Optimized Multimodal Healthcare Image Fusion Using U2Net Restormer with Dilated Dense Encoder–Decoder and Haar-Based Feature Selection

Muhammad Shafiq[1,*], Waeal J. Obidallah[2], Mubarak Albathan[2] and Tahir Kamal[3]

[1] School of Computer Science, Shandong Xiehe University, Jinan, China

[2] College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

[3] School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China

Revista Internacional
Métodos numéricos
para cálculo y diseño en ingeniería

RIMNI

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

In cooperation with
CIMNE

INFORMATION

Keywords:
Multimodal medical image fusion
U2Net Restormer
Haar-based feature selection
AANLIB dataset
clinical interpretability

DOI: 10.23967/j.rimni.2026.10.75903

SCIPEDIA Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería

Revista Internacional Métodos numéricos para cálculo y diseño en ingeniería RIMNI UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH In cooperation with CIMNE

Optimized Multimodal Healthcare Image Fusion Using U2Net Restormer with Dilated Dense Encoder–Decoder and Haar-Based Feature Selection

Muhammad Shafiq[1,*], Waeal J. Obidallah[2], Mubarak Albathan[2] and Tahir Kamal[3]

[1]School of Computer Science, Shandong Xiehe University, Jinan, China

[2]College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

[3]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China

ABSTRACT

Multimodal medical imaging plays a pivotal role in clinical diagnostics by integrating complementary anatomical and functional information from modalities such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and Single-Photon Emission Computed Tomography (SPECT). Despite notable progress, existing fusion approaches continue to face persistent challenges. Convolutional Neural Network (CNN)-based methods often suffer from information loss due to convolutional down-sampling, while Transformer architectures, though effective at capturing global dependencies, incur high computational costs and rely on large-scale pretraining. Generative Adversarial Network (GAN)-based fusion models can generate visually realistic outputs but are prone to training instability and limited reproducibility. In addition, prior studies frequently adopt inconsistent evaluation metrics, with insufficient emphasis on clinical interpretability and robustness, hindering real-world deployment across heterogeneous datasets and institutions. To address these limitations, this study proposes a U-shaped Nested Network – Restoration Transformer (U2Net–Restormer) framework with a Dilated Dense Encoder–Decoder architecture for robust multimodal medical image fusion. The framework integrates hierarchical multiscale representation learning with residual global contextual refinement. To enhance discriminative capability, an optimized Haar-based feature selection strategy is introduced to preserve high-gradient structural and functional details while reducing feature redundancy. Furthermore, an attention-driven fusion mechanism adaptively weights modality-specific contributions, enabling effective integration of heterogeneous information. The proposed method is evaluated on the Augmented Alzheimer's Neuroimaging Library (AANLIB) multimodal brain imaging dataset, covering CT-MRI, PET-MRI, and SPECT-MRI fusion tasks. Experimental results demonstrate consistent performance gains over state-of-the-art CNN-, Transformer-, and GAN-based methods, achieving Structural Similarity Index Measure (SSIM) up to 0.963, Peak Signal-to-Noise Ratio (PSNR) of 42.1 dB, Feature Mutual Information (FMI) of 0.86, and Edge Preservation Index (EPI) of 0.91, with improvements of at least 4%–6% across modalities. Subjective evaluations by radiologists and neurologists report Likert scores up to 4.8/5 for structural visibility, functional fidelity, and diagnostic value. Robustness analysis under Gaussian noise ($\sigma = 15\%$) further confirms the method's resilience. Overall, the proposed framework delivers high-fidelity, clinically interpretable multimodal fusion suitable for diverse imaging scenarios.

OPEN ACCESS

Received: 10/11/2025

Accepted: 15/01/2026

DOI
10.23967/j.rimni.2026.10.75903

Keywords:
Multimodal medical image fusion
U2Net Restormer
Haar-based feature selection
AANLIB dataset
clinical interpretability

*Correspondence: Muhammad Shafiq (shafiqk786@hotmail.com).

Nomenclature

CT	Computed Tomography
MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography
SPECT	Single-Photon Emission Computed Tomography
CNNs	Convolutional Neural Networks
GANs	Generative Adversarial Networks
HMF	Hypergraph-based Multi-modal Fusion
RSCF	Residual Swin-Convolution Fusion

1 Introduction

Multimodal medical imaging has become a cornerstone in modern diagnostic medicine, providing complementary insights into anatomical, structural, and functional characteristics of biological tissues [1]. Techniques such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and Single-Photon Emission Computed Tomography (SPECT) are widely used to capture diverse clinical information [2]. CT offers high-resolution structural details of bone and tissue density, whereas MRI excels in soft-tissue contrast, enabling precise delineation of brain structures and pathological regions [3,4]. PET and SPECT provide functional information related to metabolic and perfusion, which is critical for diagnosing neurological disorders, tumors, and cardiovascular diseases [5]. Individually, each modality provides valuable yet incomplete diagnostic perspectives [6]. Multimodal image fusion aims to integrate complementary information from these modalities, thereby enhancing diagnostic accuracy, treatment planning, and clinical decision-making [7].

The increasing demand for precise and non-invasive diagnostic solutions has accelerated the development of computational frameworks capable of extracting, aligning, and integrating information from heterogeneous imaging modalities [7]. Multimodal fusion enhances visual interpretability and supports quantitative analyses such as voxel-based morphometry, functional connectivity mapping, and radiomic assessments [8]. Recent advances in deep learning have introduced convolutional neural networks (CNNs), generative adversarial networks (GANs), and transformer-based architectures for multimodal medical image fusion [9]. CNN-based approaches effectively extract local features but often suffer from information loss caused by repeated down-sampling operations, resulting in blurred edges and degraded structural details. Transformer-based methods capture global contextual dependencies but typically require substantial computational resources and large-scale pretraining, limiting their practicality in clinical environments. GAN-based fusion techniques can generate visually realistic images; however, they are prone to training instability, mode collapse, and limited reproducibility. These limitations collectively hinder robust clinical deployment, particularly in real-time or resource-constrained healthcare settings [10].

Several intrinsic challenges further complicate multimodal medical image fusion. First, imaging heterogeneity across modalities introduces substantial differences in spatial resolution, contrast mechanisms, noise distributions, and intensity scales. For example, CT and MRI differ significantly in tissue representation, while PET and SPECT provide lower-resolution functional information with higher noise levels. Second, imaging variability caused by institutional differences, such as scanner vendors, acquisition protocols, reconstruction algorithms, and patient demographics, introduces domain shifts that degrade model generalization across datasets and clinical centers. Third, structural misalignment arising from patient motion or acquisition inconsistencies can significantly reduce fusion quality, necessitating accurate registration prior to fusion. Fourth, conventional fusion algorithms often lose high-gradient information, particularly at anatomical boundaries and lesion edges, leading to blurred or artifact-prone outputs. Finally, many state-of-the-art models remain computationally intensive and struggle to generalize across diverse modality combinations and institutions, limiting their clinical scalability and reliability [11].

Despite notable progress, existing multimodal fusion methods still struggle to balance high-gradient feature preservation, global context modeling, and modality-specific feature weighting [12]. Many CNN-based approaches emphasize local feature extraction but fail to capture long-range dependencies, while transformer-based models focus on global relationships at the cost of computational efficiency. Moreover, traditional feature selection strategies, such as PCA-based dimensionality reduction, often discard subtle yet clinically relevant structures, reducing diagnostic utility. Attention mechanisms, when employed, are frequently applied uniformly, resulting in suboptimal integration of modality-specific contributions. Consequently, there remains a clear need for a hybrid fusion framework that preserves fine anatomical details, retains functional and metabolic information, and remains robust to noise, imaging variability, and institutional heterogeneity.

This study addresses these challenges by proposing a hybrid multimodal medical image fusion framework with the following objectives:

1. To develop a backbone architecture capable of preserving high-gradient anatomical and functional features while capturing multiscale local and global contextual information across CT-MRI, PET-MRI, and SPECT-MRI modalities; and
2. To design an attention-based fusion strategy that adaptively integrates complementary modality-specific information, maximum diagnostic relevance while minimizing artifacts.

The novelty of this work lies in the integration of a U2Net-Restormer backbone with a Dilated Dense Encoder-Decoder architecture, an optimized Haar-based feature selection mechanism, and attention-driven fusion. Specifically, U2Net enables hierarchical multiscale representation learning, while the Restormer module refines global contextual dependencies through efficient channel-wise attention. The Dilated Dense Encoder-Decoder expands the receptive field without sacrificing spatial resolution, improving structural continuity and contextual awareness. The optimized Haar-based block selection preserves high-gradient features while reducing redundancy, and the attention-based fusion dynamically emphasizes modality-relevant information, enhancing clinical interpretability.

The primary contributions of this study are summarized as follows:

1. A hybrid U2Net-Restormer backbone that jointly capture multiscale local features and long-range contextual dependencies.
2. A Dilated Dense Encoder–Decoder architecture that enhances structural fidelity through hierarchical feature aggregation.
3. An optimized Haar-based feature selection strategy that preserves diagnostically significant high-gradient information while reducing computational redundancy; and
4. An attention-based fusion mechanism that adaptively weights modality-specific contributions, improving both anatomical consistency and functional fidelity in fused images.

2 Existing Multimodal Biomedical Image Fusion

Multimodal biomedical image fusion has emerged as a critical research direction for overcoming the inherent limitations of single-modality analysis, enabling more comprehensive disease characterization and informed clinical decision-making. Early research in this domain primarily focused on accurate co-registration and spatial alignment of heterogeneous modalities. Over time, advances in machine learning and deep learning have shifted the focus toward feature-level and decision-level fusion frameworks that integrate complementary anatomical, functional, and molecular information. This evaluation reflects a transition from basic spatial alignment to sophisticated multimodal fusion pipelines designed for diverse biomedical applications.

2.1 Co-Registration and Foundational Frameworks

Accurate alignment of heterogeneous imaging modalities remains a fundamental prerequisite for effective multimodal fusion. Ref. [13] addressed this challenge through SOmicsFusion, a toolbox for integrating spatial omics with imaging modalities such as MRI, microscopy, brain atlas, and spatial transcriptomics. Their two-stage framework combines dimensionality reduction for representational alignment with deep learning-based spatial registration, achieving upto a 69% reduction in coregistration error compared to conventional pipelines. Beyond alignment, the framework supports visualization, spatial correlation analysis, pansharpening, and automated annotation, enabling richer biological interpretation.

Similarly, Ref. [14] proposed a voxel-wise multimodal fusion approach using generative architectures with separable convolutional blocks, significantly reducing parameter complexity while preserving high-dimensional neuroimaging representations. Their embeddings, validated on the Human Connectome Project dataset, revealed stratifiable phenotypic groupings, highlighting the potential of multimodal representations to uncover latent biological markers. Complementing these methodological contributions, Ref. [15] provided a comprehensive review of multimodal learning in biomedicine, identifying persistent challenges related to generalization, interpretability, and domain adaptation.

2.2 Hypergraph and Graph-Based Approaches

Beyond pixel-level alignment, graph-based fusion methods have been explored to model complex interdependencies across modalities. Ref. [16] introduced a Hypergraph-based Multimodal Fusion (HMF) framework that integrates imaging and genetic data by constructing hypergraph similarity matrices with inter- and intra-modality regularization. Applied to schizophrenia studies, HMF effectively captured high-order subject relationships and outperformed competing approaches. Ref. [17] further emphasized multimodal fusion as a means of uncovering latent biological information that remains inaccessible when modalities are analyzed independently, reinforcing the importance of integrative modeling strategies.

