Received: 20 November 2025 | Revised: 13 July 2026 | Accepted: 28 August 2026 | Published online: 18 September 2026

REVIEW

Journal of Data Science and Intelligent Systems
2026, Vol. 00(00) 1–19
DOI: 10.47852/bonviewJDSIS62028331

BON VIEW PUBLISHING

Transformer-Based Multimodal Image Fusion in Healthcare: A Comprehensive Survey of Architectures, Fusion Strategies, and Clinical Applications

Nalini S Jagtap[1,*], Dilip Kumar Jang Bahadur Saini[2] and Prasun Chakrabarti[3]

[1] *Pimpri Chinchwad College of Engineering and Research, Singapore Institute of Technology, Singapore*

[2] *Dayananda Sagar University, India*

[3] *Sir Padampat Singhania University, India*

Abstract: Transformer architectures have changed how multimodal medical image fusion works. Convolutional neural network (CNN)-based methods are constrained to local receptive fields; transformers use self-attention and cross-modal interactions to capture dependencies across the full image, which matters when aligning magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), and ultrasound—modalities that differ substantially in resolution, contrast, and the anatomical features they emphasize. This survey organizes transformer-based fusion models for healthcare along four axes: modality, pairing, fusion level, and attention and architecture design. Four representative models—MACTFusion, ECFusion, MATR, and DFENet—are examined in detail, with attention to how each handles anatomical preservation, semantic reasoning, and clinical interpretability. Fusion strategies fall into pixel-level, feature-level, segmentation, and neurodegeneration staging, and the survey covers the evaluation metrics and benchmark datasets used to assess performance across these tasks. A number of problems have not been resolved. Data scarcity and modality heterogeneity continue to limit generalization; interpretability remains shallow in most deployed architectures; and clinical environments impose constraints—compute budgets, latency requirements, and regulatory oversight—that most published models are not designed for. Work on lightweight vision transformers, language-guided prompting, 3D and temporal extensions, and federated learning addresses some of these gaps, though none yet offers a comprehensive solution. The survey aims to give researchers and clinicians a structured entry point into this area.

Keywords: transformer-based fusion, multimodal medical imaging, cross-attention mechanism, medical image segmentation, hybrid CNN–transformer models

1. Introduction

Medical imaging has proven to be an essential aspect of modern-day diagnostics, allowing for non-invasive visualization of internal anatomical structures and physiological functions with specificity and precision previously unavailable. Each imaging modality—magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), and functional MRI (fMRI)—alleviates only a small part of the diagnostic puzzle on its own: MRI has high soft-tissue contrast, PET provides metabolic activity, and CT provides detailed information about bone structure. Since none of these modalities can give a full clinical picture, especially for complex diseases (e.g., cancer, neurodegeneration, or cardiovascular disease) [1, 2], their use in isolation frequently results in fragmented interpretations.

Current methods usually do not account for complementary data from different imaging modalities; however, multimodal image fusion alleviates the issue by integrating heterogeneous imaging data into a single information-rich representation. But this is not just a visual merge; it also retains some critical morphological detail while highlighting functional differences. This allows for improved diagnosis, more accurate localization of diseases, better treatment planning, and disease monitoring. Heuristic multimodal fusion methods (e.g., multi-resolution decomposition, sparse coding, or joint statistical models) for integrating multimodality mainly use hand-crafted priors (which do not generalize to real clinical data) and therefore suffer from limited performance on high-dimensional non-linear clinical data [3, 4]. The advent of deep learning, and CNNs in particular, represented a paradigm shift since they allowed for the learning of fusion

***Corresponding author:** Nalini S Jagtap, Pimpri Chinchwad College of Engineering and Research, Singapore Institute of Technology, Singapore. Email: nalini.jagtap@pccoer.in

patterns directly from data. CNNs have been quite effective in feature extraction and pixel-level mapping between different modalities. However, their inflexible receptive fields limit their capacity for long-range dependency modeling, which is vital for some medical tasks where clinically relevant features may be separated by large spatial distances [5, 6]. For instance, matching a functional PET impairment with a small structural distortion in MRI requires this cross-scale context. Although originally developed for sequential tasks in natural language processing (NLP), transformers have achieved tremendous success in computer vision thanks to their self-attention mechanisms, able to establish global contextual relationships. Transformers are well suited to modeling positional relationships, enabling both within-modality and cross-modality interaction. This allows the network to attend to diagnostically relevant regions in one modality and link them to corresponding cues in another multimodal medical fusion [7, 8]. As a specific example, a recent model combines edge-aware transformer GANs with hierarchical attention modules to unify high-frequency PET signals with MRI images [5, 6].

In this context, recent work investigates various architectural approaches to adapt transformers for healthcare fusion pipelines. Other works maintain CNN encoders while adding transformer-based fusion blocks to preserve spatial information while achieving global semantic fusion [1]. Some use cross-attention or shared transformer branches to allow strong mutual alignment of features across modalities like PET and MRI, which is useful for Alzheimer's disease classification or for brain mapping [2, 9]. Transformers have demonstrated the ability to cope with modality gaps and enhance fusion robustness through techniques such as multiscale attention [10], self-supervised fusion [11], and conditional adversarial training. In addition, hybrid innovation like MATR, consisting of structured state space models with dynamic convolution and attention, tends to find a trade-off between performance and efficiency. Similarly, models like M3IFT2.0 and MRS Fusion showcase the value of combining Swin Transformer blocks with multi-branch CNNs and attention-guided weighting for better visual and quantitative fusion performance [7, 8].

However, available literature offers few dedicated, detailed reviews related to transformer-based multimodal image fusion in the medical domain. Many existing reviews either cover AI broadly across use cases or focus only on CNN-based and traditional fusion methods [4]. This survey is for transformer-based multimodal medical image fusion specifically. A framework that classifies and compares fusion models through a four-axis taxonomy consisting of modality pairing, fusion level, attention mechanism, and architectural design is proposed. Existing literature concerning transformer architectures such as MACTFusion, ECFusion, DFENet, and MATR from various aspects of the design trade-off, clinical applicability, and benchmark performance are compared. This survey also explains what transformers provide on top of CNNs: they alleviate the locality bias of convolutions, using adaptively learned weights to fuse inter-modal features rather than static kernels, and allow for end-to-end optimization across complex fusion tasks [8, 11]. Open challenges covered include data scarcity, modality heterogeneity, and cross-center generalization, with future direction mapped out including federated fusion learning, vision-language prompting, and 3D temporal transformers.

The remainder of the survey is organized as follows. The methodology governing the literature search and selection criteria is detailed in Section 2. Section 3 presents the historical overview of the image fusion techniques. Section 4 presents a four-axis taxonomy—modality pairing, fusion level, attention mechanism, and architecture—with canonical examples. Section 5 introduces a shared evaluation framework that unifies the different measurements and benchmark datasets for reproducible comparison. Section 6 considers practical aspects of data curation, efficiency, and explainability and describes open problems. Section 7 presents the near-term research directions, with Section 8 concluding the survey.

2. Literature Review Methodology

This survey draws on a systematic search of IEEE Xplore, Scopus, and Google Scholar, covering 2010 to 2025, with title, abstract, and keyword fields as the primary search targets. IEEE Xplore's Full Text & Metadata field was queried with ("transformer" OR "attention mechanism" OR "self-attention") AND ("multimodal fusion" OR "image fusion") AND ("MRI-PET" OR "MRI-CT" OR "MRI-SPECT" OR "medical imaging"). Scopus took the equivalent TITLE-ABS-KEY form: TITLE-ABS-KEY(("transformer" OR "attention mechanism") AND ("multimodal fusion" OR "medical image fusion") AND ("MRI" OR "PET" OR "CT" OR "SPECT")). Google Scholar has no field-restricted Boolean search, so we ran transformer AND "medical image fusion" AND (MRI OR PET OR CT OR SPECT) against the title and full text and screened the first ten relevance-ranked pages by hand. Across all three databases, results were limited to peer-reviewed journal articles and conference proceedings in English.

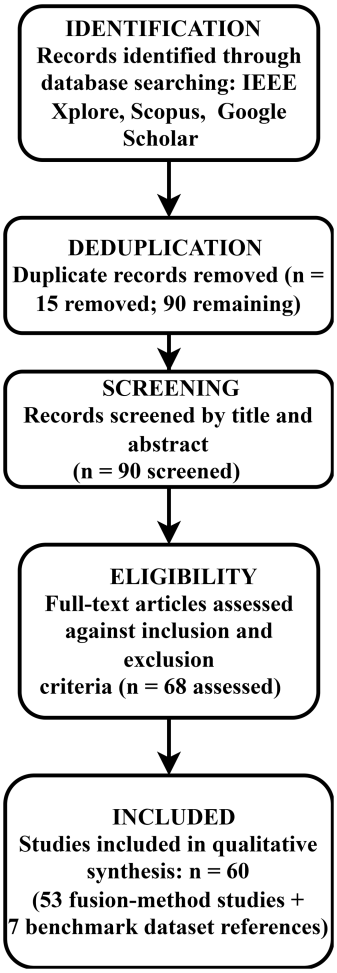
A total of 105 candidate records were identified from the initial database searches across IEEE Xplore, Scopus, and Google Scholar. Following the removal of 15 duplicates, there were 90 remaining unique records for title and abstract screening, with 22 excluded as out of scope. The remaining 68 articles were screened for full-text evaluation based on the aforementioned inclusion and exclusion criteria, out of which 8 were excluded through this process, leading to a final list of references containing 60 papers in total, comprising work proposing or evaluating transformer-based or hybrid CNN-transformer fusion architectures (n = 53) and additional supplementary references describing benchmark datasets employed for the evaluation (n = 7). The inclusion-exclusion process is illustrated in Figure 1.

