Corresponding author: Khadija Rais (e-mail: khadija.rais@univ-tebessa.dz), **Author(s) Email**: Ali-Abdelatif Betouil (e-mail: a.betouil@univ-eltarf.dz), Abdelmadjid Benmachiche (e-mail: benmachiche-abdelmadjid@univ-eltarf.dz), Amel Sahki (e-mail: a.sahki@univ-eltarf.dz), Imene Soualmia (e-mail: i.soualmia@univ-eltarf.dz)

Abstract **Medical image fusion (MIF) is a key technique in medical imaging, which combines complementary information from different imaging modalities, thereby improving the accuracy of diagnosis, particularly for lesion detection and treatment planning. Deep learning has significantly advanced this area, with the development of generative models and transformers leading to improvements in fidelity and accuracy, although the study of the influence of attention mechanisms on these models remains limited to a single type or a single architectural placement. This paper offers an analytical examination of the architectures of Pix2Pix with three attention mechanisms (spatial attention, channel attention (Squeeze-and-Excitation), and self-attention), where they are tested in three different placement strategies (encoder-only, decoder-only, and encoder-decoder), using the BraTS2020 dataset, with training supervised by a pseudo-ground-truth derived from arithmetic averaging. We fused six MRI modality pairs (FLAIR-T1, FLAIR-T1ce, FLAIR-T2, T1-T1ce, T1-T2, T1ce-T2), evaluating them using different metrics, including SSIM, PSNR, NMI, Entropy, and $Q^{AB/F}$. Results show that, in all cases, attention integration can significantly improve the quality of fusion over baseline methods, including cGAN and standard Pix2Pix. Spatial attention with encoder-decoder placement shows the best results, with SSIM values up to 0.91 and PSNR superior to 25 dB for the heterogeneous modality pair FLAIR-T1. Similarly, channel and self-attention demonstrate their effectiveness, especially with encoder-decoder placements. Based on these findings, attention-based fusion systems can be practically designed in a way that enhances MMIF, and the importance of designing attention in accordance with the nature of the modality is emphasized for optimal fusion performance. Our study demonstrates its effectiveness and may serve as a foundation for future research.**

Keywords **Multimodal medical image fusion, GANs, Pix2Pix, Attention Mechanisms, Spatial Attention, Squeeze-and-Excitation, Self-Attention, BraTS20, MRI**

I. Introduction

Multimodal medical image fusion (MMIF) has emerged as a critical technique in modern diagnostic imaging, enabling clinicians to synthesize complementary information from distinct imaging modalities into a single, more informative representation [1]. By integrating structural details from modalities such as MRI and CT with functional insights from PET or SPECT, fusion methods aim to enhance lesion delineation, tissue characterization, and overall diagnostic confidence. These integrated representations support precision medicine by enabling earlier detection of anomalies, improving diagnostic interpretation, and facilitating more accurate treatment monitoring [2].

Recent advances have shifted the paradigm from traditional signal-processing-based approaches, such as wavelet or contourlet transforms, to deep learning frameworks, particularly Generative Adversarial Networks (GANs). These models learn end-to-end mappings between input modalities and a fused output, often yielding improved structural consistency and perceptual quality. This transition aligns with broader trends in medical AI, where studies have documented the evolution from classical methods to architectures leveraging convolutional neural networks, GANs, and

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.

emerging diffusion models for multimodal synthesis [3]. Such frameworks not only improve fusion fidelity but also support downstream tasks like segmentation and classification by generating synthetic data that augment limited or imbalanced medical datasets [4]. However, their performance remains highly dependent on the availability of large, paired, and high-quality datasets, which is a significant challenge in medical imaging due to privacy constraints, annotation costs, and the rarity of certain pathologies [5].

To address data scarcity, recent studies have explored data augmentation strategies [6], including generative techniques like Deep Convolutional GANs (DCGANs). Furthermore, publicly available multimodal medical image databases such as TCIA, OASIS, ADNI, MIDAS, AANLIB, and DDSM provide diverse, curated datasets spanning multiple modalities and pathological conditions, offering researchers reproducible benchmarks for developing and evaluating fusion algorithms [7]. Despite these resources, standard fusion architectures may still fail to fully exploit the rich, heterogeneous information present across modalities. This has led to the integration of attention mechanisms, spatial attention, channel attention (e.g., Squeeze-and-Excitation), and self-attention, into generator networks to dynamically recalibrate feature representations and prioritize diagnostically relevant regions or channels [8].

Recent studies further highlight the growing importance of advanced multimodal fusion strategies in medical diagnosis. For example, prompt-level contrastive learning frameworks such as PCL-MFP demonstrate how context-aware multimodal prompts can significantly improve representation learning by integrating images, textual descriptions, and probabilistic diagnostic priors, offering a more clinically aligned fusion paradigm [9]. Similarly, attention-adaptive fusion networks such as SAFusion introduce scenario-aware architectures that dynamically adjust fusion strategies across different clinical conditions using mixture-of-experts and deformable convolutional mechanisms, improving robustness and adaptability in diverse medical imaging settings [10]. In addition, efficient image–tabular fusion frameworks such as AMF-MedIT emphasize modality confidence alignment and modulation-based fusion strategies, addressing key challenges such as feature dimension imbalance and noisy tabular clinical data through lightweight and data-efficient designs [11].

Despite these innovations, a systematic evaluation of how different attention mechanisms and their placement within Pix2Pix architectures affect multimodal MRI fusion performance remains limited. Existing studies often focus on a single attention strategy or specific modality pair, making it difficult to assess the relative effectiveness of spatial, channel, and self-attention mechanisms across diverse fusion tasks. In addition, many approaches rely on heuristic pseudo-targets due to the absence of true ground-truth fused images.

In this work, we present a comprehensive evaluation of attention-augmented Pix2Pix architectures for multimodal brain MRI fusion using the BraTS20 dataset. We investigate three attention mechanisms, spatial, channel, and self-attention, under three placement strategies: encoder-only, decoder-only, and encoder-decoder configurations, across six MRI modality pairs. Unlike previous studies that primarily introduce new architectures or evaluate a single attention design, this work provides a controlled comparative analysis within a unified Pix2Pix framework. The main contribution of this study is the systematic assessment of how attention type and placement influence fusion quality, structural preservation, and anatomical fidelity in multimodal MRI synthesis. Additionally, the findings provide practical insights for designing robust attention-enhanced GAN-based medical image fusion systems.

II. Related works

Multimodal medical image fusion has advanced by moving from handcrafted signal-processing techniques to deep learning approaches that combine CT, MRI, PET, and SPECT data for more accurate and reliable clinical diagnosis. Early methods relied on handcrafted transforms and fusion rules, such as NSCT with PCNN [12], but lacked adaptability. Later, transform-domain methods incorporated attention mechanisms, e.g., CGCEA-SCGS [13], enhancing low- and high-frequency preservation. To reduce dependence on labeled data, zero-shot and lightweight models emerged, including Yang et al. [14] and nd MMIF-VAEFusion [15], improving efficiency but with limited generalization.

With deep learning, MMIF shifted to data-driven frameworks. CNN-based models like MedFusionGAN [16] enable unsupervised fusion but suffer from instability, while contrastive approaches such as MFDCE-Fuse [17] improve feature disentanglement. Attention-enhanced CNNs (e.g., MACAN [18]) and Transformer-based models like MATR [19] and FMTFuse [20] further enhance global context and detail preservation. Hybrid architectures, including MDC-RHT [21], ECFusion [22], MSAFusion [23], and MARFusion [24], combine convolutional and attention mechanisms to improve performance. More recent methods explore diffusion models (DDFM [25]), structured GANs [26], and state-space models like MAPD-Mamba [27] for efficient long-range modeling. Beyond fusion, MMIF supports downstream tasks such as segmentation and classification. Methods like DLF [28] and Dempster-Shafer-based fusion [29] improve

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.

segmentation robustness, while quality assessment frameworks such as QANet+CASNet [30] address evaluation challenges. In classification, multimodal fusion enhances tasks like lung nodule analysis [31], with architectures like HiFuse [32] capturing both local and global features. Additionally, task-oriented synthesis models such as MIF-GAN [33] enable applications like contrast-enhanced CT generation while preserving diagnostic fidelity.

III. Methodology

A. Pix2Pix Framework

Our approach is based on a conditional generative adversarial network (cGAN) following the Pix2Pix paradigm, adapted for paired multi-modal brain MRI fusion. The system consists of an input construction stage, a U-Net generator, and a patch-based discriminator. The generator is trained using a composite loss combining adversarial and reconstruction terms (Fig. 1).

1. Input Pair Construction

For each training sample, two co-registered MRI modalities are concatenated along the channel dimension to form a multi-channel input tensor, as presented in (1) [34]:

$$x = [m_1, m_2] \quad (1)$$

Here, x denotes the concatenated multimodal input tensor, while m_1 and m_2 represent the paired co-registered MRI modalities. Since a true ground-truth fused image is not available for supervised learning, a surrogate fusion target is defined using a simple arithmetic averaging operator. Aktar et al. explicitly define the average fusion rule. Adopting this approach, we define our surrogate training target as: (2) [34]:

$$y = F(m_1, m_2) = \frac{m_1 + m_2}{2} \quad (2)$$

Here, $F(\cdot)$ denotes the surrogate fusion function, and y represents the pseudo-ground-truth fused image obtained using arithmetic averaging of the input modalities.