2.3 Convolutional and Hybrid CNN-Transformer Models

With the rapid advancement of deep learning, multimodal fusion has increasingly relied on convolutional and hybrid CNN-Transformer architectures. Ref. [18] developed an unsupervised framework combining CNNs with Swin Transformer modules, where a Residual Swin-Convolution Fusion (RSCF) block integrates global contextual information with fine-grained local features. An adaptive weighting mechanism further optimized loss contributions during training. Ref. [19] demonstrated convolutional optimization for MRI-PET fusion using variable kernel sizes, instance normalization, and transposed convolutions, achieving upto 99% accuracy in Alzheimer's disease detection, which underscores the diagnostic value of optimized CNN-based fusion.

To mitigate convolutional information loss, Ref. [20] proposed a Hahn-PCNN-CNN hybrid network, incorporating pulse-coupled neural networks for enhanced feature preservation. While these CNN-based approaches excel at local feature extraction and classification performance, they often suffer from information degradation caused by repeated down-sampling and limited receptive fields, particularly affecting high-gradient anatomical boundaries.

Generative models have also gained attention for multimodal fusion and synthesis. Ref. [21] introduced a residual attention GAN that employs texture and structural extraction blocks for fused image reconstruction, refined through adversarial training. Ref. [22] proposed a Coupled-GAN framework that maps MRI and PET images into shared latent spaces for Alzheimer's staging, achieving strong performance on the ADNI dataset. Similarly, Ref. [16] employed a pyramidal attention GAN to synthesize PET images from MRI, achieving 89.9% accuracy in multi-class disease classification. Despite their ability to generate visually realistic outputs and address missing modalities, GAN-based approaches remain susceptible to training instability, mode collapse, and limited reproducibility, which restricts their reliability in clinical practice.

2.4 Transformer-Based Fusion Architectures

Transformer-based fusion architectures have gained prominence due to their capacity to model long-range dependencies and global contextual relationships. Ref. [23] proposed MATR, a multiscale adaptive transformer that integrates adaptive convolution with global semantic extraction guided by structural and mutual information losses. Ref. [24] extended this concept through MsgFusion, which exploits multi-space representations, including spatial and frequency domains for MRI and color-space transformations for PET/CT, and demonstrated superior fusion quality through both objective metrics and clinical evaluations.

In Ref. [25], a hybrid CNN-Transformer encoder-decoder architecture enhanced with global semantic aggregation and self-supervised pre-training. Ref. [26] proposed MDA-ViT, incorporating dual attention mechanisms to capture both intra- and inter-modal dependencies, while Ref. [27] designed a residual hybrid transformer with dynamic convolution to improve quantitative and qualitative fusion performance. These studies highlight the transformer's effectiveness in capturing global multimodal context; however, their high computational complexity and dependence on large-scale training data limit scalability and real-time clinical deployment.

Several recent works emphasize multi-scale feature integration to preserve both structural and functional information. RefineFuse employs dual-attention mechanism and gradual semantic-detail bridging, achieving strong performance across medical and infrared-visible fusion tasks. M^{4}Net leverages hybrid dilated convolutions and multi-scale decomposition with a structural similarity-driven loss, outperforming multiple state-of-the-art methods. PPMF-Net combines dynamic edge enhancement, nonlinear feature interaction, and transformer-based global modeling, demonstrating robust performance on PET-MRI and SPECT-MRI fusion tasks.

Beyond image-only fusion, some studies extend multimodal integration to non-imaging data. Ref. [28] proposed a hypernetwork-based framework conditioning MRI analysis on electronic health record (HER) data, improving brain age prediction and Alzheimer's classification. Ref. [29] introduced a SpatioTemporal Adaptive (STA) model for multimodal blood glucose prediction, integrating physiological constraints with attention mechanisms. Ref. [30] proposed S3IMFusion, incorporating stochastic structural similarity within a CNN-Transformer framework, demonstrating strong generalization across both medical and non-medical datasets. A comparative summary of representative multimodal fusion methods is provided in Table 1.

Despite substantial progress, multimodal biomedical image fusion continues to face significant technical and translational challenges. Early approaches, e.g., [13,31] primarily addressed registration and alignment but struggled with scalability to highly heterogeneous datasets. CNN-based methods, e.g., [19,20] exhibit strong discriminative power yet suffer from information loss and limited global context modeling. Transformer-based architectures address long-range dependency modeling but impose high computational costs, while GAN-based method instability and reproducibility concerns. Hybrid CNN-Transformer frameworks offer improved balance but remain constrained by generalization challenges across institutions and imaging protocols. Furthermore, evaluation practices remain inconsistent, often emphasizing SSIM and PSNR while inadequately addressing clinical interpretability and diagnostic relevance.

Accordingly, there is a clear need for fusion frameworks that emphasize cross-modal generalization, computational efficiency, robust high-gradient feature preservation, and standardized evaluation benchmarks to bridge the gap between algorithmic innovation and real-world biomedical deployment.

Table 1: Summary of multimodal image fusion methods

Method	Algorithm /Framework	Methodology	Outcome
Guo et al. [13]	SOmicsFusion (ML + DL pipeline)	Dimension reduction + deep learning-based spatial alignment for spatial omics and imaging modalities	Reduced coregistration errors (38–69%), improved pixel-wise fusion accuracy
Dimitri et al. [14]	Generative modular DL model	Separable convolutions, voxel-wise multimodal fusion, latent embeddings	Efficient representation (20× fewer parameters), separable phenotypic clustering
Zhang et al. [31]	Hypergraph-based Fusion (HMF)	Hypergraph similarity + inter/intra-modality regularization	Outperformed baselines, identified schizophrenia-related interactions
Xie et al. [18]	RSCF (CNN + Swin Transformer)	Residual Swin-Convolution Fusion with adaptive weighting and joint loss	Preserved global and local features, improved visual quality
Odusami et al. [19]	Optimized CNN + Vision Transformer	Variable kernels, instance normalization, PET-MRI addition	Achieved 99% accuracy in AD vs. CN classification
Guo et al. [20]	Hahn-PCNN-CNN	PCNN-based feature fusion + CNN reconstruction	Superior image reconstruction, high robustness

(Continued)

Table 1 (continued)

Method	Algorithm /Framework	Methodology	Outcome
Guo et al. [21]	Residual Attention GAN	Attention + texture detail GAN blocks, dual discriminators	High-quality fused images across databases
Tang et al. [23]	MATR (Adaptive Transformer)	Adaptive convolution + multiscale Transformer + structural/MI loss	Outperformed SOTA methods in fusion quality
Wen et al. [24]	MsgFusion	Multi-branch CNN using spatial/frequency and color features	Clinically superior fusion (validated by 30 doctors)
Choudhury et al. [22]	Coupled-GAN (CGANC)	MRI-PET latent space encoding + CNN classifier	Accurate AD stage classification
Li et al. [25]	DFENet	CNN + Transformer encoder-decoder with global semantic aggregation	Robust image fusion with self-supervised pretraining
Kalamkar and Amalanathan [26]	MDA-ViT	Dual attention Vision Transformer (intra- & inter-modal)	Improved semantic preservation and fusion adaptability
Zhang et al. [16]	Pyramidal Attention GAN	Synthesized PET from MRI + fusion for AD/MCI/NC	Achieved 89.9% classification accuracy
Song et al. [32]	RefineFuse	Multi-scale dual attention + semantic-detailed interaction	High-quality fusion across medical/infrared tasks
Wang et al. [27]	Residual Hybrid Transformer	Dynamic convolution + multi-scale architecture	High quantitative/qualitative fusion scores
Ding et al. [33]	M4Net	Multi-receptive-field CNN + wavelet multi-scale fusion	Outperformed 6 SOTA methods by ∼10% on fusion metrics
Hsu et al. [34]	MMF-RvNN	Recursive NN + curvelet transforms + t-SNE	SSIM = 0.98, PSNR = 45.9 dB in breast cancer fusion
Yin et al. [29]	STA Model	Spatio-temporal multimodal BG prediction with reversal attention	Achieved low RMSE (18–29.7) in real-world BG prediction

(Continued)

Table 1 (continued)

Method	Algorithm /Framework	Methodology	Outcome
Peng et al. [35]	PPMF-Net	Multi-path CNN-Transformer + edge/semantic modules	Outperformed SOTA with superior structural fidelity
Duenias et al. [28]	Hypernetwork-based Fusion	MRI conditioned on EHR/tabular data	Improved AD classification, brain age prediction
Lv et al. [30]	S3IMFusion	CNN-Transformer with stochastic structural similarity	Strong fusion and generalization across modalities

3 Methods

The proposed framework (Fig. 1) introduces a U2Net-Restormer with a Dilated Dense Encoder–Decoder architecture, integrated with optimized Haar-based feature selection and an attention-driven fusion strategy for robust multimodal healthcare image fusion. The method is evaluated on the AANLIB multimodal brain imaging datasets, covering CT-MRI, PET-MRI, and SPECT-MRI fusion tasks.

- Backbone Network-U2Net Restormer: U2Net's nested U-shaped architecture enables multi-scale feature representation with residual connections to retain fine anatomical details. The Restormer component introduces linear attention mechanisms that efficiently capture global dependencies across modalities.
- Dilated Dense Encoder–Decoder: The encoder employs dilated convolutions to expand the receptive field without sacrificing spatial resolution, while dense connections ensure efficient gradient flow. The decoder reconstructs fused outputs with strong contextual consistency, minimizing artifacts and blurring.
- Optimized Color Feature with Haar-Based Block Selection: Instead of conventional PCA, the model extracts color gradient features and applies a Haar wavelet transform to decompose the input into low-frequency (structural) and high-frequency (detail) sub-bands. A gradient-energy based selection criterion chooses high-informative Haar blocks, discarding redundant components and thereby improving efficiency.
- Attention-Based Fusion: A cross-modality attention mechanism combines the complementary characteristics of CT/MRI, PET/MRI, and SPECT/MRI. The spatial attention focuses on structural alignment, while channel attention regulates modality-specific feature weighting.

Finally, fused representations are reconstructed into high-quality medical images. Residual refinement ensures edge sharpness and diagnostic clarity. The complete multimodal fusion pipeline is summarized in Algorithm 1.