Among these, 53 fusion-method studies were classified based on four axes: modality pairing, fusion level, attention mechanism, and architectural design to create the taxonomic framework introduced in the following sections. The remaining 7 references are benchmark datasets, which are used for evaluation all throughout the survey. The studies included were selected based on four main combinations of imaging modalities: MRI-PET, MRI-CT, MRI-SPECT, and multiple modalities in brain imaging with regard to clinically relevant applications such as tumor detection and neurodegeneration diagnosis and organ segmentation.

3. Historical Overview of Medical Image Fusion

The concept of image fusion started in early remote sensing and military surveillance applications dating from the late 1980s and 1990s, where data from multiple sensors, e.g., infrared, visible, and radar, were fused to improve interpretation and detection performance. This core concept of combining information from multiple complementary sources quickly moved to the medical imaging field as a way to address the limitations of single-modality scans like MRI, CT, PET, and Single Photon Emission Computed Tomography (SPECT). None of these modalities, on their

Figure 1
PRISMA-style flow diagram of the literature search and selection process



own, is able to provide a full diagnostic picture: MRI has excellent soft-tissue contrast, CT provides detail of the structural anatomy, and PET reflects associated major physiological processes, but each modality captures a separate physiological or anatomical domain.

Pixel-level or transform-based medical image fusion algorithms were conceived based upon remote-sensing methods such as pyramid decomposition, wavelet, and contourlet transforms. These classical methods (1990s–2010) were based on the concepts of multi-resolution analysis and spatial–frequency decomposition for pixel-level fusion of visual information. While the methods from the early days were simple, interpretable, and computationally efficient, they did not adapt well to more complex and non-linear intensity distributions among the medical images, and they were more sensitive to registration, especially in high-dimensional medical data.

More recently, since the early 2010s, when more multimodal medical datasets became available, researchers started to investigate sparse representation and statistical learning strategies. By using joint dictionaries or optimization-based formulations, these models were able to capture regularities from multiple modalities in shared sparse spaces while preserving contrast and robustness to noise, outperforming standard transforms. Sparse fusion methods were particularly effective in CT–MRI and MRI–SPECT fusion scenarios to enable soft-tissue delineation and organ segmentation, providing the mathematical framework for more complex data-driven fusion models.

The next paradigm (2016–2020) emerged in the era of deep learning, introducing end-to-end feature learning that eliminated most hand-crafted rules. Several CNN-based methods, such as DeepFuse, DenseFuse, and FusionGAN, proposed hierarchical feature-extraction pipelines that are able to capture both low-level texture and high-level semantics. These architectures performed better than sparse methods but still had the disadvantages of a local receptive field and small context, which limits long-range contextual reasoning, an essential component for medical imaging where spatial dependencies span across anatomical structures.

Considering the limitations of these purely convolutional-based frameworks led to the emergence of attention-guided (2014–2020) and hybrid CNN frameworks (2020–2022), which were being designed to accommodate long-range dependencies. These frameworks mix spatial-channel attention mechanisms and multi-scale modules to dynamically weight the cross-modal features. During this time, models that integrated convolutional encoders with attention blocks in an attempt to balance local detail preservation and global context understanding were also observed. This was somewhat of the inflection point where convolution fusion and some global-reasoning paradigms of transformer-based models were both on the rise. Such hybrid strategies led to the use of self-supervised and contrastive learning-based approaches, which offered strong complementary strategies for attention-based fusion. Methods combining Swin Transformers with self-supervised contrastive learning [12] demonstrated that hierarchical attention mechanisms could be further strengthened by learning modality-invariant representations without dense annotation, achieving notable improvements in SSIM, edge sharpness, and color fidelity across MRI–CT and MRI–PET fusion tasks. Multiscale attention architectures such as FAMAFuse [13] further advanced this direction by integrating spatial attention residual modules and learnable Gaussian kernels to capture both local details and global context across SPECT–MRI, PET–MRI, and CT–MRI pairs. For task-driven fusion, transformer-based frameworks for Alzheimer's disease diagnosis demonstrated that progressively fusing sMRI and PET features through improved transformer blocks could achieve expectational classification accuracies, while multimodal transformer networks [14] addressed the clinically prevalent problem of missing modalities through guided generative adversarial networks combined with cross-modal attention. On the reconstruction side, MLSDA-GAN [15] introduced modality-specific self-attention and cross-attention fusion blocks for PET reconstruction with auxiliary MRI, improving both long- and short-range feature modeling. Finally, progressive parallel frameworks such as PPMF-Net [16] combined dynamic edge enhancement, non-linear feature coupling, and transformer-based global modeling for PET–MRI fusion, demonstrating strong generalization across multiple modality pairs. Together, these developments set the stage for the specialized transformer fusion frameworks reviewed in the subsequent sections. Table 1 summarizes key events in the development of clinical imaging techniques (with representative studies, fusion strategies, and clinical relevance) for the years between 2014 and 2024 in a chronological order.

The following sections focus on established transformer-based architectures for medical image fusion, building on the advances summarized in Table 1. This section focuses on using attention mechanisms and transformer architectures to

Table 1
Review of medical image fusion studies (2014–2024)

Year	Title	Explanation/Key contribution	Limitation
2014	MRI and PET Image Fusion Using Fuzzy Logic and Image Local Features [17]	Classical pixel-level fusion combining MRI structure with PET function using local features and fuzzy logic	Sensitive to feature selection and fuzzy rule design; limited adaptability
2019	Improved Deep CNN Model for Image Fusion [18]	Integrated Gaussian–Laplacian filters with a deep CNN for fast fusion	Shallow network; limited generalization
2020	IFCNN: General Image Fusion Framework [19]	End-to-end CNN handling diverse modalities, offering strong generalization	Requires large training data; low interpretability
2020	Multimodal Medical Image Fusion Algorithm in the Era of Big Data [20]	Boundary-measured pulse-coupled neural network fusion in the non-subsampled shearlet domain for large-scale multimodal fusion	Memory-heavy; sensitive to noise
2021	DSAGAN: Dual-Stream Attention GAN for Anatomical and Functional Image Fusion [21]	GAN-based dual-stream attention mechanism fusing anatomical (MRI) and functional (PET-like) image information	Adversarial training instability; limited interpretability of attention maps
2021	EMFusion: An Unsupervised Enhanced Medical Image Fusion Network [22]	Unsupervised encoder-decoder using entropy-based information evaluation for lightweight, robust fusion	Oversharpened output; reduced global structural coherence
2021	An Effective Multimodal Image Fusion Method Using MRI and PET for Alzheimer's Disease Diagnosis [23]	Region-guided MRI–FDG PET fusion focusing on gray matter to enhance AD-relevant structural and metabolic information; validated using 3D CNNs on ADNI data	Task-specific; depends on accurate registration and tissue segmentation
2022	MsRAN: Multi-Scale Residual Attention Network [24]	Residual attention with dual-adversarial training for rich multimodal fusion detail	Complex model; training demands large datasets
2022	MATR: Multimodal Medical Image Fusion via Multiscale Adaptive Transformer [25]	Adaptive convolution and multiscale transformer blocks address long-range dependency modeling for SPECT–MRI fusion	Computational overhead increases significantly at higher spatial scales
2022	PET and MRI Image Fusion Based on a Dense Convolutional Network with Dual Attention [26]	Deep encoder–decoder with spatial and channel attention to preserve PET functionality and MRI detail	Data- and compute-intensive; reduced interpretability
2022	An Adaptive MRI–PET Image Fusion Model Based on Deep Residual Learning and Self-Adaptive Total Variation [27]	Residual learning combined with adaptive TV regularization for structure-preserving MRI–PET fusion	Risk of over-smoothing; limited large-scale validation
2023	Parameter Adaptive Unit-Linking PCNN Based MRI–PET/SPECT Image Fusion [28]	Hybrid NSST–PCNN framework with automatic parameter adaptation for multi-scale fusion across modalities	High computational cost; complex pipeline limits scalability
2023	MRSCFusion: Joint Residual Swin Transformer and Multiscale CNN for Unsupervised Multimodal Medical Image Fusion [29]	Combines Swin Transformer and CNN in a two-stage unsupervised framework; captures global context and local fine-grained texture across CT-MRI, PET-MRI, and SPECT-MRI pairs	Unsupervised training may underperform on task-specific clinical objectives; 3D scalability not addressed
2024	MMTFN: Multi-Modal Multi-Scale Transformer Fusion Network for Alzheimer's Disease Diagnosis [30]	Jointly learns MRI and PET representations using 3D multi-scale residual blocks and transformer layers; achieves 94.61% accuracy on ADNI dataset	Limited to binary/multi-class AD diagnosis; generalizability to other neurodegenerative conditions not validated

enhance global feature reasoning, cross-modality alignment, and interpretability. Key transformer-based frameworks, their architectures, taxonomy, and specific applications discussed here build the foundation of the current state of the medical image fusion field as well as its way to the future.