2. Generator Architecture:

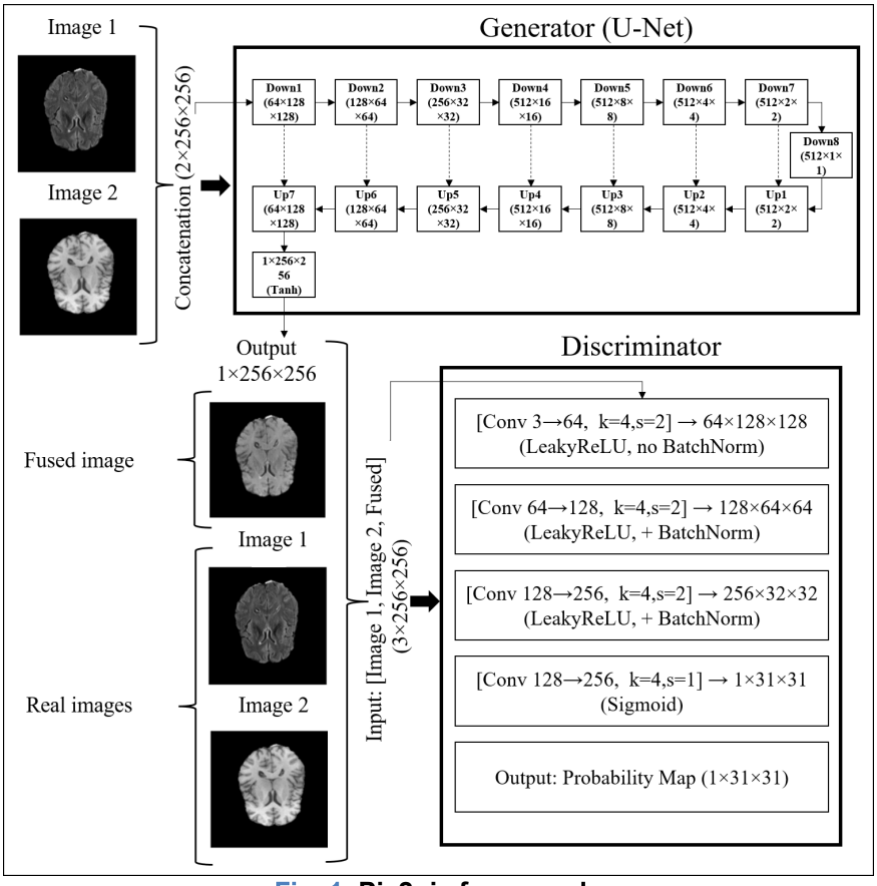

Fig. 1. Pix2pix framework.

The generator follows a U-Net encoder-decoder structure with skip connections to preserve spatial detail while enabling high-level semantic abstraction. It maps a multi-channel input $x = [m_1, m_2]$ to a single-channel fused image.

a. Encoder (Downsampling Path): The encoder contains a sequence of downsampling blocks. Each block applies a convolution operation with a

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.

stride greater than one, followed by a nonlinear activation function. Normalization is applied in all blocks except the first to preserve input dynamics. The number of feature channels increases progressively to form a multi-scale representation.

b. Bottleneck: The bottleneck layer captures high-level semantic features and does not include normalization, consistent with standard generative adversarial architectures.

c. Decoder (Upsampling Path): The decoder consists of upsampling blocks, each applying transposed convolution (or equivalent), normalization, and nonlinear activation. Dropout may be applied in early decoding stages for regularization.
Skip connections concatenate encoder and decoder features at corresponding scales, combining low-level spatial detail with high-level semantic information.

d. Output Layer: The final layer restores the original spatial resolution and applies an activation function to map outputs to a normalized intensity range.

3. Discriminator Architecture: A patch-based discriminator (PatchGAN) is used to classify local image patches as real or fake, focusing on high-frequency texture.

a. Input: The discriminator receives a conditional pair $D(x, y)$, where y corresponds either to the generated output $G(x)$ (fake) or the surrogate fused reference $\frac{m_1+m_2}{2}$ (real).

b. Convolutional Backbone: A sequence of convolutional layers progressively reduces spatial resolution while increasing feature depth. Nonlinear activations and normalization are used to improve training stability.

c. Output Layer: A sigmoid activation produces a spatial probability map, where each value represents the realism of a local patch.

4. Loss Functions

The discriminator loss is defined as binary cross-entropy (BCE) between predicted labels and ground-truth labels (real or fake), as defined in (3) [35]:

$$L_D = -\mathbb{E}_{x,y}[\log D(x,y)] - \mathbb{E}_x[\log(1 - D(x, G(x)))] \quad (3)$$

Here, L_D denotes the discriminator loss, $D(x,y)$ represents the discriminator output for real pairs, and $D(x, G(x))$ denotes the discriminator output for generated pairs. $\mathbb{E}_{(x,y)}[\cdot]$ is the expectation over real input-target pairs, while $\mathbb{E}_x[\cdot]$ is the expectation over input samples only. $G(x)$ denotes the generator output conditioned on the input x, and $\log(\cdot)$ is the logarithmic function used in the binary cross-entropy formulation. The generator loss combines an adversarial term and a reconstruction term, as presented in (4) [35]:

$$\mathcal{L}_G = \mathcal{L}_{\text{adv}} + \lambda \cdot \mathcal{L}_{\text{rec}} \quad (4)$$

Here, L_G denotes the total generator loss, L_{adv} represents the adversarial loss, L_{rec} denotes the reconstruction loss, and λ is a weighting coefficient controlling the trade-off between adversarial realism and reconstruction fidelity. The adversarial loss is defined as (5) [35]:

$$L_{adv} = -\mathbb{E}_x[\log D(x, G(x))] \quad (5)$$

Here, L_{adv} denotes the adversarial loss, $D(x, G(x))$ represents the discriminator probability that the generated image is real, and $\mathbb{E}_x[\cdot]$ denotes expectation over input samples. The reconstruction loss is a pixel-wise difference between the generated output and the surrogate fused target. In this work, we use the L1 loss, as mentioned in (6) [35]:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{x,y}[\| y - G(x) \|_1] \quad (6)$$

Here, L_{rec} denotes the reconstruction loss, $y$is the surrogate fused target image, $G(x)$ is the generator output, and $\|\cdot\|_1$ denotes the L1 norm measuring pixel-wise absolute differences.

B. Attention Mechanisms

Three attention mechanisms are integrated into the generator, as shown in Fig. 2: spatial attention, channel attention (Squeeze-and-Excitation), and self-attention. Each mechanism is evaluated under three placement strategies: encoder-only, decoder-only, and encoder-decoder.

1. Spatial Attention

Given a feature map $X \in \mathbb{R}^{B\times C\times H\times W}$, spatial attention is computed by aggregating information along the channel dimension using average ((7)(7)) and max pooling ((8)) [36]:

$$F_{\text{avg}} = \frac{1}{C}\sum_{c=1}^{C} X_c \quad (7)$$

Here, F_{avg} denotes the channel-wise average-pooled spatial descriptor, C is the number of channels, and X_c represents the feature map of the c-th channel.

$$F_{\text{max}} = \max_{c} X_c \quad (8)$$

Here, F_{max} denotes the channel-wise max-pooled spatial descriptor, and X_c represents the c-th channel feature map. These two descriptors ($F_{avg}, F_{max} \in \mathbb{R}^{B\times 1\times H\times W}$) are concatenated and passed through a convolution layer to generate a spatial attention map $M_s \in \mathbb{R}^{B\times 1\times H\times W}$.

The final refined feature map X' is obtained by element-wise multiplication of the input X and the sigmoid-normalized attention map, as defined in (9) [36]:

$$X' = X \odot \sigma(M_s) \quad (9)$$

Here, X' denotes the refined feature map, M_s represents the spatial attention map, $\sigma(\cdot)$ is the sigmoid activation function, and $\odot$ denotes element-wise multiplication.

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.

This process allows the network to focus on informative spatial regions while suppressing irrelevant background noise.

a. Encoder-only: Applied in encoder blocks after convolution to focus feature extraction on salient regions.
b. Decoder-only: Applied in decoder blocks after upsampling and before skip connections to enhance reconstructed details.
c. Encoder-decoder: Applied throughout all blocks to enforce consistent spatial attention.

2. Channel Attention (Squeeze-and-Excitation)

Given an input feature map $X \in \mathbb{R}^{B\times C\times H\times W}$, the SE block recalibrates channel-wise feature responses as follows:

Squeeze (Global Information Embedding): Compress the spatial dimensions using Global Average Pooling to generate a channel descriptor $z \in \mathbb{R}^{B\times C}$, as shown in (10) [37]:

$$z_{b,c} = \frac{1}{H \cdot W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{b,c}(i,j) \quad (10)$$

Here, $z_{(b,c)}$ denotes the squeezed channel descriptor for the sample b and channel c, $X_{(b,c)}(i,j)$ represents the feature value at spatial location (i,j), and H, W denote spatial dimensions of the feature map. Excitation (Adaptive Recalibration): Learn channel-wise dependencies using a gating mechanism with a bottleneck structure. Let r be the reduction ratio. The channel attention weights $s \in \mathbb{R}^{B\times C}$ are computed as (11) [37]:

$$s = \sigma\,(W_2 \cdot \delta(W_1 \cdot z)) \quad (11)$$

Here, s denotes the channel attention weights, W_1 and W_2 are learnable projection matrices, $\delta(\cdot)$ is the ReLU activation function, $\sigma(\cdot)$ is the sigmoid function, z is the channel descriptor, and r is the reduction ratio. Scale (Feature Reweighting): Rescale the original feature map X by the attention weights s via channel-wise multiplication (broadcasting s over spatial dimensions), as defined in (12) [37]:

$$X'_{b,c}(i,j) = s_{b,c} \cdot X_{b,c}(i,j) \quad (12)$$

Here, $s_{(b,c)}$ denotes the channel attention weight for the channel $c$and sample b, and $X_{(b,c)}(i,j)$ is the original feature response at the spatial location (i,j). Or in tensor notation, ((13)) [37]:

$$X' = X \odot s \quad (13)$$

Here, X' denotes the recalibrated feature map, X is the input feature map, and $s$represents channel-wise attention weights applied via broadcasting. This process allows the network to adaptively emphasize informative channels while suppressing less useful ones, thereby enhancing feature discriminability.

a. Encoder-only: Applied in encoder blocks to improve feature selection.
b. Decoder-only: Applied in decoder blocks to refine semantic reconstruction.
c. Encoder-decoder: Applied throughout the network for global channel optimization.

3. Self-Attention

Given an input feature map $X \in \mathbb{R}^{B\times C\times H\times W}$, we first

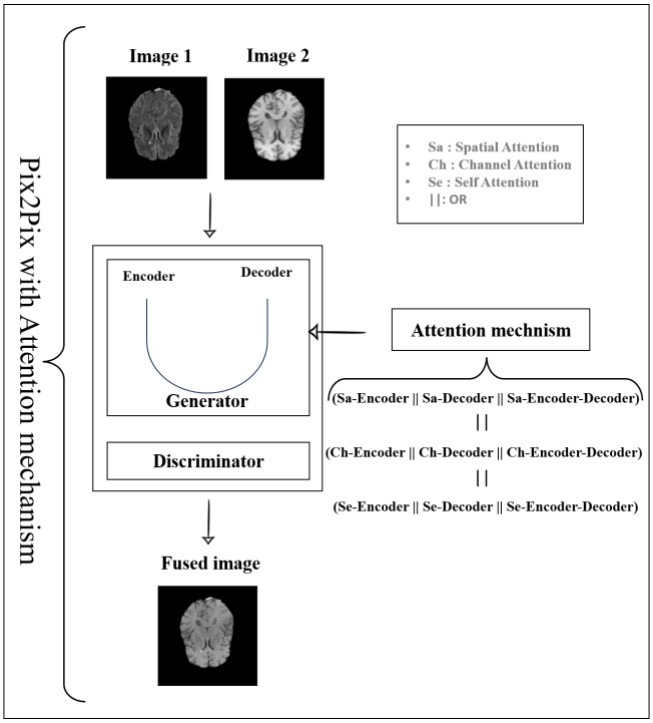

Fig. 2. Our Pix2Pix-Attention architecture.

project the features into query, key, and value representations using 1×1 convolutions (or linear projections after reshaping), as defined in (14) [38]:

$$Q = W_q X, \quad K = W_k X, \quad V = W_v X \quad (14)$$

Here, Q, K, and Vdenote the query, key, and value feature representations, while W_q, W_k, and W_v are learnable projection matrices used to compute them from input features.