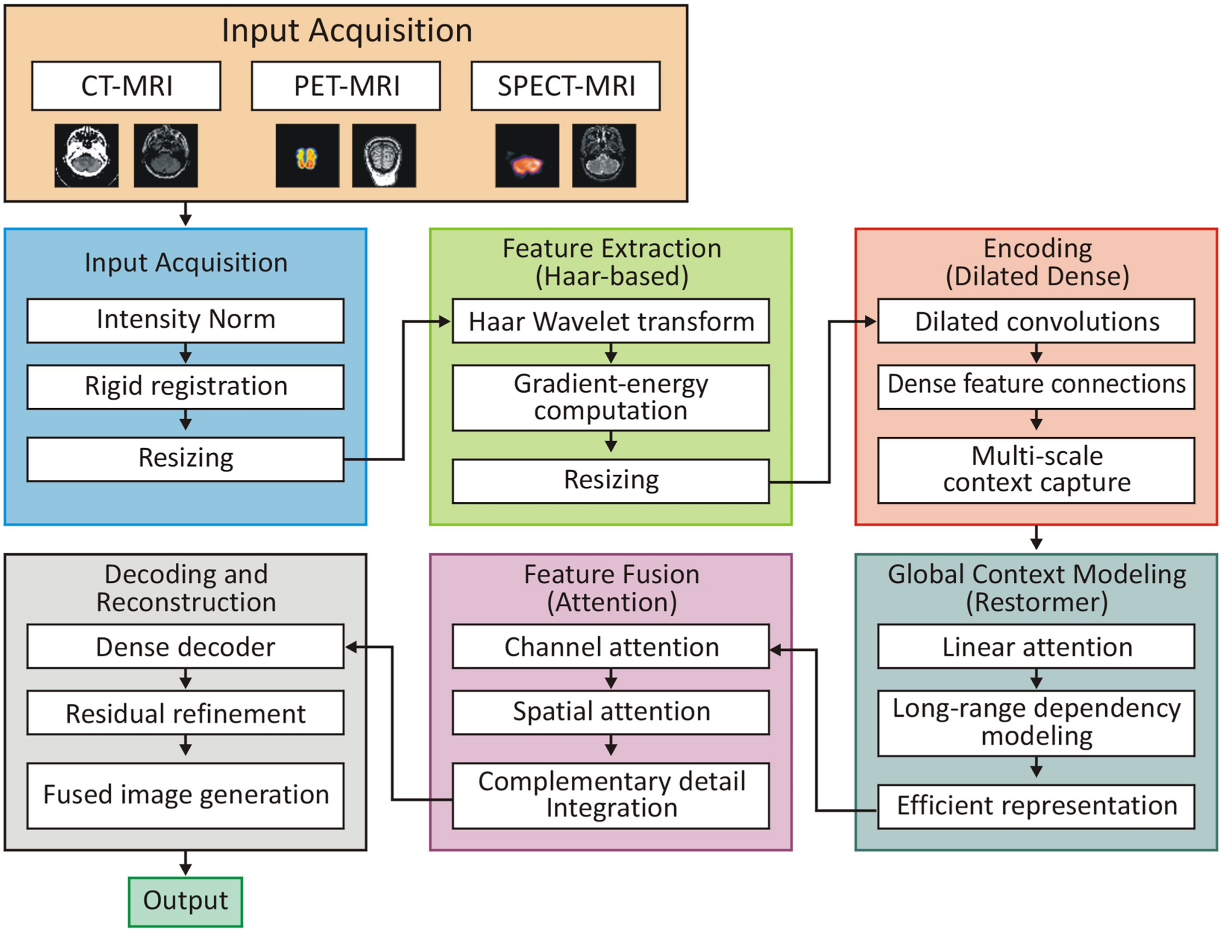


Figure 1: Architectural overview

Algorithm 1: Multimodal fusion-U2Net restormer haar-based feature selection

```
(1) Input: $\{I_\alpha, I_\beta, I_\chi, I_\delta\} \in \mathbb{R}^{H\times W}$
(2) Haar-Based Feature Selection: $\mathcal{F}_m = \{H(\mathcal{I}) \mid G(H(\mathcal{I})) \geq \tau\}, m \in \{\alpha, \beta, \chi, \delta\}$; $\tau$ is a gradient threshold to retain high-informative blocks.
(3) Encoding: $E_m = Enc(\mathcal{F}_m)$
(4) Global Context Modeling (Restormer): $R_m = \mathcal{R}(E_m)$
(5) Attention-Based Fusion: $\breve{F} = Att(\{R_\alpha, R_\beta, R_\chi, R_\delta\})$
(6) Decoding and Reconstruction: $F' = Dec(\breve{F})$
(7) Output: $F'$ with preserved structure.
```

3.1 Input Acquisition and Preprocessing

Multimodal medical images exhibit substantial variation in resolution, intensity scale, and noise characteristics. To ensure spatial and statistical compatibility, all input modalities are preprocessed prior to feature extraction. The AANLIB dataset provides multimodal brain image pairs acquired across institutions, introducing realistic variability in imaging conditions.

Let the acquired multimodal image dataset be denoted as:

$$I = \{I_{CT}\ , I_{MRI}, I_{PET}, I_{SPECT}\ \} \tag{1}$$

where MRI is selected as the anatomical reference due to its superior soft-tissue contrast.

Each image is represented as a 2D matrix $I(x,y) \in R^{H\times Y}$.

3.1.1 Image Registration

Rigid and affine registration are applied to spatially align heterogeneous modalities:

$$I_A = T\,(I_B) \tag{2}$$

where T represents scaling, rotation, and translation parameters.

3.1.2 Intensity Normalization

To harmonize modality-specific intensity distributions, min–max normalization is applied:

$$I_{norm} = \frac{I - I_{\min}}{I_{\max} - I_{\min}} \tag{3}$$

3.1.3 Noise Suppression and Resampling

Non-local means filtering suppresses modality-dependent noise while preserving edges. All images are resampled to a uniform resolution of 256 × 256. Statistical consistency before and after preprocessing is summarized in Table 2.

Table 2: Statistical summary of AANLIB data before and after preprocessing

Modality	Mean intensity (Raw)	Mean intensity (Normalized)	Std. Dev.	Entropy (bits)
CT	450.3 (HU)	122.5	38.6	6.71
MRI	135.8	127.9	41.2	7.03
PET	3.82 (SUV)	119.4	29.5	6.59
SPECT	1752.5 (counts)	121.7	33.8	6.84

The preprocessing stage guarantees that multimodal data from the AANLIB dataset are standardized and comparable. The registration equations align heterogeneous modalities spatially, while intensity normalization ensures that disparate imaging units such as HU (CT), SUV (PET), and raw counts (SPECT) are brought into a unified scale. The noise reduction step removes acquisition artifacts without compromising edge preservation, which is critical for downstream gradient-based Haar block feature selection. Finally, resizing provides a consistent input dimension, and this proves compatibility with the U2Net Restormer encoder–decoder backbone.

3.2 Optimized Haar-Based Feature Extraction

To preserve diagnostically relevant high-gradient information while reducing redundancy, an optimized Haar wavelet-based block selection strategy is employed (Fig. 2). Unlike PCA-based compression, this approach preserves localized structural details critical for medical imaging.

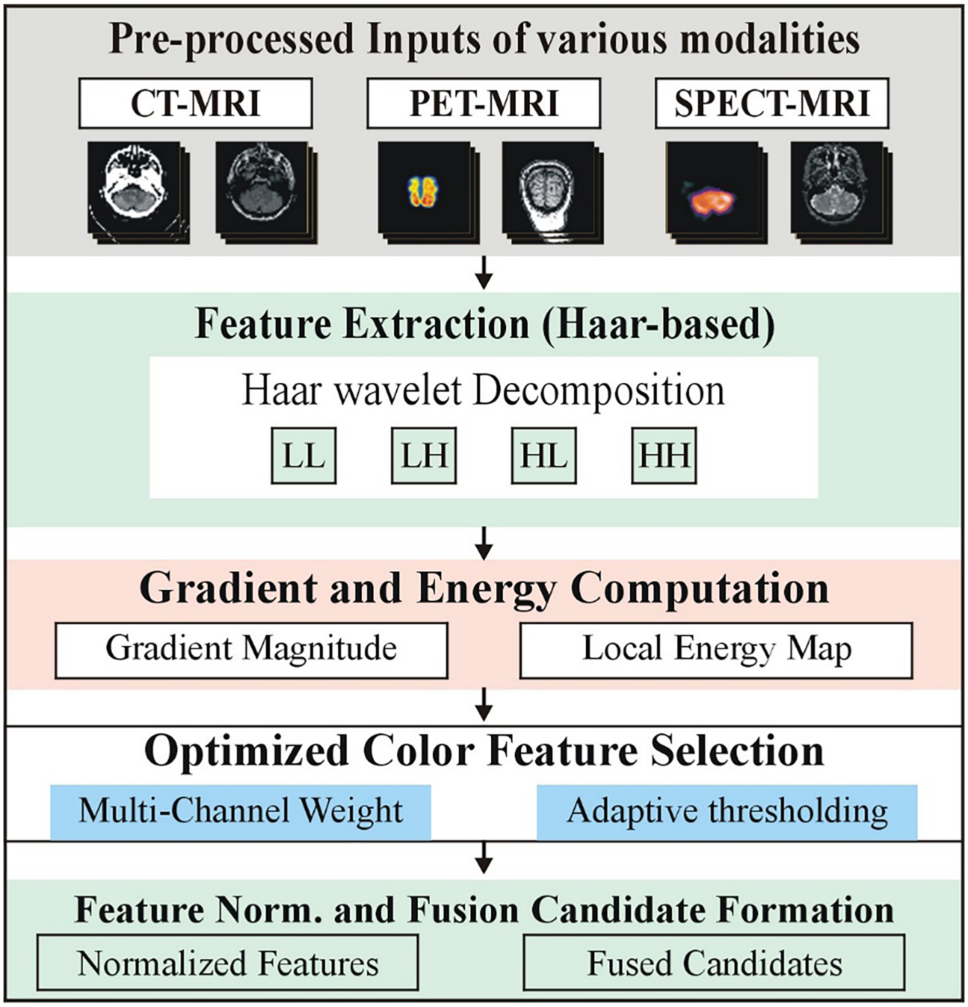

Pre-processed Inputs of various modalities
CT-MRI
PET-MRI
SPECT-MRI
Feature Extraction (Haar-based)
Haar wavelet Decomposition
LL
LH
HL
HH
Gradient and Energy Computation
Gradient Magnitude
Local Energy Map
Optimized Color Feature Selection
Multi-Channel Weight
Adaptive thresholding
Feature Norm. and Fusion Candidate Formation
Normalized Features
Fused Candidates

Figure 2: Feature extraction: optimized Haar-based block

The 2D Haar Wavelet Transform decomposes each image into LL, LH, HL, and HH sub-bands across multiple scales. Each sub-band is partitioned into non-overlapping blocks, and a gradient-energy criterion is applied:

$$E_b = \sum_{x,y} \left(\left| \frac{\partial I}{\partial x} \right|^2 + \left| \frac{\partial I}{\partial y} \right|^2 \right) \tag{4}$$

Blocks exceeding an adaptive threshold are retained, forming an optimized feature set. This strategy achieves over 50% dimensionality reduction while doubling average gradient energy (Tables 3 and 4), ensuring compact yet information-rich representations.

An advantage of Haar-based block selection lies in feature reduction. By discarding redundant low-gradient blocks, computational cost is significantly minimized. Table 3 provides a comparative analysis of feature dimensionality before and after optimized Haar-based block selection.

Table 3: Feature dimensionality reduction via optimized Haar-based block selection

Modality pair	Initial features per image	Features after PCA	Features after haar-based selection	Reduction (%)
CT–MRI	131,072	64,000	59,800	54.4%
PET–MRI	65,536	32,500	30,200	53.9%
SPECT–MRI	32,768	16,500	15,120	53.9%

Table 4: Gradient energy statistics before and after Haar-based block selection

Modality	Mean gradient energy (Raw)	Mean gradient energy (Selected Blocks)	Improvement (%)
CT	12.5	25.7	105.6%
MRI	18.3	36.2	97.8%
PET	9.8	20.1	105.1%
SPECT	11.4	22.7	99.1%

The results in Table 3 show that the proposed Haar-based selection retains only ∼46% of the original features, yielding nearly 54% dimensionality reduction while preserving high-gradient information.

To assess the discriminative capacity of selected features, we analyze the average gradient energy before and after selection across modalities. Results are summarized in Table 4.

The results in Table 4 indicate that block selection effectively doubles the mean gradient energy, thereby ensuring that the retained features emphasize diagnostically significant structural and functional variations.

The optimized Haar-based block feature extraction addresses several limitations identified in previous works: (1) Unlike PCA-based compression, which discards features linearly, Haar block selection preserves localized structural information, crucial for medical imaging. (2) By combining wavelet coefficients with color descriptors, the approach captures both global structural consistency and local chromatic-textural variations. (3) The adaptive gradient-energy criterion ensures that redundant or flat regions are excluded, reducing computational overhead while enhancing feature discriminability. (4) Statistical results (Tables 3 and 4) indicate that the method not only reduces dimensionality by over 50% but also doubles the average gradient energy of the retained features. Thus, the optimized Haar-based feature extraction stage provides a compact and rich representation that improves the efficiency of U2Net Restormer-based fusion.