4. Taxonomy of Transformer-Based Medical Image Fusion in Healthcare

Transformer-based fusion models have proven to be versatile across various medical imaging tasks, yet the diversity in model architecture, attention mechanism, and modality integration scheme obscures their common ground. The survey is organized around four-axis taxonomy: (i) modality pairing, (ii) fusion level, (iii) attention mechanism, and (iv) transformer architecture. This structured framework enables comparison of features across these models, demonstrating how architectural choices support clinical requirements and image-resolution trade-offs, and it is illustrated in Figure 2. A deeper look into each axis, including representative models and clinical relevance, is provided in the following subsections (4.1–4.6).

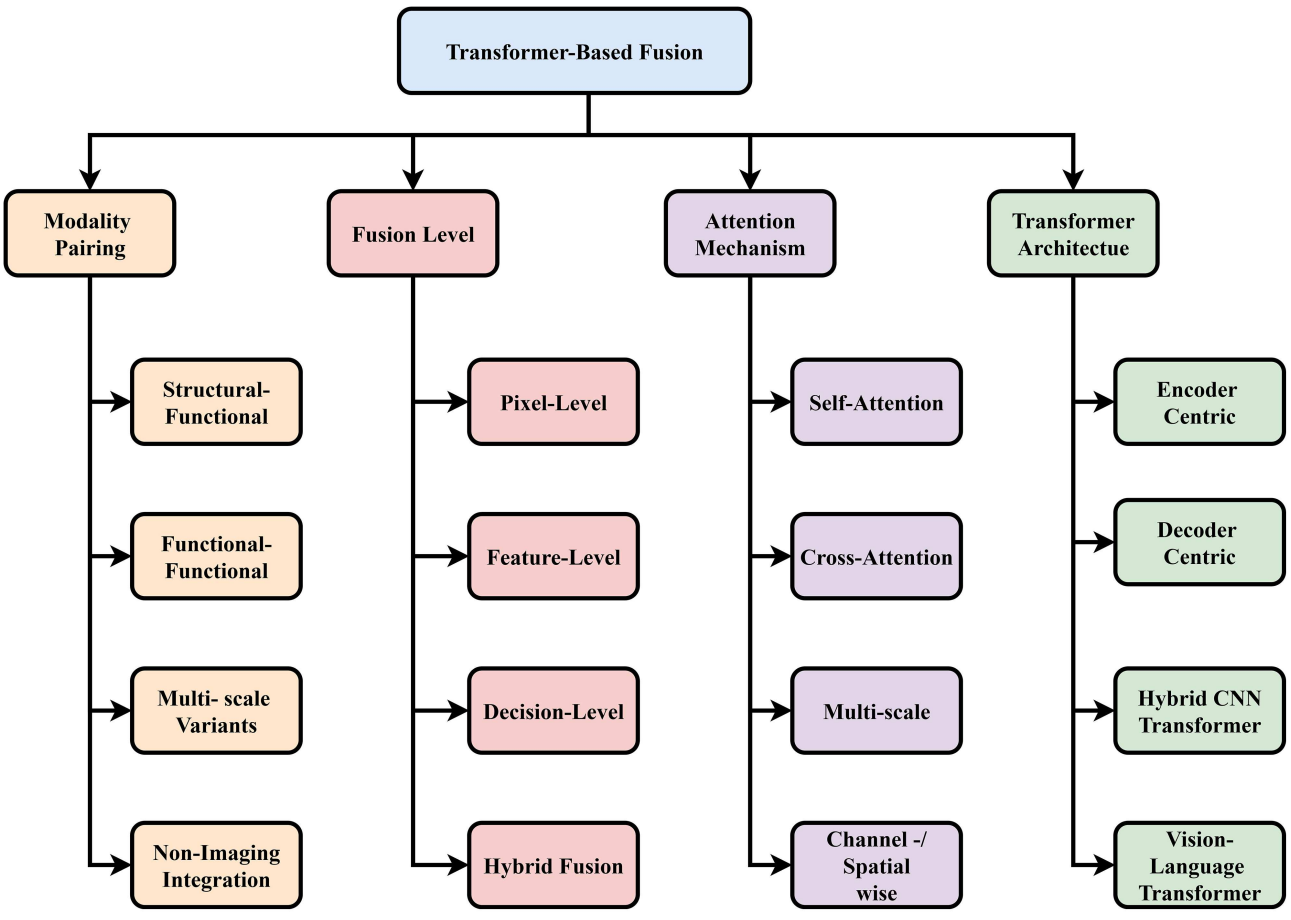
4.1. Modality pairing

The selection of imaging modalities underpins the design of a fusion model and its diagnostic characteristics. However, in medical applications, modalities are selected based not only on the complementary information each source can provide but also by usage constraints like spatial resolution, cost of acquisition, safety for the patient, and anatomical coverage.

4.1.1. Structural–functional pairing

A common approach for imaging the brain (i.e., MRI), cancer (i.e., PET), and metabolism (i.e., SPECT or fMRI) is to link structural modalities (e.g., MRI) with functional (e.g., PET) or with metabolic (e.g., SPECT or fMRI) modalities. Such pairs allow the model to associate anatomical localization with physiological activity. For example, in the case of MACTFusion [31], the cross-window and cross-grid attention that aligns structural edges from MRI and the metabolic intensities in PET is utilized to obtain the tumor boundary more clearly but with less texture degradation. A lightweight transformer architecture is used in the system that minimizes the computational cost while keeping precise multimodal feature extraction. ECFusion [32] takes this idea a step further by incorporating edge priors via a Sobel-based edge module and merging them with cross-scale transformer blocks to maintain spatial consistency when multimodal images differ in spatial resolution and contrast. However, both models are primarily validated on brain imaging datasets, limiting their generalizability to other anatomical regions and modality combinations such as CT-SPECT or ultrasound-MRI.

Figure 2
Four-axis taxonomy of transformer-based medical image fusion methods categorized by modality pairing, fusion level, attention mechanism, and architecture



4.1.2. Multi-scale and multimodal variants

MATR [25] proposes an adaptive fusion scheme that operates across multiple spatial scales to address variations in anatomical details and modality-specific resolution differences. Multiscale adaptive transformer blocks enable arbitrary contextual feature extraction, which is critical in clinical applications where different body parts have vastly different imaging properties, such as multi-organ segmentation or lesion detection. Yet, despite resisting parameter overfitting in larger pools, the computational overhead of MATR increases with spatial scale and is unwieldy for real-time clinical implementation.

4.1.3. Functional–functional fusion

Functional–functional pairings, though not utilized as often, such as fMRI with EEG can open new doors in the field of cognitive neuroscience and neuropsychiatric disorder investigations. This direction is first explored in MultiViT [33], which uses cross-attention transformers to fuse functional connectivity matrices of fMRI and gray matter maps derived from 3D structural MRI. This allows the network to learn subtle relationships between neural activity and anatomical context, which is helpful for diseases like schizophrenia or Alzheimer's that impact spatial and functional areas of the brain simultaneously. Thus, the future of functional-functional fusion is primarily limited by a lack of paired fMRI-EEG datasets and challenges associated with achieving reliable spatial co-registration across modalities that have fundamentally different resolutions.

4.1.4. Multimodal non-imaging integration

Li et al. [34] integrated imaging features with clinical parameters and histopathology-inspired cues into a transformer-based fusion model, advancing multimodal clinical image fusion and enabling personalized diagnosis from multimodal electronic health records. However, such non-imaging integration introduces significant heterogeneity, and reliance on clinical metadata limits applicability in settings where such records are incomplete or unavailable.

Table 2 summarizes representative transformer-based fusion models across common combinations of medical imaging modalities. Each combination is selected to leverage complementary clinical data, e.g., combining metabolic information from PET with structural information from MRI. These modality selections are tailored towards particular diagnostic goals, ranging from tumor tracking of neurodegeneration to multi-system classification tasks. Table 2 lists the models that best leverage each pairing.