Let $N = H \times W$. The spatial attention map $A \in \mathbb{R}^{B\times N\times N}$ is computed by measuring the affinity between all pairs of positions, ((15)) [38]:

$$A = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d}}\right) \quad (15)$$

Here, A denotes the self-attention map, $QK^{\top}$ represents the similarity between query and key features, d is the feature dimension used for scaling, and $\text{softmax}(\cdot)$ normalizes attention scores into probabilities. The refined feature map is obtained by aggregating the values weighted by the attention map, scaled by a learnable parameter γ, and adding the residual connection, as shown in (16) [38]:

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.

$$X' = \gamma \cdot \text{Reshape}(AV) + X \quad (16)$$

Here, X' denotes the output feature map, γ is a learnable scaling parameter initialized to zero, A is the attention matrix, V is the value feature representation, and $\text{Reshape}(\cdot)$ restores spatial dimensions before adding the residual connection with the input X. This process allows the network to capture long-range dependencies without incurring the prohibitive memory costs associated with high-resolution attention maps.

a. Encoder-only: Applied in encoder blocks (excluding the first) to capture global structural dependencies.
b. Decoder-only: Applied in early decoder blocks to enforce structural consistency.
c. Encoder-decoder: Applied in both the encoder and decoder to model long-range dependencies across all scales.

IV. Implementation

(7Table 1 summarizes the architectural components of the proposed fusion framework, including the generator, discriminator, and integrated attention mechanisms. The generator follows a U-Net encoder-decoder design with skip connections, while the discriminator adopts a PatchGAN structure for patch-level realism assessment. Three attention modules, spatial, Squeeze-and-Excitation, and self-attention, are incorporated to enhance feature representation. Their placement across the encoder (down) and decoder (up) stages is detailed to highlight their role in capturing both local and global dependencies during fusion.

A. Dataset and preprocessing

We use the BraTS 2020 dataset [39], [40], [41], which contains 369 multimodal 3D MRI volumes (T1, T1ce, T2, and FLAIR) stored in NIfTI format (.nii.gz). Each 3D volume is preprocessed by extracting axial slices, which are then converted into 2D images of size 256 × 256 in JPEG format. In total, 1107 axial slices are extracted from the dataset, corresponding to 277 subjects without visible tumors ("NoTumor" subset) and 830 slices with tumor regions. For this study, only the NoTumor subset is used to ensure anatomical consistency and reduce pathological variability during fusion evaluation.

To construct paired inputs, slices are matched across modalities using consistent slice indices after registration, and six modality pairs are generated: (FLAIR-T1, FLAIR-T1ce, FLAIR-T2, T1-T1ce, T1-T2, and T1ce-T2). Each input sample consists of channel-wise concatenated paired slices. To avoid data leakage, the dataset split is performed at the subject level, ensuring that slices from the same subject do not appear in different subsets. The dataset is divided into training, validation, and test sets using a ratio of 70% / 15% / 15%, respectively. This ensures that all slices belonging to a given subject are assigned exclusively to one split.

All images are normalized to the range [-1, 1] using min-max normalization applied per image. Standard preprocessing, including resizing to 256×256 and normalization is applied. However, no random data augmentation techniques (such as random flipping, rotation, or cropping) are used to ensure a fair comparison between attention mechanisms.

B. Network architecture

The generator is a U-Net with 8 encoder and 7 decoder blocks. The input to the generator consists of two concatenated MRI modalities (256 × 256 × 2), and the output is a single fused grayscale image (256 × 256 × 1). Encoder blocks use 4 × 4 convolutions (stride 2, padding 1), LeakyReLU activation (α = 0.2), and batch normalization (except the first block). The encoder feature dimensions increase progressively from 64 to 128, 256, 512, and remain at 512 for the deeper layers. Decoder blocks use 4 × 4 transposed convolutions (stride 2, padding 1), ReLU activation, and batch normalization. Dropout (p = 0.5) is applied in the first three decoder blocks. The decoder feature dimensions decrease progressively from 512 to 256, 128, and 64.

Skip connections concatenate encoder features with upsampled decoder outputs. The final layer uses a Tanh activation function to produce the fused image. The discriminator is a 3-layer PatchGAN consisting of three 4 × 4 convolution layers with 64, 128, and 256 filters (stride 2, padding 1), followed by a final 4 × 4 convolution layer (stride 1) with sigmoid activation. Batch normalization is applied except in the first discriminator layer. It outputs a patch-level probability map indicating real or fake regions. Three attention mechanisms are integrated into the generator:

a. Spatial Attention: 7×7 convolution after channel-wise average and max pooling.
b. Squeeze-and-Excitation (SE): Global average pooling followed by two 1×1 convolutions (reduction ratio 16).
c. Self-Attention: Query, key, and value projections with reduced channel dimensionality, followed by softmax attention and a residual connection.

Each attention mechanism is evaluated in three placement strategies: encoder-only, decoder-only, and encoder-decoder. For encoder-only placement, attention modules are inserted after the convolution operation in encoder blocks. For decoder-only placement, attention is applied after upsampling and before skip concatenation in decoder blocks. In the encoder-decoder configuration, attention modules are integrated into both stages.

All experiments are trained using the Adam optimizer (β_1 = 0.5, β_2 = 0.999) with a learning rate of 2

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.

× 10^{-4}, batch size of 16, and 50 training epochs. For reproducibility, random seeds are fixed for Python, NumPy, and PyTorch (seed = 42). The implementation is conducted in PyTorch on Google Colab using an NVIDIA Tesla V100 GPU (16 GB memory) with CUDA acceleration.

C. Training Protocol

Models are trained independently for each of the six modality pairs. The Adam optimizer ($\beta_1 = 0.5, \beta_2 = 0.999$) is used with a learning rate of 2×10^{-4} and a batch size of 16. Training is performed for 50 epochs on Google Colab using an NVIDIA Tesla V100 GPU (16 GB GPU memory) with the PyTorch framework and CUDA acceleration. To ensure reproducibility, random seeds are fixed for Python, NumPy, and PyTorch operations. No learning rate scheduling or early stopping is applied. The generator loss in (17) follows the composite form introduced in (4) [35]:

$$L_G = -\mathbb{E}_x[\log D(x, G(x))] + \lambda \cdot \mathbb{E}_x\left[\| G(x) - \frac{m_1 + m_2}{2} \|_1\right] \quad (17)$$

Here, L_G denotes the total generator loss. The term $-\log D(x, G(x))$ represents the adversarial loss that encourages the generator to produce realistic fused images capable of fooling the discriminator. The term $\| G(x) - (m_1 + m_2)/2 \|_1$ denotes the reconstruction loss measuring pixel-wise differences between the generated output and the surrogate fused target. $\mathbb{E}_x[\cdot]$ represents the expectation over input samples, and $\lambda = 100$ is a weighting factor controlling the trade-off between adversarial and reconstruction terms. Since a true fused ground-truth image is not available, the term $(m_1 + m_2)/2$ is used as a pseudo-ground-truth surrogate target. This provides a simple and stable supervisory signal while preserving shared structural information from both modalities. The adversarial loss term $-\log D(x, G(x))$ encourages the generator to produce realistic fused images, while the L1 loss enforces pixel-wise consistency with the surrogate target. Binary cross-entropy is used for implementing the adversarial loss. The discriminator is trained to distinguish real images from generated ones, following the standard adversarial training procedure with alternating updates between the generator and the discriminator. Specifically, the discriminator loss is defined in ((18)) [35]:

$$\mathcal{L}_D = -\mathbb{E}_x[\log D\left(\frac{m_1 + m_2}{2}\right)] - \mathbb{E}_x[\log(1 - D(G(x)))] \quad (18)$$

Here, L_D denotes the discriminator loss. The term $D((m_1 + m_2)/2)$ represents the discriminator output for the surrogate real fused image, while $D(G(x))$ denotes

Table 1. Summary of the proposed Pix2Pix-based architecture.

Category	Component	Specification
Generator	Architecture	U-Net (Encoder-Decoder)
	Input size	(256 * 256 * 2)
	Output size	(256 * 256 * 1)
	Encoder (Down)	8 blocks: Conv ((4*4), stride 2, padding 1) + LeakyReLU ((\alpha = 0.2)) + BN
	Encoder channels	64, 128, 256, 512, 512, 512, 512, 512
	Decoder (Up)	7 blocks: Transposed Conv ((4*4), stride 2, padding 1) + ReLU + BN
	Decoder channels	512, 512, 512, 512, 256, 128, 64
	Dropout	(p = 0.5) (first 3 decoder blocks)
	Skip connections	Concatenation (encoder, decoder)
	Output	Tanh activation
Discriminator	Architecture	PatchGAN (3-layer)
	Layers	3 × Conv ((4+4), stride 2) + 1 × Conv ((4*4), stride 1)
	Filters	64, 128, 256, 1
	Activation	LeakyReLU ((\alpha = 0.2))
	Output	Patch-level probability map
Attention	Spatial Attention	Avg + Max pooling, (7*7) conv, spatial mask
	Placement (Down)	Encoder blocks (after convolution)
	Placement (Up)	Decoder blocks (after upsampling, before skip)
	SE Attention	Global avg pooling, 2 × (1*1) conv, channel scaling
	Placement (Down)	Encoder blocks
	Placement (Up)	Decoder blocks (before skip)
	Self-Attention	Q, K, V projections, softmax attention, residual
	Placement (Down)	Encoder blocks (except first)
	Placement (Up)	Early decoder blocks (coarse scales)

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.

the discriminator output for the generated fused image. $\mathbb{E}_x[\cdot]$ represents the expectation over input samples, and $\log(\cdot)$ is the logarithmic function used in the binary cross-entropy formulation.