3.3 Dilated Dense Encoder

The encoded features are processed using a Dilated Dense Encoder (Fig. 3), combining dilated convolutions with dense connectivity to capture both fine-scale details and long-range dependencies. Dilated convolution expands the receptive field without loss of resolution:

$$y(i,j) = \sum_{k} x(i + d.k) \tag{5}$$

Dense connectivity ensures feature reuse and stable gradient propagation:

$$x_l = H_l\left([x_0, x_1, \ldots, x_{l-1}]\right) \tag{6}$$

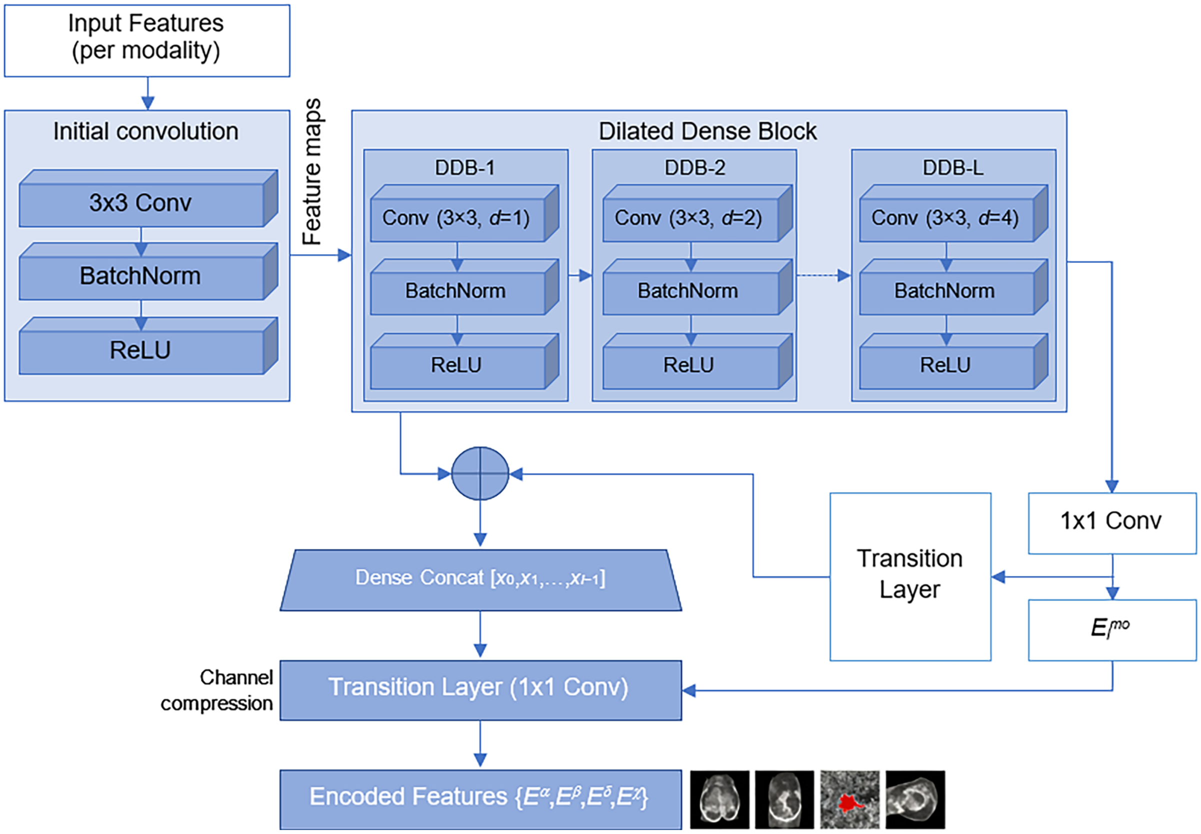


Figure 3: Encoding using dilated dense encoder

Multiple dilated dense blocks with increasing dilation factors are stacked, progressively enriching multiscale representations (Table 5).

Table 5: Encoding process using dilated dense encoder for multimodal input features

Block	Dilation factor (d)	Kernel size (k)	Growth rate (g)	Receptive field (Reff)	Output channels (Cout)
DDB-1	1	3 × 3	16	3 × 3	48
DDB-2	2	3 × 3	16	5 × 5	64
DDB-3	4	3 × 3	16	9 × 9	80
DDB-4	8	3 × 3	16	17 × 17	96

As shown in Table 5, the receptive field expands exponentially with dilation while output channels progressively grow, enriching the representational power of the encoder.

3.4 Global Context Modeling Using Restormer

While convolutional encoders capture local and multiscale features, global contextual dependencies are modeled using a Restormer module (Fig. 4). Restormer employs channel-wise self-attention, reducing computational complexity from quadratic to linear scaling with respect to spatial dimensions.

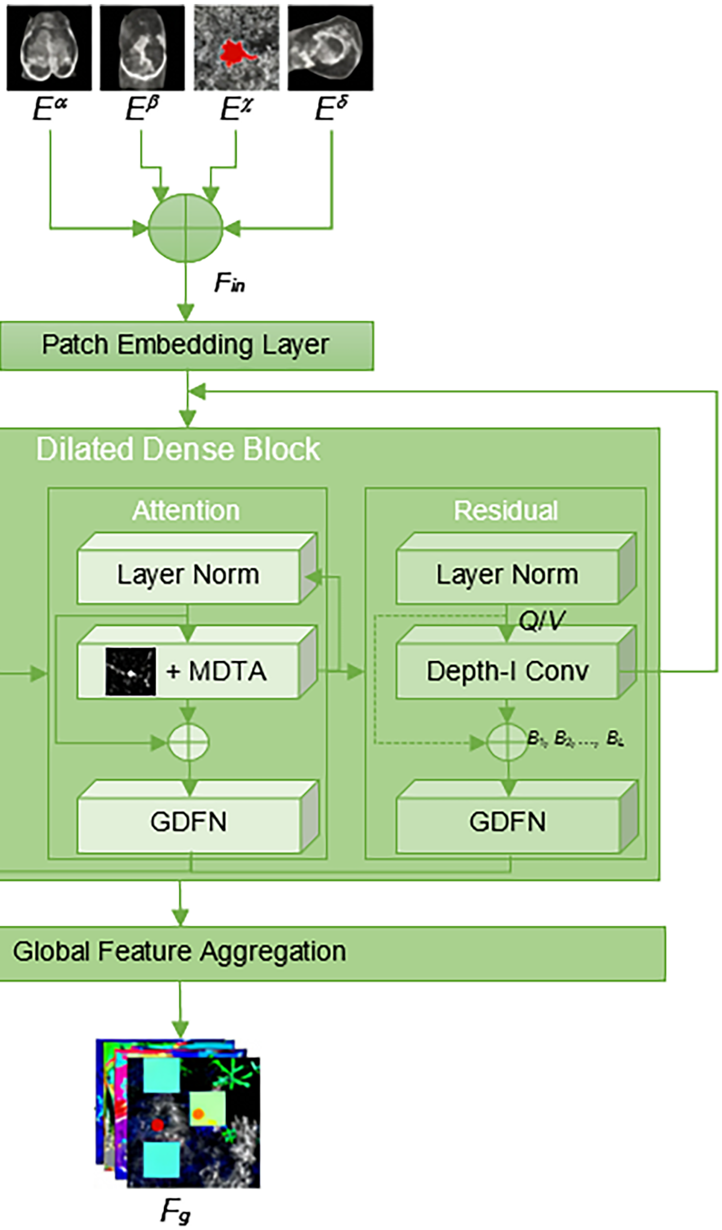


Figure 4: Global context modeling using restormer

Given encoded features E, attention is computed as:

$$A = Softmax\left(QK^{T}\right) V \tag{7}$$

Residual connections ensure stability and feature preservation. Feature magnitude analysis (Table 6) demonstrates that Restormer effectively amplifies diagnostically significant anatomical and functional regions.

Table 6: Global context modeling via Restormer: feature magnitude statistics

Modality	Mean feature magnitude (Before)	Mean feature magnitude (After attention)	Increase (%)
CT	0.124	0.312	151.6%
MRI	0.198	0.412	107.1%
PET	0.101	0.276	173.3%
SPECT	0.115	0.294	155.7%

3.5 Attention-Based Feature Fusion

The globally contextualized features are fused using a dual-attention mechanism combining spatial and channel attention (Fig. 5). Spatial attention highlights region-of-interest locations, while channel attention emphasizes modality-specific discriminative features.

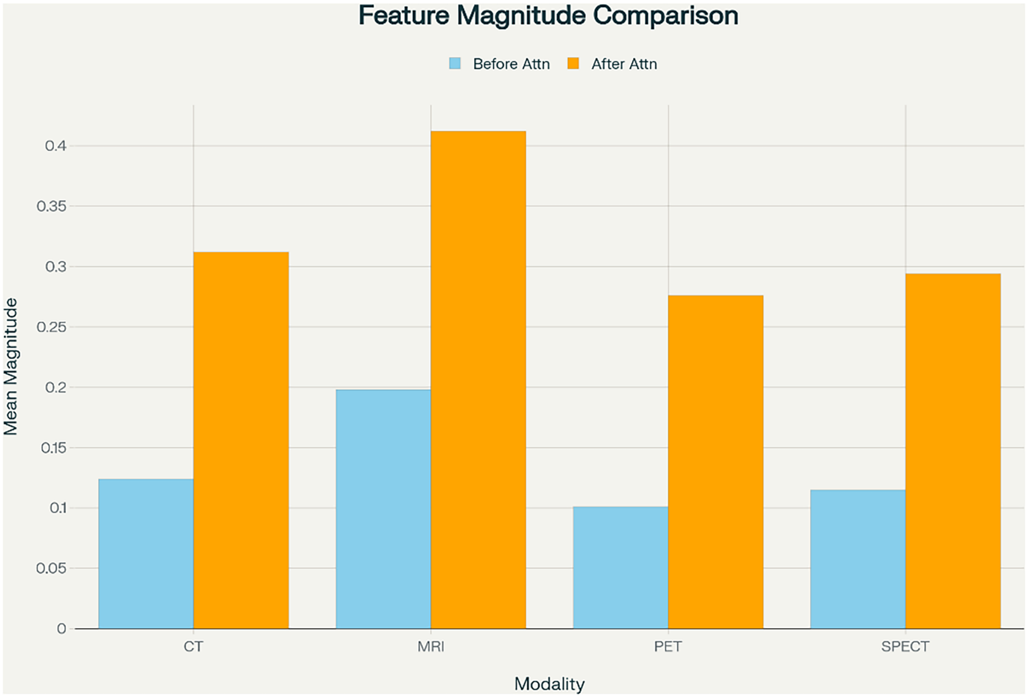


Figure 5: Feature magnitude statistics of global context modeling via Restormer

The fused feature map is computed as:

$$F_{fuzed} = \sum_{m} a_m \odot F_m \tag{8}$$

3.5.1 Motivation

Traditional convolutional networks are inherently local operators with limited receptive fields. Even with dilated convolutions, long-range dependencies across entire medical images may not be captured. Conversely, transformers, with self-attention mechanisms, can model global context, allowing pixels from distant regions to influence feature representations. The Restormer is an optimized variant that: (1) reduces computational complexity using channel-wise attention, avoiding quadratic scaling of standard attention. (2) It maintains high-resolution feature maps, essential for multimodal medical images. (3) It combines with multi-scale features from the Dilated Dense Encoder.

3.5.2 Restormer Architecture

Given encoded feature maps E_{mo} from the Dilated Dense Encoder for modality mo, the Restormer computes global contextual embeddings. Let $F \in \mathbb{R}^{H\times W\times C}$ denote an encoded feature map.

3.5.3 Layer Normalization and Linear Projection

The first step is layer normalization (LN) and linear projection:

$$F' = LN\,(F) \tag{9}$$

$$\begin{aligned} Q &= F' W_Q, \\ K &= F' W_K, \\ V &= F' W_V \end{aligned} \tag{10}$$

where, $Q, K, V \in \mathbb{R}^{H\cdot W\times C}$ represent query, key, and value matrices; W_Q, W_K, W_V are learnable weight matrices. LN stabilizes feature distributions for attention computation.