4.2. Fusion level

The term fusion level denotes the stage at which the data from different modalities are combined, i.e., raw pixels or feature maps or decision outputs. This axis of the taxonomy is especially relevant since it pairs interpretability with robustness against modality misalignments or missing acquisition. Different clinical applications impose different requirements on the level of fusion granularity desired: segmentation tasks prefer feature-level precision at a fine-grained grade, diagnosis typically calls for decision-level interpretability at a coarse granularity, and visualization pipelines often benefit from integrating pixel-wise. Transformer-based frameworks accordingly incorporate fusion methods across these levels, aligning representational design with the degree of integration required and its medical relevance. The taxonomy of the fusion-level models is shown in Figure 3, which maps pixel-, feature-, decision-, and hybrid-level schemes to their clinical application domain.

4.2.1. Pixel-level fusion

Pixel-level fusion combines multiple-modality images directly in the intensity domain before semantic encoding: input images are stacked or merged at the pixel level. Being intuitive, lightweight, and interpretable, it is well suited to quick visualization. It is, however, highly sensitive to registration error and intensity mismatch or noise and is not really suited to diagnostic reasoning tasks. A representative example is Odusami et al. [35], who apply a Vision Transformer to pixel-level fusion of MRI and PET images for early detection of Alzheimer's disease, demonstrating that pixel-wise integration can support downstream classification when combined with transfer learning. Pixel-wise fusion remains applicable where fused overlays facilitate quick radiological review, such as retinal disease visualization or portable real-time systems.

4.2.2. Feature-level fusion

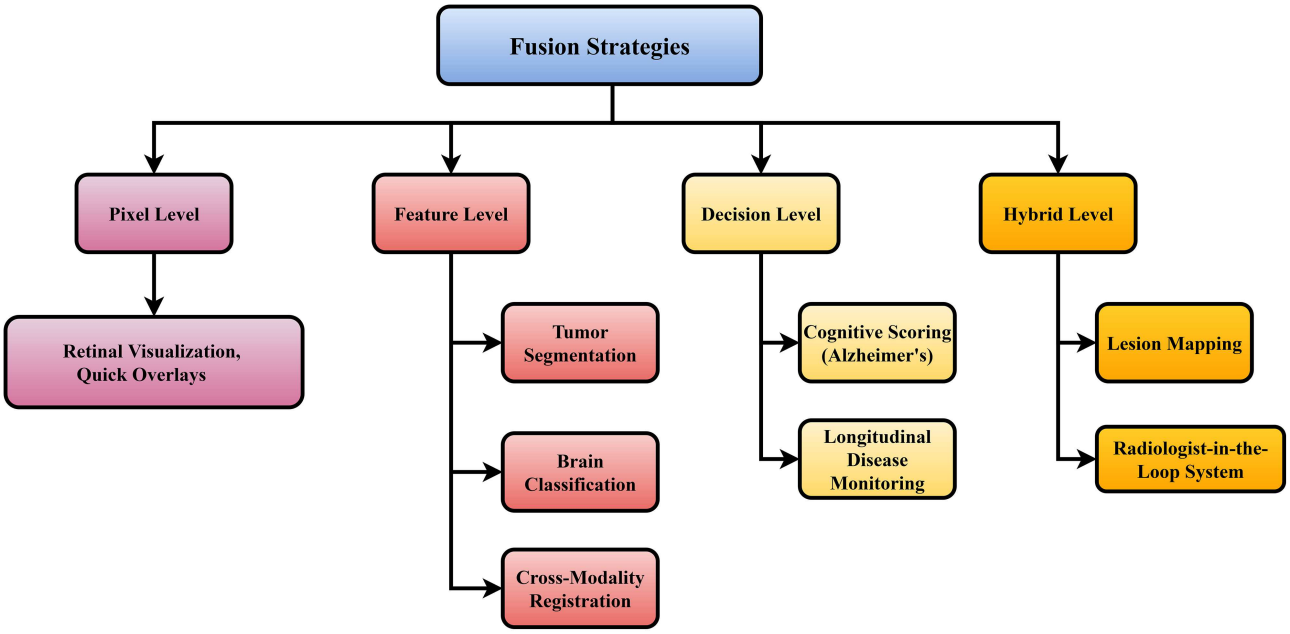
Feature-level fusion is the most widely used fusion strategy in transformer-based medical image fusion. Each modality is first encoded individually—by CNNs, Swin Transformers, or hierarchical ViTs—and the resulting feature maps are aligned and aggregated using attention. This method serves as a trade-off between interpretability and semantic abstraction, resulting in robust and semantically relevant representations.

This approach has been implemented in DFENet [36], which connects CNN-based encoders for modality-specific learning with transformer-based encoders for global context, fusing the two through a semantic aggregation block. Feature-level fusion is especially well suited for the segmentation and classification problems such as brain tumor identification, vascular tracking, and cross-modality registration. Its flexibility to combinations of imaging pairs such as CT–MRI, PET–fMRI, and fundus–OCT explains why it is the most widely used option in the most recent frameworks.

Table 2
Modality pairing and their associated clinical applications

Modality pair	Representative model(s)	Clinical application
MRI + PET	MACTFusion [31], ECFusion [32]	Neurodegeneration, tumor detection
CT + MRI	MATR [25]	Bone and soft-tissue analysis
fMRI + EEG	MultiViT [33]	Functional–functional integration
Ultrasound + Clinical Data	Li et al. [34]	Breast cancer diagnosis

Figure 3
Taxonomy of fusion strategies mapped to clinical application types



4.2.3. Decision-level fusion

Decision-level fusion combines modalities at the output level, after each has passed through a dedicated branch. The independent predictions—classification scores, segmentation masks, or risk estimates—are combined by means of ensemble averaging, attention-weighted consistency, and transformer-guided voting. An example of this is the transformer-based medical report generator by Raminedi et al. [37], in which ViT embeddings are concatenated with language tokens inside a GPT2 decoder and attend to a language summary during cross-attention to output image-grounded reports. Decision-level fusion is appealing where modular interpretability and auditability are required, since the contribution of each modality can be verified by clinicians. It has been used in the diagnosis of Alzheimer's disease, neurocognitive scoring, and longitudinal disease progression monitoring. While less popular than feature-level approaches, it is particularly important in risk scoring and multi-stage workflows where patient-specific data streams must be integrated asynchronously.

4.2.4. Hybrid fusion

Hybrid fusion involves an integration of the complementary benefits across pixel, feature, and decision levels, which can ensure higher robustness in cases where one or more modalities are missing or corrupted. An extreme of this paradigm is ECFusion [32], which combines the edge-level graph features with multi-depth image features via a cross-scale attention block to achieve fine-grained local representation while maintaining semantic consistency.

A second example is MACTFusion [31], a PET–MRI fusion framework that has been hyperparameter-tuned to balance diagnostic challenges with compute budgets. PET images are mapped to YCbCr space, the Y channel is decomposed for fusion, and MRI images are pre-processed in grayscale. The features of both inputs are based on convolutional layers and a gradient residual dense block to capture low-level textures as well as high-level semantic representations, which are arranged hierarchically in shallow and deep feature encoders for better extraction. MACTFusion is designed with a Cross Multi-Axis Transformer, which includes two symmetric mechanisms: Cross-Window Attention (CWA) and Cross-Grid Attention (CGA). CWA captures fine spatial coherence via local patch-wise attention, and CGA models long-range dependencies through larger spatial blocks. The proposed bilinear cross-modality attention flow facilitates information exchange and alignment between two modalities. These features go to a reconstruction module and produce the final image in RGB space. MACTFusion achieves a good compromise between fusion quality and computational demand, essential for real-time use in the clinic. Figure 4 [31] provides a full architectural flow representing the two-stream feature encoding, inter-stream attention, and reconstruction stages necessary for eventual real-time, clinical application.

Hybrid approaches, like MACTFusion, illustrate the utility of hybrid fusion as a means to combine pixel-level matching accuracy, feature-based richness, and decision-level modularity. They are becoming attractive for challenging (in terms of sparseness) clinical tasks such as brain tumor segmentation, lesion registration, and the whole-brain classification that call for robust handling of imperfect inputs.

Table 3 provides an overview of the taxonomy for transformer-based fusion strategies and arranges models based on pixel-level fusion, feature-level fusion, decision-level fusion, and hybrid approaches. It describes their primary design features, examples of clinical applications to which they are applied, and the rationale for each stage of integration. This overview demonstrates how various fusion levels respond to different medical tasks, from lightweight visualization overlays at the pixel level to robust multi-organ analysis with hybrid designs.