D. Pseudocode

The Algorithm 1 defines a Pix2Pix-based generator with integrated attention mechanisms. It encodes multimodal inputs, applies optional spatial, channel (SE), or self-attention at different feature levels, and reconstructs a fused image via a U-Net decoder. Attention can be selectively applied in the encoder, decoder, or both stages, enabling flexible control over feature refinement and a more adaptive fusion process.

Algorithm 1. Attention-Integrated Pix2Pix.

```
H ← X; Skips ← []
// ——— ENCODER ———
For i = 1 to 8:
    H ← Conv → Norm(i>1) → LeakyReLU
    If Placement in {Encoder, Both} AND i > 1:
        H ← Attention(H, Attention_Type)
    Skips.append(H)
    // Downsampling via stride-based convolution
    If i < 8:
        H ← Conv(stride=2)
// ——— DECODER ———
For i = 1 to 7:
    H ← ConvTranspose → Norm → ReLU
    If Placement in {Decoder, Both}:
        If NOT (Attention_Type == Self-Attention AND i > 4):
            H ← Attention(H, Attention_Type)
    H ← Concat(H, Skips.pop())
    If i ≤ 3:
        H ← Dropout(H)
// ——— OUTPUT ———
Y ← ConvTranspose(H) → Tanh
Return Y
// ——— ATTENTION SUBROUTINE ———
Function Attention(F, Type):
    If Type == Spatial:
        mask ← Sigmoid(Conv7x7(Concat(AvgPool(F), MaxPool(F))))
        Return F ⊙ mask
    If Type == SE:
        w ← Sigmoid(FC2(ReLU(FC1(GlobalAvgPool(F)))))
        Return F ⊙ Broadcast(w)
    If Type == Self-Attention:
        Q,K,V ← 1x1Conv(F) → Reshape
        A ← Softmax((Q·K^T)/√d)
        Out ← A · V
        Return γ · Reshape(Out) + F   // γ initialized to 0
```

V. Results

Visual and quantitative comparisons demonstrated notable performance variations among the evaluated fusion models. cGAN generated relatively blurry fused images with structural inconsistencies and visible artifacts, particularly around tissue boundaries and ventricular regions. VAE-MMIF produced smoother outputs but with reduced structural detail and weaker anatomical preservation. DDPM generated visually diverse images and achieved competitive PSNR values in some modality pairs; however, its lower SSIM and $Q^{AB/F}$ scores indicated weaker structural and edge preservation.

In contrast, Pix2Pix and MedFusionGAN produced sharper and more anatomically consistent fused images. As illustrated in Fig. 3, Pix2Pix preserved fine textures and edges more effectively than cGAN, achieving higher PSNR values, for example, in the FLAIR-T1 modality pair (23.57 vs. 21.67 dB), as well as higher SSIM values (0.89 vs. 0.78). Pix2Pix also achieved higher NMI, Entropy, MI, and $Q^{AB/F}$ values across all modality pairs, indicating improved

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.

Table 2. Performance Metrics for Fused MRI Images Compared with FLAIR and T1 Modalities.

	Model	PSNR (dB)	SSIM	NMI	Entropy	MI	$Q^{AB/F}$
	VAE	16.71 ± 2.03	0.65 ± 0.04	1.15 ± 0.03	4.10 ± 0.35	1.28 ± 0.07	0.48 ± 0.04
	MedFusionGAN	23.32 ± 3.34	0.88 ± 0.03	1.29 ± 0.03	5.18 ± 0.29	2.21 ± 0.05	0.76 ± 0.03
	DDPM	20.76 ± 2.29	0.44 ± 0.04	1.21 ± 0.02	4.51 ± 0.34	1.65 ± 0.06	0.52 ± 0.04
	cGAN	21.67 ± 4.07	0.78 ± 0.04	1.25 ± 0.02	4.82 ± 0.31	1.72 ± 0.06	0.61 ± 0.05
	Pix2Pix	23.57 ± 3.13	0.89 ± 0.03	1.29 ± 0.03	5.15 ± 0.28	2.18 ± 0.05	0.75 ± 0.03
	Attention-Based Pix2Pix						
SA	Encoder	23.50 ± 3.31	0.87 ± 0.04	1.29 ± 0.03	5.08 ± 0.29	2.12 ± 0.05	0.73 ± 0.03
SA	Decoder	23.74 ± 3.27	0.89 ± 0.03	1.29 ± 0.03	5.21 ± 0.28	2.25 ± 0.04	0.76 ± 0.03
SA	Encoder-Decoder	23.81 ± 3.32	0.89 ± 0.04	1.30 ± 0.03	5.28 ± 0.27	2.35 ± 0.04	0.78 ± 0.03
CA	Encoder	23.70 ± 3.30	0.89 ± 0.03	1.30 ± 0.03	5.25 ± 0.28	2.31 ± 0.04	0.79 ± 0.03
CA	Decoder	22.97 ± 3.02	0.81 ± 0.03	1.27 ± 0.02	4.92 ± 0.31	1.89 ± 0.05	0.67 ± 0.04
CA	Encoder-Decoder	23.85 ± 3.35	0.88 ± 0.03	1.30 ± 0.03	5.21 ± 0.28	2.28 ± 0.04	0.77 ± 0.03
SeA	Encoder	23.61 ± 3.28	0.89 ± 0.03	1.30 ± 0.03	5.19 ± 0.28	2.23 ± 0.05	0.76 ± 0.03
SeA	Decoder	23.76 ± 3.29	0.88 ± 0.04	1.29 ± 0.03	5.22 ± 0.29	2.26 ± 0.04	0.77 ± 0.03
SeA	Encoder-Decoder	23.66 ± 3.22	0.87 ± 0.03	1.29 ± 0.03	5.16 ± 0.29	2.21 ± 0.05	0.76 ± 0.03

information integration and edge preservation. These improvements may be attributed to the U-Net skip connections incorporated into the Pix2Pix architecture, which help preserve structural information during multimodal feature fusion. Paired t-tests confirmed that the observed improvements were statistically significant across all evaluated modality pairs ($p < 0.05$).

A. FLAIR and T1 Modalities

Table presents the quantitative evaluation of fusion models for FLAIR-T1 MRI. Among the baseline methods, VAE achieved the lowest performance with 16.71 ± 2.03 dB PSNR and 0.65 ± 0.04 SSIM, while howed moderate results at 20.76 ± 2.29 dB PSNR and 0.44 ± 0.04 SSIM. Both cGAN and Pix2Pix demonstrated stronger performance, with Pix2Pix reaching 23.57 ± 3.13 dB PSNR and 0.89 ± 0.03 SSIM, closely followed by MedFusionGAN at 23.32 ± 3.34 dB PSNR and 0.88 ± 0.03 SSIM.

Among the attention-enhanced Pix2Pix variants, all configurations yielded consistent improvements: the SA Encoder-Decoder achieved 23.81 ± 3.32 dB PSNR and 0.89 ± 0.04 SSIM, while the CA Encoder-Decoder attained the highest PSNR of 23.85 ± 3.35 dB with strong edge preservation ($Q^{AB/F}$: 0.77 ± 0.03). The SA and CA variants also maintained competitive entropy and mutual information scores, indicating robust fusion quality across complementary metrics.

B. FLAIR and T1ce Modalities

Table presents the quantitative evaluation of fusion models for FLAIR-T1ce MRI. Among the baseline methods, VAE achieved the lowest performance with 18.56 ± 2.32 dB PSNR and 0.67 ± 0.05 SSIM, while DDPM showed a high PSNR (24.31 ± 2.23 dB) but notably low structural fidelity (SSIM: 0.43 ± 0.03). Both

Table 4. Performance Metrics for Fused MRI Images Compared with FLAIR and T2 Modalities.

	Model	PSNR (dB)	SSIM	NMI	Entropy	MI	$Q^{AB/F}$
	VAE	17.57 ± 1.76	0.61 ± 0.05	1.15 ± 0.02	4.20 ± 0.34	1.31 ± 0.06	0.49 ± 0.04
	MedFusionGAN	21.93 ± 2.25	0.82 ± 0.05	1.26 ± 0.03	4.98 ± 0.29	2.12 ± 0.05	0.74 ± 0.03
	DDPM	23.56 ± 1.97	0.51 ± 0.04	1.21 ± 0.02	4.65 ± 0.32	1.70 ± 0.06	0.53 ± 0.04
	cGAN	20.60 ± 2.23	0.72 ± 0.04	1.23 ± 0.02	4.72 ± 0.33	1.61 ± 0.06	0.58 ± 0.04
	Pix2Pix	22.28 ± 2.31	0.84 ± 0.05	1.27 ± 0.04	5.08 ± 0.29	2.12 ± 0.05	0.72 ± 0.03
	Attention-Based Pix2Pix						
SA	Encoder	22.39 ± 2.34	0.84 ± 0.05	1.28 ± 0.04	5.11 ± 0.29	2.15 ± 0.05	0.73 ± 0.03
SA	Decoder	22.38 ± 2.33	0.84 ± 0.05	1.28 ± 0.04	5.13 ± 0.29	2.16 ± 0.05	0.74 ± 0.03
SA	Encoder-Decoder	23.92 ± 3.40	0.90 ± 0.04	1.31 ± 0.03	5.42 ± 0.26	2.45 ± 0.04	0.81 ± 0.03
CA	Encoder	22.29 ± 2.19	0.84 ± 0.05	1.28 ± 0.04	5.09 ± 0.29	2.13 ± 0.05	0.73 ± 0.03
CA	Decoder	22.22 ± 2.27	0.82 ± 0.05	1.26 ± 0.03	5.01 ± 0.29	2.08 ± 0.05	0.71 ± 0.04
CA	Encoder-Decoder	23.55 ± 3.18	0.86 ± 0.03	1.28 ± 0.03	5.29 ± 0.27	2.28 ± 0.04	0.77 ± 0.03
SeA	Encoder	22.35 ± 2.31	0.84 ± 0.05	1.28 ± 0.04	5.10 ± 0.29	2.14 ± 0.05	0.73 ± 0.03
SeA	Decoder	21.70 ± 2.41	0.82 ± 0.05	1.26 ± 0.04	4.98 ± 0.30	2.06 ± 0.05	0.70 ± 0.04
SeA	Encoder-Decoder	22.30 ± 2.28	0.83 ± 0.05	1.27 ± 0.04	5.06 ± 0.29	2.11 ± 0.05	0.72 ± 0.04

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.

cGAN and MedFusionGAN delivered moderate results, whereas Pix2Pix reached 24.65 ± 3.02 dB PSNR and 0.87 ± 0.03 SSIM.