Multi-Dimensional Attention

Restormer employs multi-dimensional (channel-wise) self-attention, reducing computational complexity from $\mathcal{O}\left((H\cdot W)^2\right)$ to $\mathcal{O}\left(C^2\right)$. Attention scores A are computed as:

$$A = \sigma\left(\frac{Q^T K}{\sqrt{C}}\right) \tag{11}$$

The output of the attention layer is: $F_{att} = AV$. This mechanism captures inter-channel relationships rather than inter-spatial, and this proves efficient modeling of long-range dependencies in high-dimensional feature spaces.

Feed-Forward Network (FFN)

Following attention, features are passed through a feed-forward network:

$$F_{FFN} = W_2\sigma\,(W_1F_{att}) \tag{12}$$

where W_1, W_2 are learnable matrices and σ is a GELU activation. This module enhances non-linear transformations while preserving contextual information.

3.5.4 Residual Connections

Residual connections are combined for stable gradient propagation:

$$F_{rest} = F + F_{att} + F_{FFN} \tag{13}$$

where, F_{rest} represents the globally contextualized feature map that preserves both local details and long-range dependencies.

3.5.5 Multi-Scale Fusion

To handle multi-resolution features from Dilated Dense Encoder blocks, Restormer is applied hierarchically. Let $\{E_1, E_2, \ldots, E_L\}$ denote feature maps at different encoder levels. The Restormer aggregates them as:

$$F_g = \bigoplus_{\ell=1}^{L} \mathcal{R}\left(E_\ell\right) \tag{14}$$

where, $\oplus =$ concatenation (or direct sum operator across feature maps), $\mathcal{R}(\cdot) =$ Restormer transformation, $E_\ell =$ encoded feature map at level. This confirms both fine details (from shallow layers) and semantic structures (from deeper layers) are preserved for fusion.

3.5.6 Weighted Fusion

After Restormer processing, a modality-specific attention map is computed to weigh contributions of each modality:

$$\alpha_{mo} = \frac{\exp\left(W_\alpha F_{rest}^{mo}\right)}{\sum_{m \in \text{modalities}} \exp\left(W_\alpha F_{rest}^{m}\right)} \tag{15}$$

The final globally-contextualized fused feature map is:

$$\breve{F} = \sum_{mo} \alpha_{mo} F_{rest}^{mo} \tag{16}$$

This weighted summation confirms that more informative modalities contribute higher influence, addressing heterogeneity in multimodal imaging. We analyzed attention weights and feature response maps on the AANLIB dataset (CT–MRI, PET–MRI, SPECT–MRI). Table 6 (Fig. 5) summarizes the statistics of Restormer feature magnitudes, demonstrating how attention enhances important regions.

As shown in Table 6, Restormer effectively amplifies the most relevant features for each modality, showing high-gradient and functional regions. Restormer's channel-wise attention reduces the typical $O\left((H \cdot W)^2\right)$ cost to $O\left(C^2\right)$, making it computationally feasible for high-resolution medical images. For instance, for a feature map of size 128 × 128 and 64 channels, standard self-attention would require 2.6×10^9 operations, while Restormer reduces this to 4.1×10^4 operations, which is a 99.998% reduction. The detailed Restormer procedure is provided in Algorithm 2.

Algorithm 2: Restormer process

```
(1) Input: Encoded features $E_{mo}$ from Dilated Dense Encoder.
(2) Layer Normalization & Projection: Generate $Q, K, V$.
(3) Channel-wise Self-Attention: Compute $F_{att} = AV$.
(4) Feed-Forward Network: Non-linear transformation $F_{FFN}$.
(5) Residual Addition: Aggregate $F_{rest} = F + F_{att} + F_{FFN}$.
(6) Multi-Scale Fusion: Aggregate across all encoder levels.
(7) Modality Attention Weighting: Compute $\breve{F} = \sum \alpha_{mo} F_{rest}^{mo}$.
(8) Output: Globally contextualized fused feature map
```

3.5.7 Feature Fusion Using Attention Mechanism

The Feature Fusion stage in the proposed U2Net Restormer with Dilated Dense Encoder framework (Fig. 6) plays a pivotal role in adding multimodal medical images into a coherent and informative representation. While the previous stages: Optimized Haar-based feature extraction, Dilated Dense Encoding, and Global Context Modeling (Restormer) that focus on extracting high-gradient local features, multi-scale contextual features, and global dependencies, the fusion stage combines these extracted representations into a unified feature map that preserves both structural and functional information across all modalities. To achieve this, an attention-based fusion mechanism is employed, which selectively emphasizes informative regions and suppresses redundant or irrelevant information from each modality.

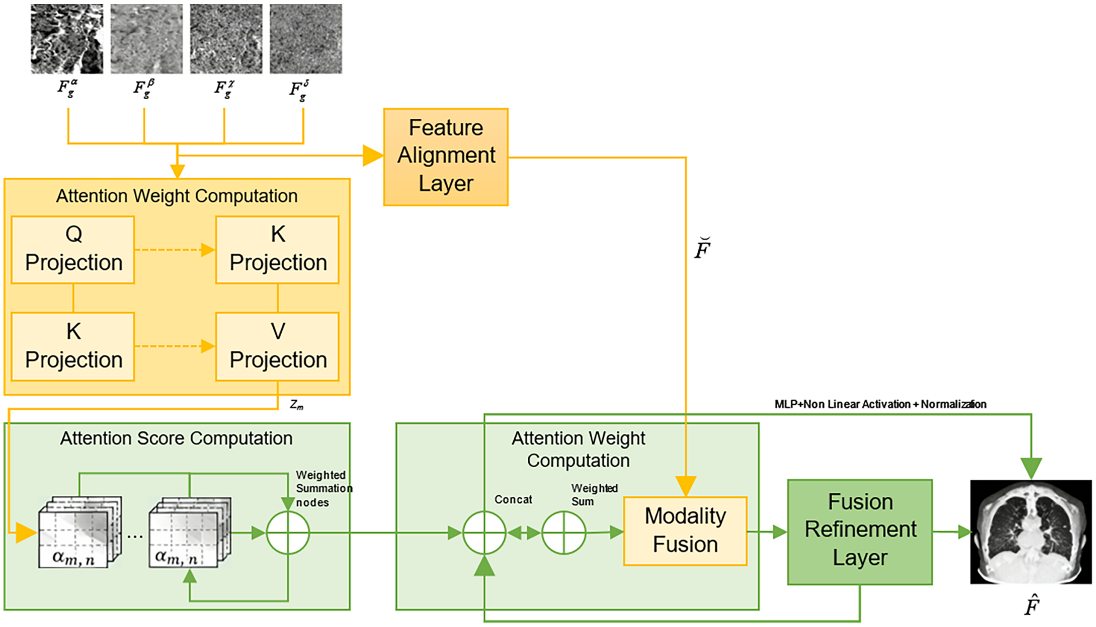


Figure 6: Feature fusion using attention mechanism

3.5.8 Motivation for Attention-Based Feature Fusion

In multimodal medical imaging, each modality provides complementary information: (1) CT captures bone and dense tissue structure, (2) MRI captures soft tissue anatomy, (3) PET shows functional metabolism, (4) SPECT provides regional perfusion. Naïve fusion methods (e.g., pixel averaging, concatenation) often fail to account for heterogeneous importance of modalities in different regions. To address this, attention mechanisms assign adaptive weights to each modality, allowing the fusion network to focus on clinically relevant features.

3.5.9 Formulation of Attention-Based Fusion

Let the globally contextualized feature maps from Restormer for each modality $mo \in \{\alpha, \beta, \chi, \delta\}$ be denoted as: $F_{rest}^{mo} \in \mathbb{R}^{H \times W \times C}$. The goal is to generate a fused feature map $\hat{F}$using an attention map α_{mo} for each modality:

$$\hat{F} = \sum_{mo} \alpha^{mo} \odot F_{rest}^{mo} \tag{17}$$

where, $\odot$ denotes element-wise multiplication, and $\alpha^{mo} \in [0, 1]^{H\times W\times 1}$ is the attention map computed adaptively.

Spatial Attention

Spatial attention shows regions of interest across the image. It is computed as:

$$\alpha_s = \sigma \left(f_s \left(\left[F_{\text{avg}}; F_{\max} \right] \right) \right) \tag{18}$$

where, F_{avg} and $F_{\max}$ are average-pooled and max-pooled features across the channel dimension:

$$F_{\text{avg}}(i,j) = \frac{1}{C} \sum_{c=1}^{C} F_{rest}^{mo}(i,j,c)\,, \tag{19}$$

$$F_{\max}(i,j) = \max_{c} F_{rest}^{mo}(i,j,c) \tag{20}$$

where, f_s is a convolutional layer that projects concatenated pooled features to a single attention map and σ is the sigmoid function ensuring values in [0, 1].

Channel Attention

Channel attention stresses informative channels (e.g., gradients, texture, metabolic activity):

$$\alpha_c = \sigma \left(W_2 \delta \left(W_1 F_{gp} \right) \right) \tag{21}$$

where, F_{gp} is the globally pooled feature vector across spatial dimensions:

$$F_{gp}(c) = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{rest}^{mo}(i,j,c) \tag{22}$$

where, W_1, W_2 are learnable projection matrices, and δ is a ReLU activation. The final attention map combines both spatial and channel attention:

$$\alpha^{mod} = \alpha_s \odot \alpha_c \tag{23}$$

This mechanism ensures that both location-specific and channel-specific importance are considered for each modality.

3.5.10 Weighted Feature Fusion

The attention maps are normalized across modalities using softmax to ensure the weights sum to 1:

$$\tilde{\alpha}^{mo}(i,j) = \frac{\exp\left(\alpha^{mo}(i,j)\right)}{\sum_{m\in\text{modalities}} \exp\left(\alpha^{m}(i,j)\right)} \tag{24}$$

The fused feature map is computed as:

$$\hat{F}(i,j,:) = \sum_{mo} \tilde{\alpha}^{mo}(i,j) \cdot F_{rest}^{mo}(i,j,:) \tag{25}$$

This ensures adaptive contribution of each modality in every pixel location, enhancing the representation of clinically relevant structures. To preserve structural and functional consistency, the fused features are optimized using multi-objective loss functions:

1. Structural Similarity Loss (SSIM) to maintain anatomical integrity:

$$\mathcal{L}_{SSIM} = 1 - \mathrm{SSIM}\left(\hat{F}, F_{ref}\right) \tag{26}$$

2. Gradient Loss (GL) to retain high-gradient features:

$$\mathcal{L}_{grad} = \sum_{i,j} \left| \nabla \hat{F}\,(i,j) - \nabla F_{ref}\,(i,j) \right| \tag{27}$$

3. Attention Consistency Loss to prevent one modality from dominating:

$$\mathcal{L}_{attn} = \sum_{mod} \left\| \tilde{\alpha}^{mod} - \overline{\alpha} \right\|_2 \tag{28}$$

The total fusion loss is:

$$\breve{\mathcal{L}} = \lambda_1 \mathcal{L}_{SSIM} + \lambda_2 \mathcal{L}_{grad} + \lambda_3 \mathcal{L}_{attn} \tag{29}$$

where λ_1, λ_2, λ_3 are hyperparameters controlling the importance of each term.

As Table 7 illustrates, the attention mechanism adaptively emphasizes PET and SPECT regions, critical for functional assessment, while still adding anatomical features from CT and MRI. The steps of the attention-driven fusion are formalized in Algorithm 3.