4.3. Attention mechanism

The attention mechanism forms the computational heart of transformer-based fusion systems. It determines how information

Figure 4
Architectural pipeline of MACTFusion showing dual-stream encoding, inter-stream attention modules, and final reconstruction for PET–MRI fusion

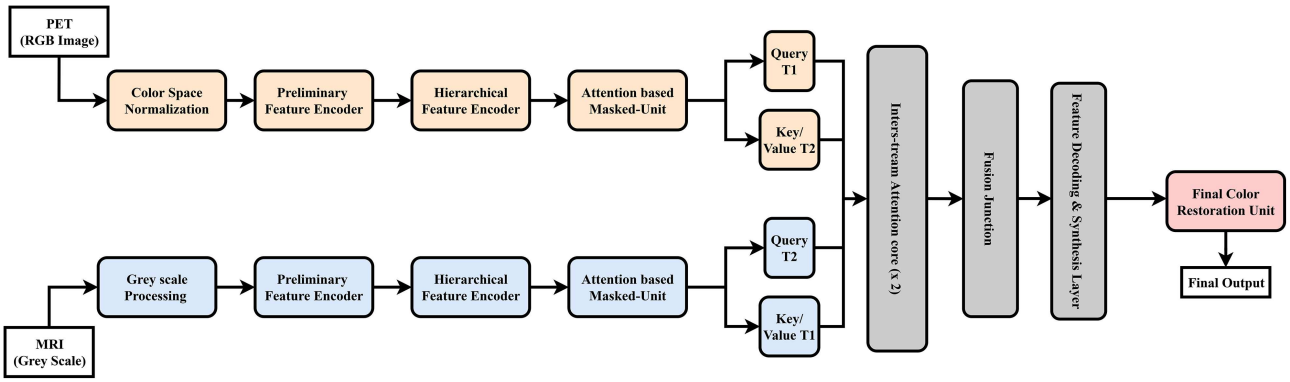


Table 3
Taxonomy of transformer-based fusion levels with representative models, core characteristics, and clinical application

Fusion level	Representative models	Core characteristics	Clinical tasks/suitability	Fusion rationale
Pixel-level	ViT-based Pixel-Level Fusion for Alzheimer's Diagnosis [35]	Fusion at raw intensity; sensitive to misalignments but preserves fine visual cues	Retinal disease visualization; low-dose enhancement; registration tasks	Provides lightweight composite overlays for rapid radiological assessment; focuses on clarity over deep semantic reasoning
Feature-level	DFENet [36]; TUFusion [38]; HCFMaNet [39]; MultiViT [33]	Features fused after encoding; balances local + global cues; enables semantic reasoning	Brain tumor segmentation; vascular mapping; whole-brain classification	Requires fine spatial attention, multiscale fusion, and semantic consistency; token-based attention improves region traceability
Decision-level	ViT-GPT2 Report Generator [37]	Independent modality outputs fused post-classification; ensemble/attention-based decision consistency	Alzheimer's/dementia diagnosis; neurocognitive scoring; longitudinal disease monitoring	Modular interpretability allows independent audit of each modality; suitable for asynchronous and progressive disease modeling
Hybrid-level	MACTFusion [31]; ECFusion [32]; MATR [25]	Multi-stage fusion at pixel, feature, and decision levels; combines local fidelity and global reasoning	Brain tumor analysis; lesion/vessel registration; multi-organ imaging	Enhances robustness and improves consistency by combining early pixel detail, mid-level semantic reasoning, and decision supervision

from different spatial locations and modalities is weighted and aggregated during the fusion process.

4.3.1. Self-attention

Self-attention [40] enables intra-modality interactions by modeling dependencies between all token pairs within a modality. This is particularly beneficial when dealing with modalities with internally complex textures, such as MRI or dermoscopy images.

4.3.2. Cross-attention

Cross-attention underlies most multimodal integration work. MultiViT [33] uses it to align spatial and functional features learned from image pairs: one modality queries the other to refine its own representation. The mechanism handles complementary information well and limits redundancy, but it does not resolve the more basic problem of keeping alignments interpretable across heterogeneous modalities. This is sharpest in functional imaging, where metabolic or perfusion maps must register accurately onto anatomical structures. Cross-attention helps, but if registration is error-prone or the inputs carry a semantic mismatch, the mechanism has little to work with.

4.3.3. Channel-wise and spatial attention

Beyond spatial alignment, channel-wise attention controls how much weight different feature dimensions receive. This is relevant when modalities encode largely independent information—structural density and perfusion rate, for instance, share little statistical overlap—and blunt aggregation would suppress one at the expense of the other.

Table 4
Comparison of attention mechanisms used in transformer-based fusion models, highlighting their purpose and application relevance

Attention type	Example models	Purpose	Best for
Self-attention	Self-attention transformer [40]; Dual-attention multimodal fusion [41]	Enhances modality-specific context	Texture refinement, contrast enhancement
Cross-attention	MultiViT [33]; AdaFuse [42]	Learns inter-modal dependencies	PET–MRI alignment, schizophrenia classification
Channel-wise	ECINFusion [43]	Reweights semantic feature bands	Cardiac imaging, multi-organ fusion

The attention mechanism chosen also determines how the model selects, weights, and aggregates features across spatial and modality axes. Table 4 summarizes strategies spanning self-attention, cross-attention, multi-scale, and channel-wise approaches, with representative models and their clinical applications. Together they address spatial misalignment, modality imbalance, and resolution heterogeneity—the three problems that recur most consistently across fusion tasks.

4.4. Transformer architecture

Architectural design shapes computational cost, scalability, and how well a fusion system performs diagnostically. The main configurations are encoder-centric, decoder-centric, and hybrid.

4.4.1. Encoder-centric models

Encoder-centric models apply transformer layers immediately after feature extraction. DFENet [36] is a representative model that combines CNN and transformer modules in its encoder to learn both local and global representations. This architecture benefits from reduced inference latency and is ideal for embedded or edge applications in IoMT settings.

4.4.2. Decoder-centric models

Decoder-centric architectures incorporate transformers in the reconstruction or interpretation phase. These systems excel in cases where the primary goal is downstream clinical interpretation, such as lesion delineation or semantic segmentation.

4.4.3. Hybrid architectures

Hybrid transformer architectures are currently the most widespread in medical fusion research. MACTFusion [31] is a prime example, employing CNN blocks for local feature extraction and transformers for modality alignment and global reasoning. Such designs combine the strengths of both paradigms: the low-level fidelity of convolution and the long-range dependency modeling of transformers.

4.4.4. Transformer-augmented language models

Recent designs pair transformer vision encoders with language models such as GPT-2 to generate interpretable radiology reports. Cross-attention operates not only across image modalities but also between visual and textual inputs—an extension that makes these systems relevant to clinical decision support, not just image fusion. The choice of architectural backbone determines where efficiency, flexibility, and diagnostic performance are prioritized, since gains in one typically come at a cost to another. Architectures can be grouped by where transformer modules appear in the pipeline; Table 5 covers the main classes, their design characteristics, and their advantages. The grouping reflects a practical reality: low-latency inference, semantic comprehension, and report generation place different demands on the system, and no single design satisfies all three equally.

4.5. Transformer-based architectures in medical image fusion

Transformer architectures in medical image fusion offer a diverse set of design paradigms, moving beyond the local receptive fields of convolutions to capture global semantic interaction. This shift has led to highly expressive and task-specific fusion frameworks capable of aligning spatial, structural, and contextual information across medical images.

Table 5
Categorization of transformer-based fusion architectures with representative models, core principles, and clinical strengths

Architecture	Representative models	Core idea	Strengths
Encoder-centric	DFENet [36]	CNN + Transformer encoding before fusion	Lightweight, ideal for IoMT, early fusion
Decoder-centric	TransFusion [3]	Transformer-based reconstruction of fused anatomical-functional images in the decoding stage	Improved semantic understanding, end-to-end anatomical-functional fusion
Hybrid (CNN + ViT)	MATR [25]	Combines CNN local features + transformer global reasoning	Balanced modeling, robust performance
Multimodal vision-language	ViT-GPT2 Report Generator [37]	Vision transformer encodes images, fused with text decoder	Report generation, interpretable AI outputs

4.5.1. ECFusion: edge-augmented cross-scale transformer for enhanced contrast preservation

ECFusion [32] (introduced in section 4.1.1 as a structural-functional fusion example) addresses a frequent limitation of multimodal fusion by ensuring sharp structural edges and contrast, which is critical for high-precision clinical tasks such as organ delineation or tumor localization. This architecture processes both content and edge information in two streams. Input modalities are mapped to the YCbCr domain to separate luminance, with encoders extracting features in parallel. The first encoder is equipped with an Edge-Augmented Module (EAM), which uses the Sobel operator for edge map extraction. These maps are then fed through residual blocks of varying depths to encode multi-scale edge features.

Fusion is managed by a Cross-Scale Transformer Fusion (CSTF) module with a Hierarchical Cross-Scale Embedding Layer (HCEL). Across the stack, the transformer builds attention maps at different resolutions, integrating global context and local structure across the input via this layer. The CSTF uses normalization and feed-forward units between self-attention layers to iteratively refine the feature maps. The decoder then concatenates the multi-scale fused features and applies residual enhancement to preserve intensity and edge information during reconstruction.