Among the attention-enhanced Pix2Pix variants, all configurations yielded consistent improvements: the SA Encoder-Decoder achieved the highest PSNR of 25.13 ± 3.28 dB, SSIM of 0.89 ± 0.03, and mutual information of 2.42 ± 0.04, while the CA Encoder and SeA Encoder-Decoder also attained strong performance with PSNR values of 25.01 ± 3.23 dB and 25.02 ± 3.22 dB, respectively. The SA and SeA variants further maintained competitive entropy and edge preservation scores ($Q^{AB/F}$ up to 0.80), indicating robust fusion quality across complementary metrics.

C. FLAIR and T2 Modalities

Table presents the quantitative evaluation of fusion models for FLAIR-T2 MRI. Among the baseline methods, VAE achieved the lowest performance with 17.57 ± 1.76 dB PSNR and 0.61 ± 0.05 SSIM, while DDPM showed a relatively high PSNR (23.56 ± 1.97 dB) but notably low structural fidelity (SSIM: 0.51 ± 0.04). Both cGAN and MedFusionGAN delivered moderate results, whereas Pix2Pix reached 22.28 ± 2.31 dB PSNR and 0.84 ± 0.05 SSIM.

Among the attention-enhanced Pix2Pix variants, the SA Encoder-Decoder configuration achieved the strongest performance with 23.92 ± 3.40 dB PSNR, 0.90 ± 0.04 SSIM, mutual information of 2.45 ± 0.04, and edge preservation ($Q^{AB/F}$) of 0.81 ± 0.03. The CA Encoder-Decoder also attained competitive results with 23.55 ± 3.18 dB PSNR and 0.86 ± 0.03 SSIM, while other attention variants yielded modest improvements over the standard Pix2Pix baseline. These results indicate that spatial attention with encoder-decoder integration provides the most consistent gains for FLAIR-T2 fusion quality across complementary metrics.

D. T1 and T1ce Modalities

Table presents the quantitative evaluation of fusion models for T1-T1ce MRI. Among baseline methods, VAE achieved the lowest performance with 16.06 ± 2.63 dB PSNR and 0.64 ± 0.04 SSIM, while DDPM showed moderate PSNR (19.85 ± 1.13 dB) but low structural fidelity (SSIM: 0.46 ± 0.04). MedFusionGAN, cGAN, and Pix2Pix delivered comparable results, with Pix2Pix reaching 22.15 ± 3.66 dB PSNR and 0.88 ± 0.04 SSIM.

Among the attention-enhanced variants, the SA Encoder and Encoder-Decoder configurations achieved the strongest performance, with PSNR values of 22.42 ± 3.95 dB and 22.38 ± 3.90 dB, SSIM of 0.91 ± 0.04, and mutual information up to 2.43 ± 0.04, alongside improved edge preservation ($Q^{AB/F}$: 0.80-0.81). CA and SeA variants also yielded modest gains over the standard Pix2Pix baseline, confirming that spatial attention integration provides the most consistent improvements for T1-T1ce fusion quality.

E. T1 and T2 Modalities

Table presents the quantitative evaluation of fusion models for T1-T2 MRI. Among the baseline methods, VAE achieved the lowest performance with 16.79 ± 1.74 dB PSNR and 0.65 ± 0.04 SSIM, while DDPM showed a moderate PSNR (19.70 ± 1.07 dB) but low structural fidelity (SSIM: 0.44 ± 0.03). MedFusionGAN and Pix2Pix delivered the strongest baseline results, with MedFusionGAN reaching 20.25 ± 2.18 dB PSNR and 0.78 ± 0.05 SSIM.

Among the attention-enhanced Pix2Pix variants, improvements over the standard Pix2Pix baseline were modest but consistent: SA Encoder and CA Encoder achieved the highest PSNR values of 20.24 ± 2.14 dB and 20.18 ± 2.21 dB, respectively, with SSIM of 0.80 ± 0.05 and edge preservation ($Q^{AB/F}$) of 0.71 ± 0.04. All attention configurations yielded similar performance across metrics, indicating that while spatial and channel attention provide incremental benefits for T1-T2 fusion.

F. T1ce and T2 Modalities

Table presents the quantitative evaluation for the fusion of T1ce-T2 MRI modalities. Among baseline methods, VAE achieves the lowest performance with 13.45 ± 2.24 dB PSNR and 0.59 ± 0.05 SSIM. In contrast, cGAN exhibits a severe performance collapse, obtaining only 11.43 ± 1.27 dB PSNR and 0.08 ± 0.01 SSIM, indicating poor structural preservation and ineffective fusion in this modality pair. DDPM shows improved PSNR (22.39 ± 1.37 dB) but a relatively low SSIM (0.57 ± 0.05), reflecting limited structural consistency despite acceptable intensity reconstruction.

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.

Table 5. Performance Metrics for Fused MRI Images Compared with T1 and T1ce Modalities.

	Model	PSNR (dB)	SSIM	NMI	Entropy	MI	$Q^{AB/F}$
	VAE	16.06 ± 2.63	0.64 ± 0.04	1.14 ± 0.02	4.11 ± 0.36	1.30 ± 0.07	0.49 ± 0.04
	MedFusionGAN	22.10 ± 4.00	0.89 ± 0.04	1.33 ± 0.04	5.21 ± 0.28	2.22 ± 0.04	0.76 ± 0.03
	DDPM	19.85 ± 1.13	0.46 ± 0.04	1.25 ± 0.03	4.68 ± 0.32	1.72 ± 0.06	0.55 ± 0.04
	cGAN	22.26 ± 5.85	0.82 ± 0.06	1.29 ± 0.03	4.98 ± 0.32	1.85 ± 0.05	0.65 ± 0.04
	Pix2Pix	22.15 ± 3.66	0.88 ± 0.04	1.33 ± 0.04	5.16 ± 0.28	2.25 ± 0.04	0.75 ± 0.03
	Attention-Based Pix2Pix						
SA	Encoder	22.42 ± 3.95	0.91 ± 0.04	1.36 ± 0.05	5.34 ± 0.27	2.41 ± 0.04	0.80 ± 0.03
SA	Decoder	22.25 ± 3.72	0.90 ± 0.04	1.35 ± 0.04	5.29 ± 0.27	2.37 ± 0.04	0.79 ± 0.03
SA	Encoder-Decoder	22.38 ± 3.90	0.91 ± 0.04	1.36 ± 0.05	5.38 ± 0.27	2.43 ± 0.04	0.81 ± 0.03
CA	Encoder	21.33 ± 2.80	0.89 ± 0.04	1.35 ± 0.05	5.27 ± 0.27	2.35 ± 0.04	0.78 ± 0.03
CA	Decoder	22.10 ± 3.62	0.86 ± 0.04	1.32 ± 0.03	5.19 ± 0.28	2.27 ± 0.04	0.76 ± 0.03
CA	Encoder-Decoder	21.93 ± 3.30	0.89 ± 0.04	1.34 ± 0.04	5.31 ± 0.27	2.39 ± 0.04	0.80 ± 0.03
SeA	Encoder	22.32 ± 3.84	0.90 ± 0.04	1.34 ± 0.04	5.26 ± 0.28	2.33 ± 0.04	0.78 ± 0.03
SeA	Decoder	22.30 ± 3.82	0.90 ± 0.04	1.34 ± 0.04	5.29 ± 0.27	2.36 ± 0.04	0.79 ± 0.03
SeA	Encoder-Decoder	22.31 ± 3.87	0.89 ± 0.04	1.33 ± 0.04	5.24 ± 0.28	2.31 ± 0.04	0.77 ± 0.03

Table 6. Performance Metrics for Fused MRI Images Compared with T1 and T2 Modalities.

	Model	PSNR (dB)	SSIM	NMI	Entropy	MI	$Q^{AB/F}$
	VAE	16.79 ± 1.74	0.65 ± 0.04	1.17 ± 0.02	4.10 ± 0.35	1.31 ± 0.07	0.49 ± 0.04
	MedFusionGAN	20.25 ± 2.18	0.78 ± 0.05	1.25 ± 0.02	4.81 ± 0.31	2.04 ± 0.05	0.72 ± 0.04
	DDPM	19.70 ± 1.07	0.44 ± 0.03	1.19 ± 0.01	4.53 ± 0.33	1.70 ± 0.06	0.52 ± 0.04
	cGAN	18.78 ± 3.41	0.69 ± 0.05	1.23 ± 0.02	4.62 ± 0.35	1.52 ± 0.06	0.55 ± 0.05
	Pix2Pix	20.08 ± 2.01	0.79 ± 0.05	1.25 ± 0.03	5.08 ± 0.30	2.02 ± 0.05	0.69 ± 0.04
	Attention-Based Pix2Pix						
SA	Encoder	20.24 ± 2.14	0.80 ± 0.05	1.26 ± 0.03	5.14 ± 0.29	2.06 ± 0.05	0.71 ± 0.04
SA	Decoder	20.11 ± 2.11	0.77 ± 0.05	1.25 ± 0.03	5.06 ± 0.30	2.01 ± 0.05	0.69 ± 0.04
SA	Encoder-Decoder	20.12 ± 2.12	0.78 ± 0.05	1.25 ± 0.03	5.09 ± 0.30	2.03 ± 0.05	0.70 ± 0.04
CA	Encoder	20.18 ± 2.21	0.80 ± 0.05	1.26 ± 0.03	5.12 ± 0.29	2.05 ± 0.05	0.71 ± 0.04
CA	Decoder	20.15 ± 2.14	0.79 ± 0.05	1.25 ± 0.03	5.08 ± 0.30	2.02 ± 0.05	0.69 ± 0.04
CA	Encoder-Decoder	20.12 ± 2.15	0.79 ± 0.05	1.26 ± 0.03	5.10 ± 0.29	2.04 ± 0.05	0.70 ± 0.04
SeA	Encoder	20.18 ± 2.12	0.79 ± 0.05	1.26 ± 0.03	5.11 ± 0.29	2.05 ± 0.05	0.70 ± 0.04
SeA	Decoder	20.18 ± 2.11	0.79 ± 0.05	1.25 ± 0.03	5.11 ± 0.29	2.05 ± 0.05	0.70 ± 0.04
SeA	Encoder-Decoder	20.07 ± 2.18	0.79 ± 0.05	1.26 ± 0.03	5.12 ± 0.29	2.05 ± 0.05	0.70 ± 0.04

Table 7. Performance Metrics for Fused MRI Images Compared with T1ce and T2 Modalities.