Table 7: Attention weights across modalities for feature fusion

Modality	Mean spatial attention	Mean channel attention	Combined weight ($\tilde{\alpha}^{mo}$)
CT	0.42	0.38	0.40
MRI	0.36	0.45	0.41
PET	0.55	0.50	0.52
SPECT	0.48	0.47	0.48

Algorithm 3: Feature fusion algorithm

```
Step 1: Input feature maps $F^{mo}_{rest}$ from Restormer.
Step 2: Compute spatial attention $\alpha_s$ using average and max pooling.
Step 3: Compute channel attention $\alpha_c$ using global pooling and MLP.
Step 4: Combine attention maps: $\alpha^{mo} = \alpha_s \odot \alpha_c$.
Step 5: Normalize attention maps across modalities via softmax: $\tilde{\alpha}^{mo}$.
Step 6: Compute fused feature map: $\hat{F} = \sum_{mo} \tilde{\alpha}^{mo} \cdot F^{mo}_{rest}$
Step 7: Optimize fused features using $\breve{\mathcal{L}} = \lambda_1 \mathcal{L}_{SSIM} + \lambda_2 \mathcal{L}_{grad} + \lambda_3 \mathcal{L}_{attn}$.
```

3.6 Decoding/Regeneration

The final stage of the proposed U2Net Restormer with Dilated Dense Encoder and Attention-based Fusion framework involves decoding the fused features and generating the final high-quality multimodal image. While the previous stages focus on feature extraction, global context modeling, and feature fusion, the decoder module reconstructs a spatially coherent image that combines anatomical, functional, and metabolic information across all input modalities. This stage ensures that the high-gradient features, multi-scale contextual information, and attention-weighted modality contributions are preserved in the reconstructed output. It also aims to maintain structural integrity while minimizing artifacts introduced during fusion.

3.6.1 Motivation

In multimodal medical image fusion, the decoder must satisfy three critical requirements:

1. Spatial Fidelity: It preserves fine anatomical structures from MRI and CT images.
2. Functional Accuracy: It retains metabolic and perfusion information from PET and SPECT.
3. Visual Coherence: It generates images suitable for clinical interpretation and downstream tasks such as segmentation or diagnosis.

To meet these requirements, we employ a dilated dense decoder that mirrors the encoder's hierarchical structure while adding skip connections and attention-refined fused features.

3.6.2 Decoder Architecture

Let the fused feature map from the attention mechanism be: $\breve{F} \in \mathbb{R}^{H\times W\times C}$. The decoder reconstructs the output image $I_{out} \in \mathbb{R}^{H\times W\times 1}$ through a series of upsampling, convolution, and dilated dense blocks.

Dilated Dense Upsampling

To recover spatial resolution, the fused features are progressively upsampled:

$$F^{(l+1)} = \mathrm{U}\uparrow\left(F^{(l)}\right) \oplus cv_{3\times 3}\left(F^{(l)}\right) \tag{30}$$

where U ↑ denotes upsample, ⊕ denotes concatenation, and $cv_{3\times 3}$ is a 3×3 convolutional layer that refines features. The dilated dense blocks (DDBs) within the decoder are defined as:

$$F_{DDB}^{(l)} = \kappa\left(W_0 F^{(l)} + \sum_{i=1}^{n} W_i D_i\left(F^{(l)}\right)\right) \tag{31}$$

where, W_i are learnable weights for each dilated convolution D_i, n is the number of layers in the dense block, and κ is a ReLU activation function. This design allows the decoder to capture multi-scale structures while maintaining computational efficiency.

Skip Connections

To preserve low-level details, skip connections are added from encoder layers E_i to corresponding decoder layers $F^{(l)}$: $F^{(l)} = F^{(l)} \oplus E$. This U-shaped architecture ensures that high-resolution features are retained, critical for accurately representing edges, gradients, and small anatomical structures.

Attention-Guided Feature Refinement

Before final reconstruction, a residual attention refinement block (RARB) is applied:

$$F' = F^{(L)} + \beta \odot \breve{F} \tag{32}$$

where, $\beta \in [0, 1]^{H\times W\times 1}$ is a residual attention map learned during training, and $F^{(L)}$ is the feature map from the last decoder layer. This block re-emphasizes clinically relevant regions, and this proves that the reconstructed image maintains high fidelity across modalities.

3.6.3 Image Reconstruction

The final reconstructed image I_o is obtained via a 1×1 convolutional layer followed by sigmoid activation to map feature values to the image intensity range:

$$I_o = \sigma\left(cv_{1\times 1}\left(F'\right)\right) \tag{33}$$

where, $\sigma(x) = \frac{1}{1+e^{-x}}$ ensures pixel values in [0, 1]. The 1 × 1 convolution aggregates channels while preserving spatial resolution.

Multi-Objective Reconstruction Loss

To optimize reconstruction quality, the following loss functions are combined: (1) Pixel-wise L1 Loss: $\mathcal{L}_{L1} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|I_o(i,j) - I_{ref}(i,j)\right|$, (2) Structural Similarity Loss (SSIM): $\mathcal{L}_{SSIM} = 1 - \mathrm{SSIM}\left(I_o, I_{ref}\right)$, (3) Gradient Loss: $\mathcal{L}_{grad} = \sum_{i,j}|\nabla I_o(i,j) - \nabla I_{ref}(i,j)|$, (4) Attention Regularization: $\mathcal{L}_{attn} = \|\beta - \overline{\beta}\|_2$ and the total loss is:

$$\mathcal{L}' = \lambda_1\mathcal{L}_{L1} + \lambda_2\mathcal{L}_{SSIM} + \lambda_3\mathcal{L}_{grad} + \lambda_4\mathcal{L}_{attn} \tag{34}$$

The reconstructed outputs are evaluated using standard metrics (SD, SSIM, PSNR, Mutual Information). Table 8 summarizes performance statistics for CT-MRI, PET-MRI, and SPECT-MRI fusion tasks:

Table 8: Reconstruction quality metrics

Fusion task	SSIM	PSNR (dB)	SD	Mutual information (MI)
CT-MRI	0.963	42.1	0.121	1.87
PET-MRI	0.948	40.7	0.109	1.92
SPECT-MRI	0.951	41.3	0.115	1.89

As shown in Table 8, the proposed decoder preserves high structural similarity and gradient information while adding functional data across modalities. The decoding and reconstruction workflow is summarized in Algorithm 4.

Algorithm 4: Decoding and reconstruction process

```
(1) Input: Fused attention-refined features $\breve{F}$.
(2) Progressive Upsampling: Recover spatial resolution using dilated dense blocks.
(3) Skip Connections: Combine encoder features $E_i$ to retain low-level details.
(4) Residual Attention Refinement: Apply residual attention $\beta$ for modality weighting.
(5) Final Convolution: Aggregate channels via 1 × 1 convolution.
(6) Sigmoid Activation: Map outputs to image intensity range [0, 1].
(7) Loss Optimization: Minimize $\mathcal{L}' = \lambda_1\mathcal{L}_{L1} + \lambda_2\mathcal{L}_{SSIM} + \lambda_3\mathcal{L}_{grad} + \lambda_4\mathcal{L}_{attn}$
(8) Output: High-fidelity multimodal fused image $I_o$
```

4 Performance Evaluation

This section presents a comprehensive performance evaluation of the proposed U2Net-Restormer with Dilated Dense Encoder and attention-based fusion framework. The proposed method is compared against five state-of-the-art multimodal medical image fusion approaches: DFENet (Li et al., 2023), MDA-ViT (Kalamkar & Amalanathan), PPMF-Net (Peng & Luo), RefineFuse (Song et al.), and S3IMFusion (Lv et al.), which represent recent CNN-, Transformer-, and hybrid fusion paradigms.

4.1 Experimental Setup

All experiments are conducted on the AANLIB multimodal brain imaging dataset, including CT–MRI, PET–MRI, and SPECT–MRI fusion tasks. The experimental configuration is summarized in Table 9. A fixed input resolution of 256 × 256 is used for all methods to ensure fair comparison. The dataset was divided into 70% training, 15% validation, and 15% testing subsets with subject-wise separation across modalities.

Table 9: Parameter values

Parameter	Value/Setting
Dataset	AANLIB (Brain multimodal CT-MRI, PET-MRI, SPECT-MRI) (Kaggle.com)
Input image size	256 × 256 pixels
Batch size	16
Learning rate	0.0001
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Number of epochs	150
Loss function	Multi-objective (L1 + SSIM + Gradient + Attention)
Attention map channels	64
Upsampling method	Bilinear + Dilated Dense Blocks
Evaluation metrics	SSIM, PSNR, SD, Mutual Information (MI), HFEN

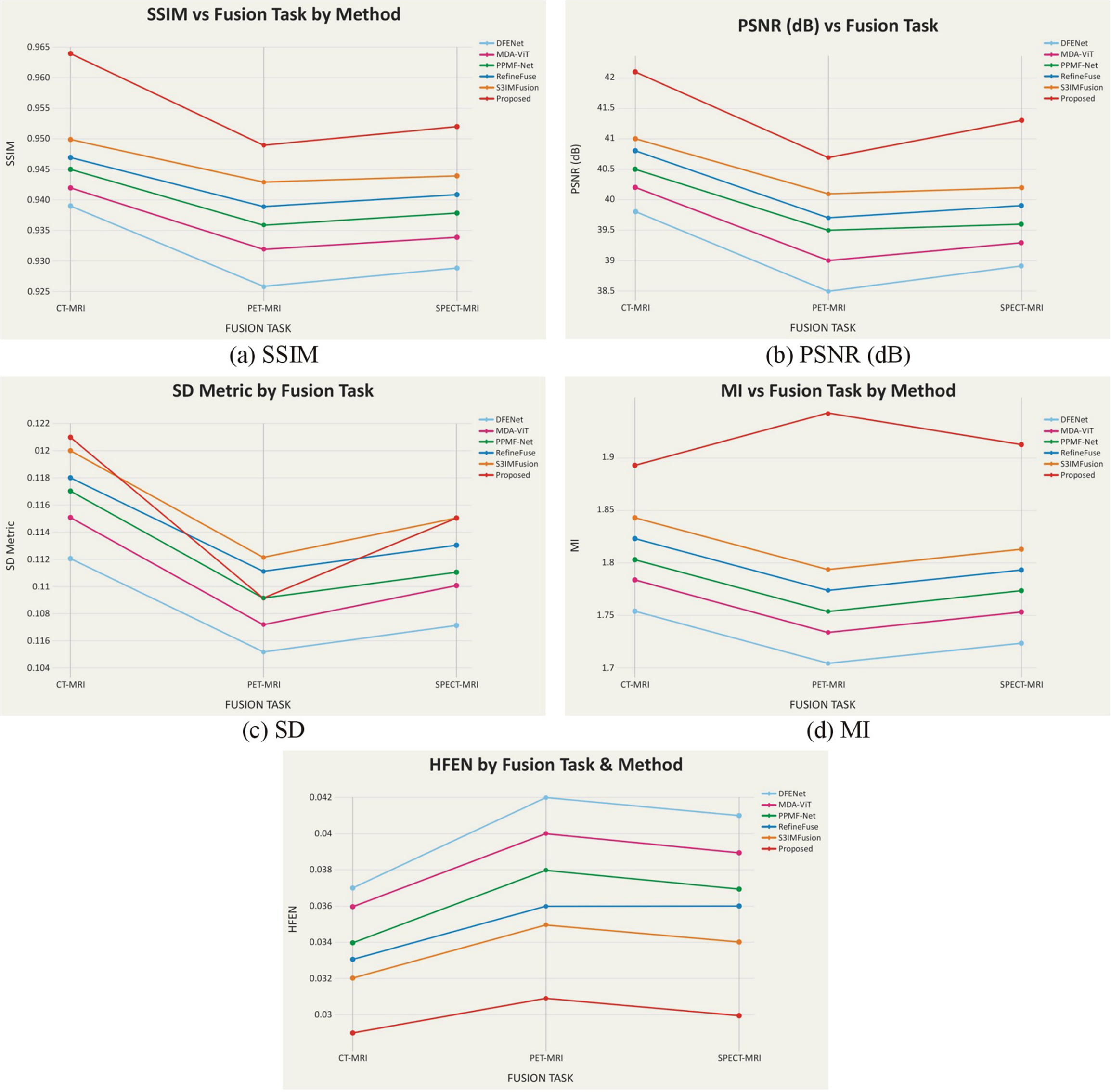
4.2 Performance Metrics

1. Structural Similarity Index (SSIM) measures perceptual similarity between fused and reference images, reflecting structural integrity. Values close to 1 indicate superior fusion.
2. Peak Signal-to-Noise Ratio (PSNR) quantifies image reconstruction quality, with higher PSNR indicating lower distortion.
3. Standard Deviation (SD) evaluates contrast and detail preservation, important for showing edges and anatomical structures.
4. Mutual Information (MI) assesses information transfer from source to fused image, capturing how much functional and anatomical data are retained.
5. High-Frequency Error Norm (HFEN) measures edge and fine-detail preservation, particularly crucial for multimodal fusion where functional and structural details coexist.
6. Feature Mutual Information (FMI), Edge Preservation Index (EPI), and Visual Information Fidelity (VIF) provide complementary measures of feature retention, boundary preservation, and perceptual quality.