The full architectural layout of the ECFusion model—highlighting the dual-stream EAM processing, cross-scale fusion blocks, and final decoder—is illustrated in Figure 5 [32]. ECFusion performs well even if inputs are not aligned or differ in intensity scale. However, ECFusion may underperform when Sobel-based edge priors are applied to low-contrast modalities where edge boundaries are inherently ambiguous, such as PET or fMRI.

4.5.2. MATR: multiscale adaptive transformer for semantic context integration

MATR [25] (introduced in Section 4.1.2 as a multi-scale fusion example) addresses the limitation of conventional convolution-based fusion networks by capturing long-range dependencies and global context that such networks cannot model. At the core, the model starts with a strong multi-resolution convolutional-encoder-based feature extraction pipeline. Rather than conventional convolutional layers with static kernels, MATR uses adaptive convolution whereby the kernels are modulated using predicted global semantic priors so that the local filters can adapt to the visual context.

The centerpiece of MATR is the Multiscale Adaptive Transformer Block (MATB), which uses multi-head attention and dynamic positional encoding to extract multiscale features across resolution. Each of these transformer layers works independently per scale, and their outputs are fused in a channel-wise manner and combined through a residual feed-forward network. The architecture can be regularized with a dual-objective loss function: a structural similarity loss for preserving anatomical integrity and a region mutual information loss for enforcing inter-modality complementarity.

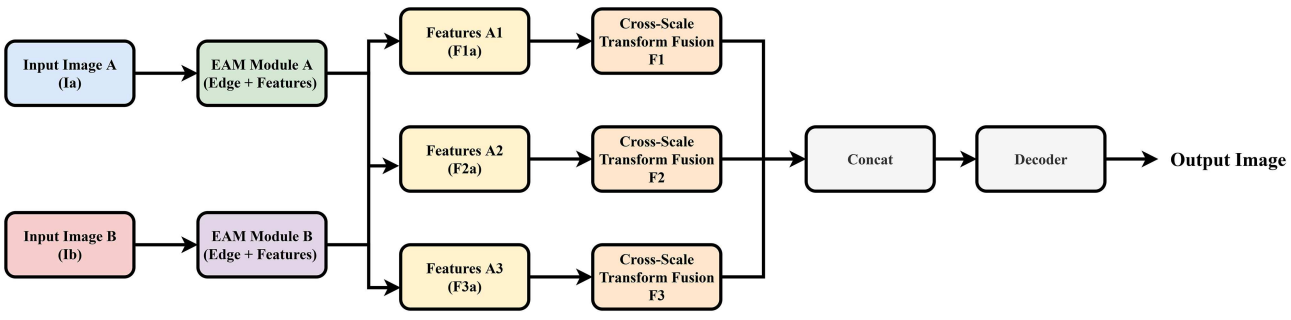
This architecture enables MATR to robustly handle clinical problems, such as PET–MRI co-registration in oncology or CT–MRI co-registration in neurosurgical planning. Its multi-scale design enables its strong performance against significant differences in resolution or intensities between the modalities. Despite its strong multi-scale performance, MATR's adaptive convolution and dual-objective loss increase training complexity, and its validation remains largely limited to brain imaging datasets.

4.5.3. Breast tumor fusion model: transformer-based clinical-image integration

The breast tumor fusion model (introduced in Section 4.1.4 as a non-imaging integration example) presents a solution for combining imaging and clinical data to classify breast tumor malignancy. It integrates ultrasound image data with structured clinical metadata for full diagnostic modeling [34]. It has three main input streams: ROI-based B-mode ultrasound images, entire B-mode images, and patient metadata (demographic and clinical characteristics). Each data stream is processed by a dedicated encoder: EfficientNet extracts features from the image data, while XGBoost processes the tabular features. The resulting feature representations are then passed to a Vision Transformer (ViT), where they are transformed into token embeddings and fused across modalities.

The multimodal embeddings are then modeled using layers of self-attention in the ViT, which can model complex relations between visual and clinical features. Moreover, positional encodings are applied to preserve modality identity and spatial continuity. Fully connected layers are applied to the final embeddings to perform binary classification, predicting benign versus malignant tumors. The remarkably high sensitivity across all validation cohorts implies that the fusion technique based on a transformer trained on standard radiomic and clinical features data can be easily integrated into routine clinical work. The design was modular and therefore could also be extended to other domains such as cardiovascular risk prediction or liver disease stratification. However, this model is only applicable in the

Figure 5
ECFusion architecture illustrating dual-stream edge-enhanced encoding, cross-scale transformer fusion modules, and the final decoding pathway



presence of full-scale data and for binary classification, as it relies on structured clinical metadata.

4.5.4. MDC-RHT: multimodal fusion using residual hybrid transformers with dynamic convolutions

MDC-RHT [44] is a transformer-based model that fuses PET-MRI. It converts PET images to the YCbCr space to match greyscale MRI and introduces a dynamic convolutional block (DCB) for feature extraction and spatial fidelity.

These outputs are then passed to Residual Hybrid Transformer (RHT) modules combining global attention and the local convolutions within a hierarchical and parallel model so that contexts can be preserved without loss of resolution. The RHT branches are combined through a residual bottleneck, and the Y channel is merged with the Cb and Cr channels to reconstruct a high-quality RGB image. As MDC-RHT is modality-specific but retains semantic consistency, it can also be used for both tumor detection and metabolic analysis, as shown in Figure 6 [44]. Nevertheless, its modality-specific approach for PET-MRI leads to limitations in generalizability, and the integration of dynamic convolution blocks increases inference time, making it less compatible with real-time clinical deployment.

4.5.5. DFENet: dual-branch feature enhancement for robust multimodal fusion

DFENet [36] (introduced in Section 4.2.2 as a feature-level fusion example) tackles a well-known gap in CNN-based fusion—the inability to capture context beyond a local receptive field—by pairing a CNN branch with a Vision Transformer in a shared encoder-decoder. The CNN side uses a Res2Net backbone with dilated convolutions at three scales to extract fine texture and edge detail. The overlapped patch merging is used in the transformer side part to construct a global context of different resolutions. They are interconnected via a Global Semantic Feature Aggregation Module (GSFAM) to aggregate multi-scaled outputs of transformers into a single spatially coherent representation. In fusion, a Local Energy and Gradient Maximum (LEGM) strategy chooses features based on local energy and gradient activity; this makes the approach more robust against noisy inputs. The decoder uses dense skip connections to recover spatial detail during reconstruction. Together, these five architectures show a shift in how the field tries to close the gap between benchmark performance and clinical usability. ECFusion and MATR attack the problem at the feature level: an edge prior in one case and an adaptive multiscale block in the other, integrated into an otherwise standard transformer pipeline. Both report strong image-quality metrics, and neither touches deployment cost. MDC-RHT takes the same feature-level logic further, adding a fourth convolutional dimension and overlapping cross-attention, and the added complexity increases inference latency rather than reducing it. Quality metrics, in other words, are being optimized at the direct expense of the efficiency real-time or intraoperative use would actually require. DFENet takes a different route: it borrows capacity from a non-medical domain through MS-COCO pretraining instead of adding architecture. That sidesteps data scarcity, but it introduces a domain-gap risk none of the other four models face: ECFusion, MATR, MDC-RHT, and the breast tumor model are all trained and validated end-to-end on medical data. The breast tumor fusion model sits apart from the rest of the group. It is the only one built around structured clinical metadata rather than a second image modality, and its transformer runs on modality-agnostic tokens rather than paired image patches. That design trades generalizability, since it only works when tabular clinical data is available, for an output a clinician can actually use: a binary malignancy prediction rather than a fused image. The pattern across all five shows architectural sophistication outpacing clinical validation. Only the breast tumor model reports an outcome a clinician would recognize as decision-relevant; the other four are validated almost entirely on image-quality metrics that say little about how they would perform in an actual diagnostic workflow.

Table 6 compares seven representative transformer-based fusion models on MRI-PET, setting image-quality metrics against design assumptions and clinical readiness. No single design dominates; each entry reflects a different set of trade-offs.

ECFusion's MI figure is substantially higher than the rest, but the gap is a measurement artifact. When ECFusion's authors re-ran EMFusion under their own MI formula, EMFusion's score rose from 0.7105 to over 3.0. Log base, bin count, and normalization choices differ across papers, which means MI values in this column cannot be read as directly comparable. VIF and SSIM are more consistent across implementations.