	Model	PSNR (dB)	SSIM	NMI	Entropy	MI	$Q^{AB/F}$
	VAE	13.45 ± 2.24	0.59 ± 0.05	1.00 ± 0.00	4.10 ± 0.41	1.13 ± 0.07	0.40 ± 0.05
	MedFusionGAN	22.51 ± 2.25	0.81 ± 0.06	1.27 ± 0.03	5.04 ± 0.30	2.21 ± 0.05	0.75 ± 0.03
	DDPM	22.39 ± 1.37	0.57 ± 0.05	1.20 ± 0.02	4.81 ± 0.32	1.82 ± 0.06	0.60 ± 0.04
	cGAN	11.43 ± 1.27	0.08 ± 0.01	1.02 ± 0.00	4.52 ± 0.38	1.21 ± 0.06	0.41 ± 0.05
	Pix2Pix	22.91 ± 2.35	0.81 ± 0.06	1.27 ± 0.03	5.08 ± 0.29	2.18 ± 0.05	0.72 ± 0.03
	Attention-Based Pix2Pix						
SA	Encoder	23.07 ± 2.36	0.83 ± 0.06	1.28 ± 0.04	5.22 ± 0.28	2.28 ± 0.04	0.76 ± 0.03
SA	Decoder	23.04 ± 2.36	0.83 ± 0.06	1.28 ± 0.04	5.19 ± 0.28	2.25 ± 0.04	0.75 ± 0.03
SA	Encoder-Decoder	23.10 ± 2.37	0.83 ± 0.06	1.28 ± 0.04	5.39 ± 0.26	2.38 ± 0.04	0.78 ± 0.03
CA	Encoder	23.04 ± 2.47	0.83 ± 0.06	1.28 ± 0.04	5.34 ± 0.27	2.35 ± 0.04	0.77 ± 0.03
CA	Decoder	22.56 ± 2.60	0.78 ± 0.06	1.26 ± 0.03	5.01 ± 0.29	2.12 ± 0.05	0.71 ± 0.04
CA	Encoder-Decoder	23.04 ± 2.37	0.82 ± 0.06	1.28 ± 0.04	5.26 ± 0.28	2.31 ± 0.04	0.76 ± 0.03
SeA	Encoder	22.69 ± 2.33	0.81 ± 0.05	1.27 ± 0.03	5.14 ± 0.28	2.22 ± 0.05	0.74 ± 0.03
SeA	Decoder	22.75 ± 2.21	0.81 ± 0.06	1.27 ± 0.03	5.11 ± 0.29	2.20 ± 0.05	0.73 ± 0.03
SeA	Encoder-Decoder	22.99 ± 2.33	0.82 ± 0.06	1.27 ± 0.03	5.24 ± 0.28	2.29 ± 0.04	0.76 ± 0.03

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.

MedFusionGAN and Pix2Pix achieve strong and stable performance, with MedFusionGAN reaching 22.51 ± 2.25 dB PSNR and 0.81 ± 0.06 SSIM, and Pix2Pix achieving 22.91 ± 2.35 dB PSNR and 0.81 ± 0.06 SSIM.

Among attention-enhanced Pix2Pix variants, all configurations consistently improve fusion quality across metrics. The SA-Pix2Pix Encoder-Decoder model achieves the best overall performance with 23.10 ± 2.37 dB PSNR, 0.83 ± 0.06 SSIM, and the highest mutual information (2.38 ± 0.04), along with strong edge preservation ($Q^{AB/F}$: 0.78 ± 0.03). The CA and SeA variants also demonstrate competitive performance, with PSNR values ranging from 22.56 ± 2.60 dB to 23.04 ± 2.47 dB and SSIM values between 0.78 ± 0.06 and 0.83 ± 0.06.

The quantitative results across all modality pairs consistently demonstrate the superiority of attention-enhanced Pix2Pix variants over baseline methods. While VAE-MMIF and cGAN generally produce lower reconstruction quality, and DDPM shows inconsistent structural preservation despite competitive PSNR in some cases, Pix2Pix and MedFusionGAN provide stable and strong baseline performance across all metrics. Notably, attention mechanisms further enhance fusion quality, with SA-based Encoder-Decoder configurations achieving the best overall results in most modality pairs in terms of PSNR, SSIM, mutual information, and $Q^{AB/F}$ scores. In particular, fusion involving T1ce-T2 and FLAIR-T1ce modalities appears more challenging, where the benefits of attention mechanisms are more pronounced. These findings confirm that incorporating spatial and channel attention significantly improves both structural consistency and information preservation in multimodal MRI fusion.

VI. Discussion

Our comprehensive evaluation across six MRI modality pairs reveals consistent performance trends among the evaluated fusion models and demonstrates the effectiveness of integrating attention mechanisms into the Pix2Pix framework. Quantitative results show that attention-enhanced Pix2Pix variants consistently outperform the baseline cGAN, VAE-MMIF, DDPM, and standard Pix2Pix models, particularly in terms of structural preservation, statistical similarity, and edge retention.

Across nearly all modality pairs, spatial attention variants achieved the highest SSIM and NMI values, indicating stronger anatomical consistency between fused images and source modalities. For example, in the FLAIR-T2 modality pair, the SA Encoder-Decoder configuration achieved 23.92 ± 3.40 dB PSNR, 0.90 ± 0.04 SSIM, 1.31 ± 0.03 NMI, and 2.45 ± 0.04 MI, outperforming the standard Pix2Pix baseline (22.28 ± 2.31 dB PSNR and 0.84 ± 0.05 SSIM). Similarly, for FLAIR-T1ce fusion, the SA Encoder-Decoder model achieved the highest PSNR of 25.13 ± 3.28 dB with 0.89 ± 0.03 SSIM and 2.42 ± 0.04 MI, demonstrating improved preservation of complementary structural and intensity information. These improvements indicate that spatial attention effectively suppresses irrelevant background responses while emphasizing anatomically informative regions during feature fusion.

The differential effectiveness of attention mechanisms across modality pairs can be explained by MRI contrast characteristics. In FLAIR-T2 fusion, FLAIR suppresses cerebrospinal fluid (CSF) signals while T2 highlights fluid-sensitive regions, producing strong complementary information. In this setting, spatial attention significantly improves structural boundary preservation and edge consistency, as reflected in the highest $Q^{AB/F}$ score of 0.81 ± 0.03 obtained by the SA Encoder-Decoder configuration. Conversely, modality pairs with greater structural similarity, such as T1-T1ce, exhibit smaller but more stable improvements. In this case, SA-based configurations achieved SSIM values up to 0.91 ± 0.04, and MI values up to 2.43 ± 0.04, indicating that attention mechanisms remain beneficial even when source modalities share similar anatomical characteristics.

Channel attention variants demonstrated more variable behavior across experiments. While CA Encoder configurations frequently achieved competitive PSNR values, decoder-only channel attention generally produced weaker structural fidelity. For example, in FLAIR-T1ce fusion, the CA Decoder model achieved only 0.76 ± 0.04 SSIM compared with 0.88 ± 0.04 for the CA Encoder configuration. Similarly, in T1ce-T2 fusion, the CA Decoder configuration achieved lower edge preservation ($Q^{AB/F}$: 0.71 ± 0.04) than the CA Encoder model (0.77 ± 0.03). These findings suggest that channel recalibration alone is insufficient to resolve complex spatial inconsistencies between modalities, particularly in highly heterogeneous fusion tasks. Encoder-decoder CA integration generally improved PSNR and entropy values but occasionally resulted in slightly lower SSIM values, indicating a trade-off between pixel-level reconstruction accuracy and perceptual structural quality.

Self-attention enhanced (SeA) variants produced stable but comparatively moderate improvements across most modality pairs. For instance, in T1-T2 fusion, SeA Encoder-Decoder achieved 22.30 ± 2.28 dB PSNR and 0.83 ± 0.05 SSIM, closely matching the standard Pix2Pix baseline while remaining below SA-based configurations. This behavior suggests that global dependency modeling is more beneficial when modalities share strong structural correspondence,

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.

whereas local spatial attention mechanisms are more effective for modality pairs exhibiting substantial contrast variability and heterogeneous anatomical emphasis.

Table 8. Qualitative Impact of Attention Mechanism and Placement.

Attention Mechanism	Encoder	Decoder	Encoder-Decoder
Spatial Attention (SA)	High	Moderate	High
Channel Attention (CA)	High	Low-Moderate	Moderate
Self-Attention (SeA)	Moderate	Moderate	Moderate-High

The placement of attention modules within the generator architecture also strongly influenced performance, Table . Encoder-only attention improved feature representation during downsampling, decoder-only attention enhanced reconstruction sharpness, and combined encoder-decoder attention consistently achieved the strongest overall balance between global structure preservation and local texture refinement. This trend was particularly evident in FLAIR-T1ce and FLAIR-T2 fusion tasks, where encoder-decoder SA configurations achieved the highest PSNR, MI, entropy, and $Q^{AB/F}$ values simultaneously. These results indicate that distributing attention across multiple network stages enables more effective multimodal feature aggregation and reconstruction.

Comparison with recent deep learning-based medical image fusion studies further highlights the competitiveness of the proposed framework. Prior studies commonly report PSNR values ranging from approximately 20-24 dB and SSIM values between 0.80 and 0.88 for multimodal MRI fusion tasks. In contrast, the proposed SA-enhanced Pix2Pix variants consistently exceeded these ranges in several modality pairs, achieving PSNR values above 23-25 dB and SSIM values reaching 0.90-0.91. In addition, improvements in MI and $Q^{AB/F}$ demonstrate enhanced information integration and edge preservation compared with conventional GAN-based fusion strategies. Unlike many previous approaches that investigate a single attention mechanism or placement strategy, this study systematically evaluates encoder, decoder, and encoder-decoder attention integration, providing a broader understanding of how attention design influences fusion quality.

Performance differences across modality pairs further reveal the importance of modality characteristics in determining fusion difficulty. The T1ce-T2 modality pair represents the most challenging scenario for all evaluated models. Baseline cGAN performance degraded substantially in this setting, achieving only 11.43 ± 1.27 dB PSNR and 0.08 ± 0.01 SSIM, indicating severe structural inconsistency and ineffective fusion. This degradation likely results from strong contrast differences between enhancing lesions in T1ce images and edema-related intensity distributions in T2 images. In this challenging scenario, spatial attention mechanisms provided the largest relative improvements, with the SA Encoder-Decoder configuration increasing PSNR to 23.10 ± 2.37 dB and SSIM to 0.83 ± 0.06. In contrast, structurally similar modality pairs such as T1-T1ce showed smaller but more stable gains, suggesting that attention mechanisms are particularly advantageous for heterogeneous multimodal fusion tasks.