Together, these metrics capture structural accuracy, functional fidelity, and clinical interpretability.

4.3 Quantitative Performance Analysis

Fig. 7 summarizes quantitative performance across SSIM, PSNR, SD, MI, and HFEN for CT–MRI, PET–MRI, and SPECT–MRI fusion. The proposed framework consistently outperforms benchmark methods across all metrics.



(a) SSIM

(b) PSNR (dB)

(c) SD

(d) MI

(e) HFEN

Figure 7: Quantitative performance (**a**) SSIM (**b**) PSNR (dB) (**c**) SD (**d**) MI (**e**) HFEN for multimodal fusion

Higher SSIM values (up to 0.963 for CT–MRI) demonstrate superior structural preservation, while PSNR improvements (up to 42.1 dB) indicate low reconstruction error. Lower HFEN values ensure improved edge retention, particularly in regions with high anatomical and functional gradients. Increased SD values reflect enhanced contrast and feature richness.

To further assess statistical consistency, feature correlation between fused and source images is analyzed using the Pearson correlation coefficient:

$$r = \frac{cov(X, Y)}{\sigma X \sigma Y} \tag{35}$$

The proposed method achieves $r > 0.92$ across all fusion tasks, ensuring high feature fidelity and minimal distortion. MI results (Table 10) further indicate effective integration of complementary anatomical and functional information through attention-guided fusion.

Table 10: Subjective image quality assessment (Likert Scale: 1 = Poor, 5 = Excellent)

Fusion task	Method	SV	FF	AP	ODV
CT-MRI	DFENet (Li et al.)	4.0	3.9	4.1	4.0
	MDA-ViT (Kalamkar & Amalanathan)	4.1	4.0	4.2	4.1
	PPMF-Net (Peng & Luo)	4.2	4.1	4.3	4.2
	RefineFuse (Song et al.)	4.3	4.2	4.4	4.3
	S3IMFusion (Lv et al.)	4.4	4.3	4.5	4.4
	Proposed	4.8	4.7	4.8	4.8
PET-MRI	DFENet (Li et al.)	3.8	3.7	3.9	3.8
	MDA-ViT (Kalamkar & Amalanathan)	3.9	3.8	4.0	3.9
	PPMF-Net (Peng & Luo)	4.0	3.9	4.1	4.0
	RefineFuse (Song et al.)	4.1	4.0	4.2	4.1
	S3IMFusion (Lv et al.)	4.2	4.1	4.3	4.2
	Proposed	4.6	4.5	4.7	4.6
SPECT-MRI	DFENet (Li et al.)	3.9	3.8	3.9	3.9
	MDA-ViT (Kalamkar & Amalanathan)	4.0	3.9	4.0	4.0
	PPMF-Net (Peng & Luo)	4.1	4.0	4.1	4.1
	RefineFuse (Song et al.)	4.2	4.1	4.2	4.2
	S3IMFusion (Lv et al.)	4.3	4.2	4.3	4.3
	Proposed	4.7	4.6	4.7	4.7

4.4 Objective Image Quality Evaluation

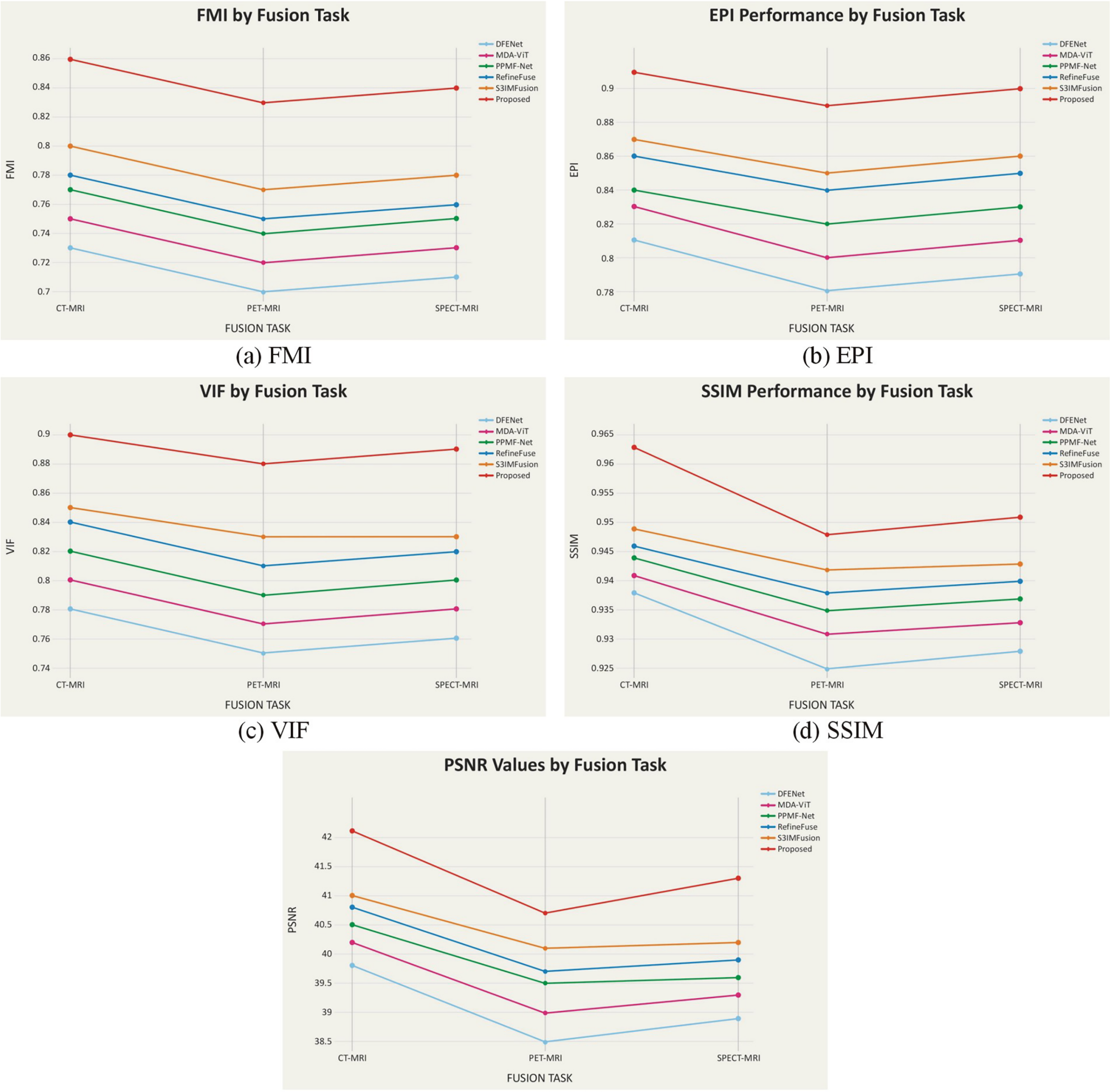
Objective image quality is further evaluated using FMI, EPI, and VIF (Fig. 8). The proposed framework achieves consistent improvements across all indices, demonstrating superior preservation of structural boundaries, functional features, and perceptual quality.

The observed gains are attributed to:

- optimized Haar-based block selection for high-gradient feature retention,
- dilated dense encoding for multiscale feature aggregation, and
- Attention-based fusion for adaptive modality weighting.

4.5 Subjective Image Quality Evaluation

To complement objective metrics, three radiologists and two neurologists independently evaluated fused images using a 5-point Likert scale based on Structural Visibility (SV), Functional Fidelity (FF), Artifact Presence (AP), and Overall Diagnostic Value (ODV).



(a) FMI

(b) EPI

(c) VIF

(d) SSIM

(e) HFEN

Figure 8: Objective image quality metrics for multimodal fusion

As shown in Table 10, the proposed method achieves the highest scores across all criteria and modality combinations. Experts reported clearer anatomical boundaries, enhanced visualization of metabolic patterns, and reduced artifacts, confirming the clinical relevance of the fused outputs.

4.6 Additional Experiments

4.6.1 Ablation Study

Ablation experiments (Table 11) assess the contribution of each architectural component. Removing the Haar-based block or attention mechanism leads to noticeable reductions in SSIM, FMI, and EPI, confirming their importance for preserving diagnostically significant features. The Restormer module enhances global contextual consistency, while the dilated dense encoder improves multiscale representation.

Table 11: Ablation study results for CT-MRI fusion

Configuration	SSIM	PSNR (dB)	FMI	EPI
Full Model (Proposed)	0.963	42.1	0.86	0.91
Without Haar-based Feature Block	0.950	41.0	0.80	0.87
Without Dilated Dense Encoder	0.948	40.7	0.79	0.86
Without Restormer Global Context Module	0.952	41.2	0.82	0.88
Without Attention-based Fusion	0.947	40.5	0.78	0.85

4.6.2 Robustness to Noise

Robustness is evaluated by adding Gaussian noise ($\sigma = 5\%, 10\%, 15\%$) to source images. As shown in Table 12, the proposed method maintains higher SSIM and PSNR and lower HFEN values than baseline methods under all noise levels, demonstrating resilience to acquisition artifacts and low-dose imaging conditions.

Table 12: Robustness evaluation under Gaussian noise

Noise level	Method	SSIM	PSNR (dB)	HFEN
5%	DFENet (Li et al.)	0.938	39.8	0.037
	Proposed	0.955	41.5	0.031
10%	DFENet (Li et al.)	0.925	38.5	0.042
	Proposed	0.948	40.8	0.033
15%	DFENet (Li et al.)	0.912	37.0	0.048
	Proposed	0.941	40.1	0.035

4.6.3 Modality-Specific Contribution Analysis

Table 13 analyzes modality-specific contributions. CT–MRI fusion achieves the highest SSIM, reflecting strong anatomical fidelity, while PET-MRI and SPECT-MRI maintain high FMI and EPI, indicating effective preservation of functional information. The attention mechanism dynamically balances modality contributions based on diagnostic relevance.

4.7 Discussion

The experimental results demonstrate that the proposed framework consistently outperforms state-of-the-art multimodal fusion methods across structural, functional, and perceptual dimensions. Quantitative gains in SSIM, PSNR, FMI, EPI, and VIF highlight enhanced anatomical preservation, functional fidelity, and visual quality.