KD^{3}MT and DFENet record the highest VIF scores, which is unsurprising given their architectures: KD^{3}MT leans on a teacher network; DFENet is pre-trained on MS-COCO before fine-tuning on medical data. The performance advantage comes with a cross-domain transfer risk—neither model encounters the small medical datasets used here until after training on non-medical sources. EMFusion and MDC-RHT sidestep this

Figure 6
Architecture of MDC-RHT showing YCbCr decomposition, dynamic convolutional encoding, hybrid residual transformer layers, and RGB image reconstruction

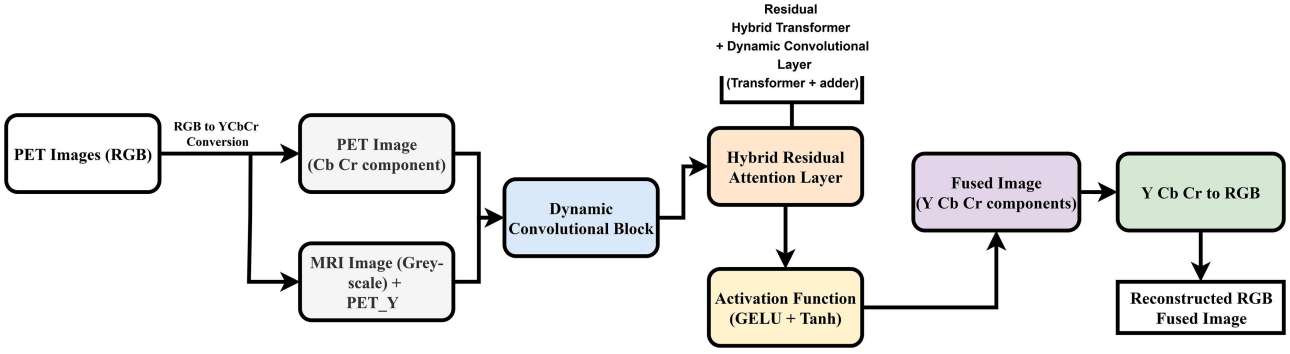


Table 6
Quantitative performance comparison of representative transformer-based fusion models on MRI–PET

Method (Year)	Architecture/Key innovation	MRI-PET Metrics	Parameters (M)	FLOPs (G)	Inference time (s)	Strengths	Weaknesses	Clinical feasibility
KD^3MT [45]	Knowledge distillation-driven dynamic mixer transformer; wavelet domain fusion (WDF) module	MI: 3.97; VIF: 0.35; MS-SSIM: 0.76	0.06	3.64	0.35	Teacher-student distillation improves generalization with limited labeled data; captures both local and global correlations via dynamic mixer transformer	Requires pre-trained teacher network, adding training complexity; wavelet domain fusion may be sensitive to modality-specific frequency characteristics	Moderate—strong quantitative performance but no reported clinical latency benchmarks
MATR (2022) [25]	Multiscale adaptive transformer; adaptive convolution replaces vanilla conv	MI: 0.8151; VIF: 0.4187; SSIM: 0.5364	N/R	N/R	N/R	Adaptive conv captures global complementary context; multi-scale design	Local texture defective in magnified regions; higher parameter complexity	Moderate—unsupervised; no latency analysis reported
EMFusion (2021) [22]	Unsupervised encoder-decoder; entropy-based information evaluation	MI: 0.7105; VIF: 0.5396; SSIM: 0.5348	0.30	N/R	N/R	Lightweight; good edge similarity; robust entropy-based weighting	Oversharpened output; dark PET color; poor global structural coherence	Good—lightweight architecture
DFENet (2023) [36]	Dual-branch CNN + ViT encoder-decoder; GSFAM; LEGM fusion strategy	MI: 0.9504; VIF: 0.6903; SSIM: 0.7910	N/R	N/R	10.228	Integrates local texture and global semantics; LEGM robust to noise	Very high inference time (~10 s/image); needs MS-COCO pretraining	Limited—~10 s/image impractical for real-time use
ECFusion (2025) [32]	Edge-Augmented Module (Sobel) + Cross-Scale Transformer (CSTF); multi-path fusion	MI: 4.5539; VIF: 0.8788; SSIM: 0.6487	0.92	14.56	0.013	Explicit edge preservation; avoids pixel-shift artifacts; best VIF on MRI-PET	Sobel maps limited for complex lesions	Moderate—efficient but no clinical latency benchmarks
MACT-Fusion (2025) [31]	Lightweight cross-Transformer; cross-window + cross-grid attention; SAFM	MI: 1.691; VIF: 0.571; SSIM: 0.931	0.571	~30–50	N/R	Fewest FLOPs; best MS-SSIM	Lower VIF than ECFusion; 3D fusion not addressed	Clinical brain tumor validation
MDC-RHT (2024) [44]	MDC (4-dimensional dynamic conv) + Residual Hybrid Transformer; overlapping cross-attention; multi-scale unsupervised	MI: 0.9381; VIF: 0.6111; SSIM: 0.5516	N/R	N/R	N/R	MDC captures rich contextual cues across all kernel dimensions; RHT overcomes shift-window blocking; strong color fidelity	High attention complexity limits deployment; interpretability needs work	Limited—authors flag attention complexity as barrier; pruning identified as future work

by training without supervision, though structural consistency suffers relative to MACTFusion.

ECFusion achieves the best VIF overall. Its Sobel-based edge module assumes that diagnostically relevant information tracks sharp intensity boundaries, an assumption that holds less well on low-contrast modalities such as PET or fMRI. MACTFusion takes the opposite approach, prioritizing efficiency, and records the highest SSIM in the table—though all evaluations have been conducted on brain tumor data, and generalization to other anatomical regions has not been tested.

The efficiency columns expose a comparability problem of a different kind. Four of the seven models report none of the three metrics. Among the remaining five, disclosure is uneven: KD^{3}MT provides all three figures (0.06 M parameters, 3.64 GFLOPs, 0.35 s); ECFusion reports parameters and FLOPs (0.92 M, 14.56 GFLOPs) but no inference time; EMFusion reports only parameters (0.30 M); DFENet gives only inference time (10.228 s); and MACTFusion lists parameters (0.571 M) and an approximated FLOPs range read from a chart. None of the papers specify a common hardware configuration, so even the figures that exist are not on equal footing. The near-800-fold gap between ECFusion's 13.4 ms and DFENet's 10.228 s is the starkest illustration of this, but the practical consequence is simpler: a reader cannot determine from the published record which of these models is computationally viable for a specific clinical setting. This is a field-wide gap, not a limitation of these seven papers in particular.

Only MACTFusion reports a clinical validation outcome—brain tumor classification. The other six are assessed on image-quality metrics without a corresponding diagnostic task.

4.6. Cross-model critical comparison

Table 6 benchmarks seven models on a shared MRI-PET evaluation. Several other models discussed in Sections 4.1 and 4.3 address related cross-modal fusion problems and report different evaluation outcomes, summarized here for comparison.

MultiViT and AdaFuse both use cross-attention but report different types of validation. MultiViT fuses fMRI functional network connectivity with sMRI gray matter maps via cross-attention and reports an AUC of 0.833 on schizophrenia classification, a disease-specific diagnostic outcome. AdaFuse applies cross-attention between spatial and frequency-domain features on CT-MRI and PET-MRI pairs, reporting performance using image-quality metrics (entropy, PSNR, mutual information, correlation coefficient, and feature mutual information) without an accompanying diagnostic or classification task. Of the two, only MultiViT's reported results connect fusion performance to a clinical outcome measure.

TUFusion and ECINFusion differ in their intended scope. TUFusion is described by its authors as a multi-domain fusion algorithm, with reported applicability spanning medical imaging, industrial inspection, security, and other domains, rather than being designed and evaluated exclusively for medical fusion. ECINFusion is evaluated specifically on CT-MRI and PET-MRI pairs using image-quality metrics, with its design centered on an explicit channel-wise interaction mechanism for cross-modal feature exchange. ECINFusion's evaluation is therefore narrower in domain scope but more directly targeted at the medical modality pairs this survey addresses.

Across these four models, only MultiViT reports a disease-specific classification outcome; AdaFuse and ECINFusion report image-quality metrics on medical modality pairs without a corresponding diagnostic task, and TUFusion's reported evaluation spans medical and non-medical domains rather than being medical-specific.

5. Evaluation Metrics

The quantification of image fusion models largely relies on the task type (i.e., structural preservation, segmentation, or classification) and on the interpretability needs of the medical domain. For applications centered around fusion, such as CT–MRI or PET–MRI, structural similarity (SSIM), peak signal-to-noise ratio (PSNR), and visual information fidelity (VIF) are the most widely used metrics. All models utilizing cross-attention and residual-guided fusion achieve consistently high SSIM values (>0.95) and perform well on other metrics. For segmentation-based models, especially for multi-organ CT or brain MRI data, Dice coefficient and Hausdorff distance (HD95) are typical choices. Fully transformer-based architectures also perform very well on the Dice (>83%) in the recent multi-organ benchmarks, alongside low HD95, indicating more accurate localization of anatomical boundaries. Classification-based approaches utilizing decision-level fusion strategies employ accuracy, AUC, or sometimes an attention entropy for explainability.

These models frequently produce highly stable diagnostic predictions and leverage transformer heads that localize predictive attention to clinically relevant areas. Models that incorporate adversarial or GAN-based modules also use Fréchet Inception Distance (FID) as a realism metric, especially when the fused image feeds into a human-in-the-loop process. The most widely used metrics are presented in Table 7.