Several limitations should be acknowledged. First, supervised training relied on a pseudo-ground-truth generated using the arithmetic average of the source modalities, defined as $((m_1 + m_2)/2)$. Although this strategy enables controlled quantitative comparison, it does not necessarily represent a clinically optimal fusion target. Consequently, absolute PSNR and SSIM values should be interpreted cautiously, while relative performance comparisons remain valid because all models were evaluated under identical conditions. Second, the current framework processes MRI data using 2D slice-based architectures, which limits the ability to capture volumetric inter-slice contextual dependencies present in full 3D MRI acquisitions.

Future work should investigate 3D or pseudo-3D fusion architectures capable of exploiting volumetric anatomical continuity. In addition, unsupervised and weakly supervised fusion strategies, including cycle-consistent learning and perceptually guided optimization, may improve the clinical realism and diagnostic utility of fused images. Incorporating task-driven objectives, such as tumor segmentation or disease classification losses, could further enhance the relevance of multimodal fusion for downstream clinical applications. Overall, the experimental results demonstrate that attention mechanisms significantly improve multimodal MRI fusion performance within the Pix2Pix framework. Spatial attention, particularly when integrated in both encoder and decoder stages, consistently achieves the best balance between structural preservation, information integration, and perceptual reconstruction quality. The findings further indicate that optimal attention design should be adapted to modality-specific characteristics, balancing global structural consistency and local texture preservation to achieve robust multimodal fusion performance quality.

VII. Conclusion

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.

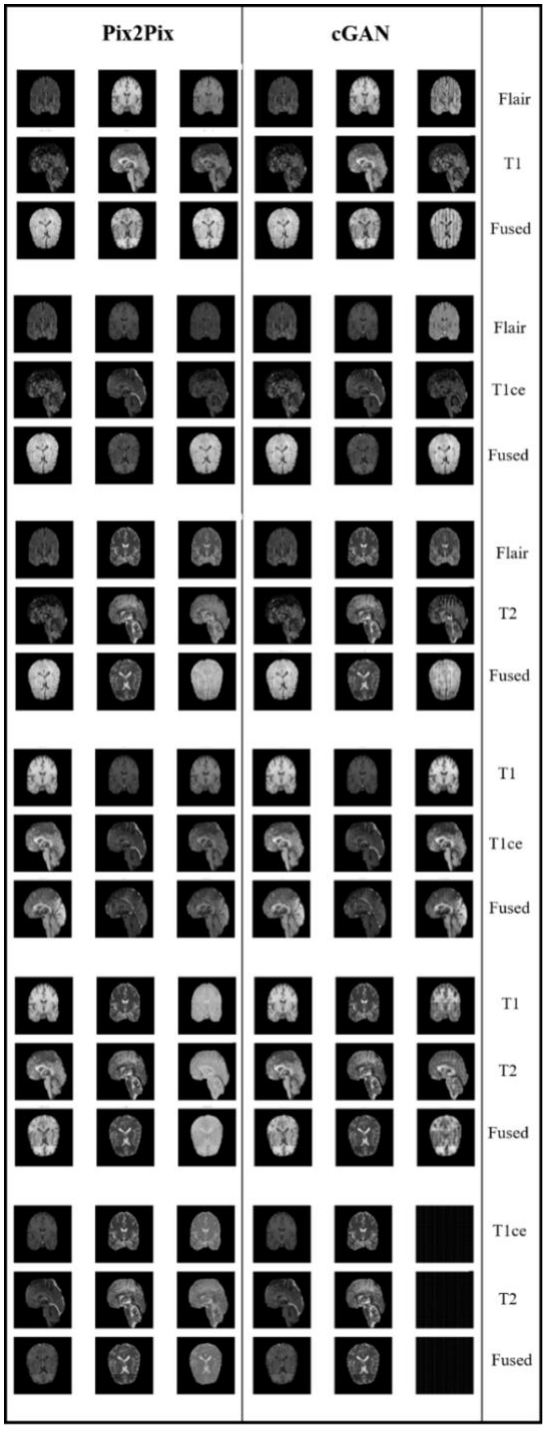

Fig. 3. cGAN VS Pix2Pix Samples.

This study systematically examined the effects of three types of attention mechanisms, i.e., spatial attention, channel attention, and self-attention mechanisms in a Pix2Pix framework for brain MRI fusion on the BraTS20 dataset.

All variants were tested in six combinations of modalities The results show that, in contrast to the cGAN and Pix2Pix models, attention mechanisms offer a consistent improvement in fusion quality. In comparison with all approaches, the SA-Pix2Pix-Encoder-Decoder outperforms other approaches with the highest overall score of SSIM of 0.9125, NMI reaching up to 1.3621, and a good PSNR value of approximately 25.12 dB for some modality pairs. The most difficult case is T1ce-T2 fusion. As the difficulty of all basic models increases significantly, the proposed attention-based model achieves the best performance (SSIM ≈ 0.8331, PSNR ≈ 23.10 dB). Structurally similar pairs like T1-T1ce are, on the other hand, the most stable and are scored best by all models. In addition, channel attention and self-attention also improve, but the extent of the improvement depends on the pair of modalities and the placement of the network. The encoder-decoder model is found to be the best, in both SSIM and NMI, as it simultaneously extracts features and reconstructs the image. Generally, the encoder-decoder configuration performs better in both SSIM and NMI than the encoder-only and decoder-only configurations. Although the model is trained on a simple pseudo-ground-truth (the average of the two input modalities), the attention-augmented architectures still generate more coherent and artifact-free fused images. This implies that, although the training targets are imperfect, the architecture can compensate for their shortcomings, and the relative performance holds well across all metrics. Future research will focus on eliminating the need for pseudo-ground-truth by investigating unsupervised and/or cycle-consistent learning frameworks. Moreover, we will apply this method to full 3D MRI volumes and design new hybrid attention mechanisms that fuse spatial and channel attention for improved performance and clinical applications.

Acknowledgment

The authors would like to express their sincere gratitude to the Artificial Intelligence Center at Chadli Benjedid University for providing the necessary support and research environment that contributed to the completion of this work.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

This study used the publicly available BraTS 2020 dataset. The dataset is accessible through the Center for Biomedical Image Computing & Analytics (CBICA).

Author Contribution

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.

Ali-Abdelatif Betouil and Abdelmadjid Benmachiche contributed to the conceptualization and design of the study. Khadija Rais conducted the literature review, performed the experiments and analysis, and drafted the manuscript. Amel Sahki and Imene Soualmia contributed to supervision, validation of results, and critical revision of the manuscript. All authors read and approved the final manuscript.

Declarations

Ethical Approval

Not applicable. This study does not involve human participants or animal subjects.

Consent for Publication Participants.

Not applicable.

Competing Interests

The authors declare that they have no competing interests.

References

[1] G. Ali, S. Shah, M. AElAffendi, M. Asim, and M. Hammad, "The evolving landscape of few shot learning in medical image diagnosis a scoping review," *Discov. Appl. Sci.*, vol. 8, no. 2, p. 213, Jan. 2026, doi: 10.1007/s42452-025-08188-3.

[2] S. Ullah Khan, M. Ahmad Khan, M. Azhar, F. Khan, Y. Lee, and M. Javed, "Multimodal medical image fusion towards future research: A review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 8, p. 101733, Sep. 2023, doi: 10.1016/j.jksuci.2023.101733.

[3] Z. Zhou, J. Wu, J. Jiang, M. Zhou, W. Guo, and Y. Hu, "A review of multimodal medical image fusion Developments in traditional, model-based and learning-based approaches," *Perioper. Precis. Med.*, pp. 152–167, Dec. 2025, doi: 10.61189/617079irudnn.

[4] A. Bouamrane, M. Derdour, A. Bennour, A. Benmachiche, and M. Gasmi, "Machine Learning for Medical Image Analysis," in *AI for Medical Image Analysis: Reconciling Innovation and Ethical Considerations*, N. Ben Aoun, S. Ahmad, and M. Hammad, Eds., Cham: Springer Nature Switzerland, 2026, pp. 97–125. doi: 10.1007/978-3-032-02963-8_4.

[5] Q. Zhang *et al.*, "Multimodal Fusion on Low-quality Data: A Comprehensive Survey," *Inf. Fusion*, p. 104437, May 2026, doi: 10.1016/j.inffus.2026.104437.

[6] K. Rais, M. Amroune, M. Y. Haouam, A. Benmachiche, and S. Abid, "Dynamic feature context activation and data augmentation for enhanced medical image segmentation," *Multimed. Tools Appl.*, vol. 85, Feb. 2026, doi: 10.1007/s11042-026-21296-5.

[7] M. Haribabu, V. Guruviah, and P. Yogarajah, "Recent advancements in multimodal medical image fusion techniques for better diagnosis: an overview," *Curr. Med. Imaging Rev.*, vol. 19, no. 7, pp. 673–694, 2023, doi: 10.2174/1573405618666220606161137.

[8] T. Tirupal, B. C. Mohan, and S. S. Kumar, "Multimodal Medical Image Fusion Techniques – A Review," *Curr. Signal Transduct. Ther.*, vol. 16, no. 2, pp. 142–163, Aug. 2021, doi: 10.2174/1574362415666200226103116.

[9] G. Dai *et al.*, "Prompt-level contrastive learning for context-aware multi-modal image representation in medical diagnosis," *Pattern Recognit.*, vol. 174, p. 113027, Jun. 2026, doi: 10.1016/j.patcog.2025.113027.

[10] W. Li, P. Jia, D. He, S. Liu, G. Wang, and Y. Huang, "SAFusion: Scenario-Adaptive Network for Multimodal Medical Image Fusion," *IEEE J. Biomed. Health Inform.*, pp. 1–14, 2026, doi: 10.1109/JBHI.2026.3651957.

[11] C. Yu, J. Ye, Y. Liu, X. Zhang, and Z. Zhang, "AMF-MedIT: An efficient align-modulation-fusion framework for medical image–tabular data," *Biomed. Signal Process. Control*, vol. 118, p. 109772, 2026, doi: 10.1016/j.bspc.2026.109772.

[12] Sa. I. Ibrahim, M. A. Makhlouf, and Gh. S. El-Tawel, "Multimodal medical image fusion algorithm based on pulse coupled neural networks and nonsubsampled contourlet transform," *Med. Biol. Eng. Comput.*, vol. 61, no. 1, pp. 155–177, Jan. 2023, doi: 10.1007/s11517-022-02697-8.

[13] M. Rafiq, A. Maurya, P. Singh, and M. Diwakar, "Laplacian Pyramid-Based Fusion with Contrast–Entropy Attention and Sign-Consistent Softmax for Enhanced Multimodal Medical Imaging," in *2026 2nd International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI)*, Feb. 2026, pp. 702–706. doi: 10.1109/IC3ECSBHI67834.2026.11468941.