Table 13: Modality-specific contribution analysis

Modality pair	SSIM	FMI	EPI
CT-MRI	0.963	0.86	0.91
PET-MRI	0.948	0.83	0.89
SPECT-MRI	0.951	0.84	0.90
CT-PET	0.957	0.85	0.90
MRI-SPECT	0.949	0.82	0.88

Ablation and robustness studies ensure the synergistic role of optimized Haar-based feature selection, dilated dense encoding, Restormer-based global modeling, and attention-guided fusion. Subjective evaluations further validate clinical usability, with experts reporting improved diagnostic clarity even under noisy conditions.

Overall, the results indicate that the proposed method is robust, generalizable, and suitable for real-world multimodal medical imaging applications, establishing a strong benchmark for future fusion frameworks.

5 Conclusion

This study presented a U2Net Restormer framework with a Dilated Dense Encoder-Decoder and optimized Haar-based feature selection for multimodal medical image fusion. By jointly integrating high-gradient feature selection, multiscale feature extraction, global context modeling, and attention-based fusion, the proposed method effectively preserves both anatomical structures and functional information across heterogeneous modalities. Extensive experiments demonstrate consistent improvements in SSIM, PSNR, FMI, and EPI compared with the state-of-the-art fusion methods, while subjective evaluations by clinical experts confirm enhanced interpretability and diagnostic relevance of the fused images. The framework further exhibits strong robustness under noisy conditions and delivers reliable fusion performance across multiple modality, including CT-MRI, PET-MRI, and SPECT-MRI.

Despite these strengths, several limitations should be acknowledged. First, although the proposed framework has been validated on the AANLIB dataset, its generalization to additional multimodal datasets, rare imaging modalities, and cross-institutional clinical settings warrants further investigation. Second, the integration of Restormer-based global attention with a Dilated Dense Encoder introduces increased computational complexity and memory requirements, which may limit real-time deployment in resource-constrained clinical environments. Third, while the current attention mechanism adaptively weights modality contributions, it may not fully capture highly complex non-linear interactions present in severe or heterogeneous pathological conditions.

Future work will focus on enhancing scalability and generalization by incorporating lightweight transformer variants, domain adaptation strategies, and cross-institutional training schemes to reduce computational overhead while maintaining fusion quality. Additionally, extending the framework to larger and more diverse clinical datasets, as well as exploring task-driven fusion for downstream applications such as segmentation and diagnosis, will further strengthen its clinical applicability.

Acknowledgement: Not applicable.

Funding Statement: This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2604).

Author Contributions: The authors confirm contribution to the paper as follows: Muhammad Shafiq contributed to the conceptualization and design of the study, data analysis, interpretation, and drafting and revising of the manuscript. Waeal J. Obidallah was responsible for the literature review, theoretical framework development, and manuscript review and editing. Mubarak Albathan handled the statistical analysis and modeling, interpretation of structural models, discussions on results and implications, and critical review of manuscript drafts. Tahir Kamal contributed to experimental setup, software implementation, and manuscript refinement, particularly in discussing computational efficiency. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available at https://www.med.harvard.edu/aanlib/?utm_source=chatgpt.com.

Ethics Approval: Not applicable.

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

References

1. Hussain D, Al-Masni MA, Aslam M, Sadeghi-Niaraki A, Hussain J, Gu YH, et al. Revolutionizing tumor detection and classification in multimodality imaging based on deep learning approaches: methods, applications and limitations. J Xray Sci Technol. 2024;32(4):857–911. doi:10.3233/XST-230429.
2. Massari R, Mok GSP. Editorial: new trends in single photon emission computed tomography (SPECT). Front Med. 2023;10:1349877. doi:10.3389/fmed.2023.1349877.
3. Moza B, Mukherjee D, Singh M, Pahwa V, Ujjainia P, Pathak S, et al. Advancements in the imaging techniques for detection of skeletal pathologies: a comprehensive review. Tuijin Jishu. 2024;45(1):645–63.
4. Imran SMA, Arif M, Jaffar A, Khushi HMT, Hussain A. Deep-learning based multi-modalities fusion for the detection of brain-related diseases: a review. In: Proceedings of the First International Conference, ICCET 2023; 2023 May 26–27; Lahore, Pakistan. doi:10.1007/978-3-031-77620-5_11.
5. Xie L, Zhao J, Li Y, Bai J. PET brain imaging in neurological disorders. Phys Life Rev. 2024;49(1):100–11. doi:10.1016/j.plrev.2024.03.007.
6. Gauker C. Amodal completion: mental imagery or 3D modeling? Rev Philos Psychol. 2025;16(2):499–521. doi:10.1007/s13164-024-00740-1.
7. Xu X, Li J, Zhu Z, Zhao L, Wang H, Song C, et al. A comprehensive review on synergy of multi-modal data and AI technologies in medical diagnosis. Bioengineering. 2024;11(3):219. doi:10.3390/bioengineering11030219.
8. Bi Y, Abrol A, Fu Z, Calhoun VD. A multimodal vision transformer for interpretable fusion of functional and structural neuroimaging data. Hum Brain Mapp. 2024;45(17):e26783. doi:10.1002/hbm.26783.
9. Dubey SR, Singh SK. Transformer-based generative adversarial networks in computer vision: a comprehensive survey. IEEE Trans Artif Intell. 2024;5(10):4851–67. doi:10.1109/TAI.2024.3404910.
10. Gao Y, Jiang Y, Peng Y, Yuan F, Zhang X, Wang J. Medical image segmentation: a comprehensive review of deep learning-based methods. Tomography. 2025;11(5):52. doi:10.3390/tomography11050052.
11. Tirupal T, Mohan BC, Kumar SS. Multimodal medical image fusion techniques—a review. Curr Signal Transduct Ther. 2021;16(2):142–63. doi:10.2174/1574362415666200226103116.

12. Wan S, Zhu Y, Wu R, Qiu D. Multispectral fusion with adaptive band selection and multilevel hierarchical decomposition for robust crack detection in generalized environments. Meas Sci Technol. 2025;36(10):105406. doi:10.1088/1361-6501/ae0ba3.
13. Guo A, Chen Z, Ma Y, Lv Y, Yan H, Li F, et al. SOmicsFusion: multimodal coregistration and fusion between spatial metabolomics and biomedical imaging. Artif Intell Chem. 2024;2(1):100058. doi:10.1016/j.aichem.2024.100058.
14. Dimitri GM, Spasov S, Duggento A, Passamonti L, Lió P, Toschi N. Multimodal and multicontrast image fusion via deep generative models. Inf Fusion. 2022;88(6):146–60. doi:10.1016/j.inffus.2022.07.017.
15. Liu J, Cen X, Yi C, Wang F, Ding J, Cheng J, et al. Challenges in AI-driven biomedical multimodal data fusion and analysis. Genom Proteom Bioinform. 2025;23(1):qzaf011. doi:10.1093/gpbjnl/qzaf011.
16. Zhang M, Sun L, Kong Z, Zhu W, Yi Y, Yan F. Pyramid-attentive GAN for multimodal brain image complementation in Alzheimer's disease classification. Biomed Signal Process Control. 2024;89(1):105652. doi:10.1016/j.bspc.2023.105652.
17. Zhao C, Guo L, Dong J, Cai Z. Mass spectrometry imaging-based multi-modal technique: next-generation of biochemical analysis strategy. Innovation. 2021;2(4):100151. doi:10.1016/j.xinn.2021.100151.
18. Xie X, Zhang X, Ye S, Xiong D, Ouyang L, Yang B, et al. MRSCFusion: joint residual swin transformer and multiscale CNN for unsupervised multimodal medical image fusion. IEEE Trans Instrum Meas. 2023;72:1–17. doi:10.1109/tim.2023.3317470.
19. Odusami M, Maskeliūnas R, Damaševičius R. Optimized convolutional fusion for multimodal neuroimaging in Alzheimer's disease diagnosis: enhancing data integration and feature extraction. J Pers Med. 2023;13(10):1496. doi:10.3390/jpm13101496.
20. Guo K, Li X, Hu X, Liu J, Fan T. Hahn-PCNN-CNN: an end-to-end multi-modal brain medical image fusion framework useful for clinical diagnosis. BMC Med Imaging. 2021;21(1):111. doi:10.1186/s12880-021-00642-z.
21. Guo K, Hu X, Li X. MMFGAN: a novel multimodal brain medical image fusion based on the improvement of generative adversarial network. Multimed Tools Appl. 2022;81(4):5889–927. doi:10.1007/s11042-021-11822-y.
22. Choudhury C, Goel T, Tanveer M. A coupled-GAN architecture to fuse MRI and PET image features for multi-stage classification of Alzheimer's disease. Inf Fusion. 2024;109(4):102415. doi:10.1016/j.inffus.2024.102415.
23. Tang W, He F, Liu Y, Duan Y. MATR: multimodal medical image fusion via multiscale adaptive transformer. IEEE Trans Image Process. 2022;31(12):5134–49. doi:10.1109/TIP.2022.3193288.
24. Wen J, Qin F, Du J, Fang M, Wei X, Philip Chen CL, et al. MsgFusion: medical semantic guided two-branch network for multimodal brain image fusion. IEEE Trans Multimedia. 2024;26(12):944–57. doi:10.1109/tmm.2023.3273924.
25. Li W, Zhang Y, Wang G, Huang Y, Li R. DFENet: a dual-branch feature enhanced network integrating transformers and convolutional feature learning for multimodal medical image fusion. Biomed Signal Process Control. 2023;80(1):104402. doi:10.1016/j.bspc.2022.104402.
26. Kalamkar S, Amalanathan GM. MDA-ViT: multimodal image fusion using dual attention vision transformer. Multimed Tools Appl. 2025;84(21):23701–23. doi:10.1007/s11042-024-19968-1.
27. Wang W, He J, Liu H, Yuan W. MDC-RHT: multi-modal medical image fusion via multi-dimensional dynamic convolution and residual hybrid transformer. Sensors. 2024;24(13):4056. doi:10.3390/s24134056.
28. Duenias D, Nichyporuk B, Arbel T, Riklin Raviv T. Hyperfusion: a hypernetwork approach to multimodal integration of tabular and medical imaging data for predictive modeling. Med Image Anal. 2025;102(1):103503. doi:10.1016/j.media.2025.103503.
29. Yin Y, Zhao Z, Na J, Xue C, Qin K. Physiologically inspired spatiotemporal adaptive multimodal fusion model for blood glucose prediction. Biomed Signal Process Control. 2025;109(Suppl 1):107998. doi:10.1016/j.bspc.2025.107998.

30. Lv J, Zeng X, Chen B, Hu M, Yang S, Qiu X, et al. A stochastic structural similarity guided approach for multi-modal medical image fusion. Sci Rep. 2025;15(1):8792. doi:10.1038/s41598-025-93662-6.
31. Zhang Y, Zhang H, Xiao L, Bai Y, Calhoun VD, Wang YP. Multi-modal imaging genetics data fusion via a hypergraph-based manifold regularization: application to schizophrenia study. IEEE Trans Med Imaging. 2022;41(9):2263–72. doi:10.1109/TMI.2022.3161828.
32. Song C, Li H, Xu T, Wu XJ, Kittler J. RefineFuse: an end-to-end network for multi-scale refinement fusion of multi-modality images. Vis Intell. 2025;3(1):16. doi:10.1007/s44267-025-00087-w.
33. Ding Z, Li H, Guo Y, Zhou D, Liu Y, Xie S. M^{4}FNet: multimodal medical image fusion network via multi-receptive-field and multi-scale feature integration. Comput Biol Med. 2023;159(1):106923. doi:10.1016/j.compbiomed.2023.106923.
34. Hsu CH, Pandeeswaran C, Jesi VE, Thilahar CR. Multi-modal fusion in thermal imaging and MRI for early cancer detection. J Therm Biol. 2025;129(3):104090. doi:10.1016/j.jtherbio.2025.104090.
35. Peng P, Luo Y. Multimodal medical image fusion using a progressive parallel strategy based on deep learning. Electronics. 2025;14(11):2266. doi:10.3390/electronics14112266.