[14] F. Yang, M. Jia, L. Lu, and M. Yin, "Adaptive zero-learning medical image fusion," *Biomed. Signal Process. Control*, vol. 84, p. 105008, Jul. 2023, doi: 10.1016/j.bspc.2023.105008.

[15] X. Feng *et al.*, "MMIF-VAEFusion: An end-to-end multi-modal medical image fusion network using vector quantized variational auto-encoder," *Biomed. Signal Process. Control*, vol. 102, p. 107407, Apr. 2025, doi: 10.1016/j.bspc.2024.107407.

[16] M. Safari, A. Fatemi, and L. Archambault, "MedFusionGAN: multimodal medical image fusion using an unsupervised deep generative adversarial network," *BMC Med. Imaging*, vol. 23,

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.

no. 1, p. 203, 2023, doi: 10.1186/s12880-023-01160-w.

[17] T. Zhang, X. Yang, R. Lu, D. Zhang, X. Xie, and Z. Zhu, "Modal Feature Disentanglement and Contribution Estimation for Multimodality Image Fusion," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–16, 2025, doi: 10.1109/TIM.2025.3545534.

[18] J. Huang, T. Tan, X. Li, T. Ye, and Y. Wu, "Multiple attention channels aggregated network for multimodal medical image fusion," *Med. Phys.*, vol. 52, no. 4, pp. 2356–2374, 2025, doi: 10.1002/mp.17607.

[19] W. Tang, F. He, Y. Liu, and Y. Duan, "MATR: Multimodal Medical Image Fusion via Multiscale Adaptive Transformer," *IEEE Trans. Image Process.*, vol. 31, pp. 5134–5149, 2022, doi: 10.1109/TIP.2022.3193288.

[20] Z. Zhang, T. Zhang, and Y. Sun, "FMTFuse: Edge Fourier-Enhanced Multi-Scale Transformer for Multi-Modal Image Fusion," in *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2026, pp. 12102–12106. doi: 10.1109/ICASSP55912.2026.11463388.

[21] W. Wang *et al.*, "MDC-RHT: Multi-Modal Medical Image Fusion via Multi-Dimensional Dynamic Convolution and Residual Hybrid Transformer," *Sensors*, vol. 24, no. 13, Jun. 2024, doi: 10.3390/s24134056.

[22] F. Luo, D. Wu, L. R. Pino, and W. Ding, "A novel multimodel medical image fusion framework with edge enhancement and cross-scale transformer," *Sci. Rep.*, vol. 15, no. 1, p. 11657, Apr. 2025, doi: 10.1038/s41598-025-93616-y.

[23] R. He *et al.*, "Multiscale self-attention convolution and adaptive fusion for enhanced multimodal medical image fusion," *Expert Syst. Appl.*, vol. 299, p. 129967, Mar. 2026, doi: 10.1016/j.eswa.2025.129967.

[24] D. Cao, J. Wang, J. Yan, Z. Chen, X. Liao, and H. Cheng, "Neighborhood-Attention-Based Multiscale Alignment and Hierarchical Reconstruction for Multimodal Medical Image Fusion," *ACM Trans Multimed. Comput Commun Appl*, février 2026, doi: 10.1145/3797039.

[25] Z. Zhao *et al.*, "DDFM: Denoising Diffusion Model for Multi-Modality Image Fusion," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France: IEEE, Oct. 2023, pp. 8048–8059. doi: 10.1109/ICCV51070.2023.00742.

[26] G. C. Kumar, K. M, J. S, and N. S, "Structured constraints based Deep guided Generative adversarial network(GAN) for deformable multimodal medical image fusion(MMIF) and enhancement," in *2025 2nd International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS)*, Jul. 2025, pp. 1–5. doi: 10.1109/ICCAMS65118.2025.11234098.

[27] H. Song, Y. Mao, J. Feng, and M. Ye, "MAPD-Mamba: Modality-Adaptive Perception-Driven Mamba Fusion Network," in *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2026, pp. 5306–5310. doi: 10.1109/ICASSP55912.2026.11462025.

[28] L. Xie *et al.*, "Deep label fusion: A generalizable hybrid multi-atlas and deep convolutional neural network for medical image segmentation," *Med. Image Anal.*, vol. 83, p. 102683, Jan. 2023, doi: 10.1016/j.media.2022.102683.

[29] L. Huang, T. Denoeux, P. Vera, and S. Ruan, "Evidence Fusion with Contextual Discounting for Multi-modality Medical Image Segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds., Cham: Springer Nature Switzerland, 2022, pp. 401–411. doi: 10.1007/978-3-031-16443-9_39.

[30] L. Tang *et al.*, "GAN-Guided Few-Shot Attention Network for Medical Images Fusion Quality Assessment," *IEEE Trans. Med. Imaging*, vol. 44, no. 11, pp. 4292–4306, Nov. 2025, doi: 10.1109/TMI.2025.3572511.

[31] Y. Wang *et al.*, "RFSC: Multimodal medical image alignment fusion diagnostic classification network based on de discriminator image translation," *Biomed. Signal Process. Control*, vol. 109, p. 107905, Nov. 2025, doi: 10.1016/j.bspc.2025.107905.

[32] X. Huo, G. Sun, S. Tian, Y. Wang, L. Yu, J. Long, W. Zhang, and A. Li, "HiFuse: Hierarchical multi-scale feature fusion network for medical image classification," *Biomedical Signal Processing and Control*, vol. 87, part A, art. no. 105534, Jan. 2024, doi: 10.1016/j.bspc.2023.105534.

[33] J. Yin, J. Peng, X. Li, and J. Wang, "Enhanced Aortic CT Synthesis Based on Multiscale Information Fusion," *IEEE Multimed.*, vol. 32, no. 2, pp. 75–84, Apr. 2025, doi: 10.1109/MMUL.2025.3546908.

[34] Mst. N. Aktar, A. J. Lambert, and M. Pickering, "An automatic fusion algorithm for multi-modal medical images," *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 6, no. 5, pp. 584–598, Sep. 2018, doi: 10.1080/21681163.2017.1304244.

[35] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.

(CVPR), Jul. 2017, pp. 5967–5976. doi: 10.1109/CVPR.2017.632.

[36] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Computer Vision – ECCV 2018*, vol. 11211, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., in Lecture Notes in Computer Science, vol. 11211. , Cham: Springer International Publishing, 2018, pp. 3–19. doi: 10.1007/978-3-030-01234-2_1.

[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7132–7141. doi: 10.1109/CVPR.2018.00745.

[38] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. doi: 10.5555/3295222.3295349.

[39] S. Bakas *et al.*, "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge," Apr. 23, 2019, *arXiv*: arXiv:1811.02629. doi: 10.48550/arXiv.1811.02629.

[40] B. H. Menze *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015, doi: 10.1109/TMI.2014.2377694.

[41] S. Bakas *et al.*, "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," *Sci. Data*, vol. 4, p. 170117, Sep. 2017, doi: 10.1038/sdata.2017.117.

Author Biography

Ali Abdelatif Betouil is an Algerian researcher and academic specializing in computer science and artificial intelligence. He is an Associate Professor (MCB) in the Department of Computer Science at Chadli Bendjedid University of El Tarf. He holds a PhD in Computer Science and an engineering degree in Computer Science. He teaches and conducts research within the Faculty of Science and Technology at the university. His research interests focus on deep learning, artificial intelligence, cloud computing, combinatorial optimization, and multicriteria decision-making. Dr. Betouil has published several scientific papers on advanced topics, including intelligent anomaly detection in autonomous vehicles using hybrid learning methods. He actively contributes to scientific research through peer-reviewed publications and participation in academic and scientific activities at both national and international levels.

Abdelmadjid Benmachiche is a Professor of Computer Science at Chadli Bendjedid University in El Tarf, Algeria, and serves as the Director of the House of Artificial Intelligence. His research interests cover a wide range of domains, including artificial intelligence, cybersecurity, e-learning systems, robotics, medical imaging, and IoT. Over the years, he has made substantial contributions to deep learning, recommendation systems, autonomous navigation, and intelligent educational platforms. His work emphasizes developing innovative, real-world AI-driven solutions that address complex interdisciplinary challenges. Professor Benmachiche has authored and co-authored numerous scientific publications in reputable peer-reviewed journals and international conferences. In addition to his research output, he actively leads and participates in national and international research projects and collaborations. Through these efforts, he continues to advance knowledge, foster innovation, and support the integration of artificial intelligence technologies.

Khadija Rais is a researcher in Artificial Intelligence and Medical Imaging at the Laboratory of Mathematics, Informatics and Systems (LAMIS), Echahid Cheikh Larbi Tebessi University, Algeria. She holds a PhD in Artificial Intelligence and a master's degree in Multimedia and Systems. Her research focuses on deep learning for medical imaging, with expertise in generative models, data augmentation, and intelligent healthcare systems. She has also contributed to interdisciplinary areas, including cybersecurity and e-learning systems. She has authored and co-authored numerous publications in peer-reviewed journals and at national and international conferences, reflecting her active engagement in advancing applied AI research and developing innovative solutions.

Amel Sahki is an Associate Professor (MCB) in the Department of Computer Science at the Faculty of Science and Technology, Chadli Bendjedid University of El Tarf, since September 2023. She holds a PhD in Electronics, specializing in instrumentation and information processing. She teaches electronics, numerical methods, virtualization, and cloud computing at undergraduate and master's levels. Her research interests include sensors, electronic noses (E-Nose), fault detection and

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.

prediction, fire prediction, and artificial intelligence. She has published research on gas sensor modeling and regularly presents at national and international scientific conferences. She supervises final-year student projects and is actively involved in academic research and mentoring activities within her laboratory, contributing to innovation in intelligent sensing systems and AI-based predictive technologies.

Soualmia Imene is a second-year PhD candidate in Computer Science at Chadli Bendjedid University and a researcher at the LIMA Laboratory. Her research interests include artificial intelligence, machine learning, cybersecurity, and intelligent security systems. She has participated in several national and international scientific conferences and is actively involved in research on AI-based cyber threat detection and secure intelligent systems. Her current work focuses on applying advanced artificial intelligence techniques for cybersecurity, anomaly detection, and fraud detection in complex digital environments. She is also interested in developing robust and adaptive models for real-world security applications, with an emphasis on improving detection accuracy and system resilience.

Manuscript received 29 April 2026; Revised 20 May 2026; Accepted 25 May 2026; Available online 01 June 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1720

Note: the original version contained 16 self-citations and that these citations have since been removed.