Table 7
Common evaluation metrics used in transformer-based medical image fusion, detailing their purpose and typical clinical use cases

Metric	Purpose/Focus	Example use cases
SSIM	Structural preservation	CT–MRI fusion, PET–MRI alignment
PSNR	Signal reconstruction/noise reduction	Cross-modality image enhancement
VIF	Perceptual fidelity based on HVS	Low-contrast organ fusion
MI	Information retention	PET–fMRI fusion, entropy matching
Dice Score	Mask overlap accuracy	Brain tumor, liver segmentation
HD95	Boundary precision	Vascular lesion and stroke mapping
AUC/Accuracy	Classification performance	Alzheimer's disease staging
FID	Realism of generated outputs	Transformer–GAN fused reconstruction
Attention Entropy	Interpretability & attention focus	ViT-based fusion visualization

Table 8
Common benchmark datasets for multimodal medical image fusion, detailing modalities, task types, and clinical focus areas

Dataset	Modalities	Task type	Clinical domain/Focus
BRATS, [46]	T1, T2, FLAIR, T1ce	Tumor segmentation	Brain glioma detection and sub-region labeling
ADNI [47]	PET, MRI	Cognitive classification	Alzheimer's progression and dementia diagnosis
CHAOS [48]	T1/T2 MRI, CT	Liver segmentation	Multi-organ abdominal analysis
OASIS [49]	Structural MRI	Neuro-classification	Age-related cognitive decline and dementia studies
ISLES [50]	DWI, PWI	Stroke lesion mapping	Ischemic core and penumbra segmentation
TCIA [51]	CT, MRI, PET	Multi-task (fusion, detection)	Oncology-focused lesion and tissue mapping
IXI	T1, T2, PD MRI, CT	Modality registration	Structural brain mapping and anatomical alignment
EyePACS [52]	Fundus RGB images	Disease detection	Diabetic retinopathy screening and grading

5.1. Benchmark datasets

To maintain reliable evaluation in medical fusion, studies should use standardized, high-quality datasets that support a wide variety of imaging modalities and diagnostic domains. These datasets will necessarily be specific to clinical endpoints—whether tumor segmentation, tracking neurodegeneration, or detecting retinal disease—dictating model architecture choices at the point of fusion. For instance, in neuro-oncology and glioma detection, the datasets are commonly composed of four MRI modalities (FLAIR, T1, T1ce, and T2), allowing for multi-resolution feature extraction across sequences. Statistical measures such as the mean and standard deviation for both tumor ore and edema indicate that transformer models trained on these samples consistently yield better performance across Dice and mIoU metrics than their convolutional counterparts. PET–MRI linked datasets are widely used in research on cognitive decline in the elderly.

At the disease level, these datasets support decision-level fusion for diagnosis across progression stages, typically incorporating attention pooling and class-token embeddings. Retinal disease screening, by contrast, relies on specialized retinal imaging datasets, while abdominal organ analysis depends on datasets specifically designed to account for spatial constraints.

Here, either VIF or MS-SSIM is used for evaluation, with the VIF measuring structural similarity retention and the latter measuring perceived realism. Table 8 outlines widely used benchmark datasets across a variety of clinical domains, paired with their associated modalities and target tasks.

These datasets shape the transformer design in significant ways. High-resolution multi-organ segmentation tasks typically leverage full-depth encoder-decoder architectures, while cross-modality or low-resource tasks benefit from hybrid CNN-transformer variants with adaptive token refinement.

5.1.1. Data resources

Publicly available medical imaging datasets commonly referenced in the literature for benchmarking and evaluation purposes include BRATS, ADNI, OASIS, ISLES, TCIA, IXI, CHAOS, and EyePACS. Access to these datasets is subject to the respective terms and conditions of the original data providers. MRI data for age-related cognitive decline analysis were obtained from the Open Access Series of Imaging Studies (OASIS). Stroke lesion segmentation data were sourced from the ISLES (Ischemic Stroke Lesion Segmentation) challenge dataset. Multimodal oncological imaging data, including CT, MRI, and PET scans, were accessed via the Cancer Imaging Archive (TCIA). Additional structural brain MRI datasets were obtained from the IXI repository. Abdominal organ segmentation experiments used the CHAOS challenge dataset, while retinal fundus images for diabetic retinopathy screening were obtained from the EyePACS dataset.

6. Challenges, Limitations, and Open Research Problems

Despite the remarkable progress in transformer-based multimodal fusion for healthcare, several technical and translational hurdles remain. These challenges span across data availability, modality heterogeneity, computational scalability, clinical trust, and fusion granularity design. In this section, these issues are categorized and discuss emerging research directions.

6.1. Data scarcity and label imbalance

One of the most pressing challenges in medical fusion is the lack of larger, well-annotated, paired inter-modality datasets, except for a few. This contrasts with natural image tasks, where aligned pairs with pixel-level labels can be collected on demand; medical paired datasets require considerable efforts and expenses to acquire and are typically restricted to small patient cohorts. As a result, models are often trained on small cohorts, which restricts generalization. To bridge this gap, recent architectures have proposed either self-supervised pretraining or semi-paired fusion models.

For instance, some methods only need pixel correspondence to the coarse level (not pixel-wise), and they simulate the consistency between different modalities through synthetic alignment. Unfortunately, these approaches could end up learning domain-specific, artifact-based, or confounding-based structures rather than actually the clinically relevant structures. Recent reviews also stress that although deep learning has advanced rapidly, the vast majority of fusion pipelines have not been designed originally for medical data, which exacerbates the lack of available data, in turn limiting the robustness across disease populations [53].

To address this, a number of optimization-driven strategies have been presented for such problems; in particular, hybrid evolutionary approaches can dynamically optimize fusion weights

and improve convergence speed so that sensitivity to dataset size is minimized while preserving its quality [54]. Class imbalance is also a common phenomenon in medical datasets, particularly for disease staging or rare condition detection, and can bias fusion attention towards majority classes. This is especially problematic for decision-level fusion, where final predictions may be skewed by outputs from the more heavily represented class.

6.2. Modality heterogeneity and misalignment

Medical modalities differ in more than resolution or contrast. MRI captures soft-tissue structure, PET reflects metabolic function, and fundus imaging reveals surface vascularity. Each encodes a different kind of information, and designing transformers that correlate these channels coherently is still unsolved.

Semantic mismatch causes noisy attention weights even when inputs are spatially aligned and incoherent feature representations even when features share a spatial layout. Models with fixed positional encoding or uniform fusion blocks handle this poorly. Modality-specific encoders and hybrid attention layers have been proposed [54–56], but the evidence suggests they manage the problem rather than resolve it—cross-modal integration remains noisy when the inputs encode phenomena with little shared structure. Multi-scale and frequency decomposition-based attention reduce that noise to a degree, yet semantically coherent fusion across modalities remains an open question.

6.3. Interpretability and clinical trust

Transformers learn long-range dependencies well, but attention is not inherently interpretable to clinicians. Attention heatmaps and class-token relevance maps visualize what the model attended to; they do not explain whether that attention reflects clinically sound reasoning. Few existing methods incorporate medical ontology or domain-specific priors into their interpretability pipelines, which leaves this gap largely unaddressed. Saliency-based fusion methods narrow it somewhat—they identify regions of visual relevance—but cannot produce findings a clinician could act on [56]. For decisions such as cancer grading or surgical planning, that is a meaningful limitation. Radiologist-in-the-loop systems that co-design model outputs and causality-aware transformers that reason about why certain regions were fused or ignored are two directions that could address it more directly.

6.4. Computational burden and clinical deployability

Many transformer-based fusion models, specifically the ones with deep ViT backbones (a type of transformer model) or large token embeddings, are both memory- and compute-expensive. This becomes a significant bottleneck for clinical tools that aim to work in real-time or can be deployed on edge devices. In recent architectures, token pruning, patch coarsening, or depth reduction of transformer blocks is employed. On the other hand, overly aggressive pruning may sacrifice spatial fidelity, which is critical in segmentation tasks. For clinical adoption, transformer-based models need to advance in the direction of parameter efficiency, design that facilitates quantization, and runtime-agnostic backbones that can be run on MRI consoles or mobile diagnostic units.

6.5. Granularity-aware fusion design

Another drawback is the choice to select the fusion level, among pixel-, feature-, and decision-level fusion. For the vast majority of existing models, fusion is either hardcoded very early (pixel-level) or very late (decision-level) without any dynamic adaptation when it comes to the relevance of the modality and/or the specific demands of the task. While there is a lot of room for improvement, one of the promising directions would be to adaptively design fusion that dynamically determines how deep to fuse modalities. Fusing low-resolution anatomical features with latent semantic concepts that is only later fused (hybrid) is the direction some hybrid architectures are beginning to explore via multi-scale attention routing. This level of adaptivity could be embedded within transformers, perhaps driven by reinforcement learning or meta-controllers, and stands as a frontier that might bridge the need for clinical granularity with technical depth.

6.6. Generalization across centers and protocols

Clinical images also show high variability based on hospital, scanner hardware, acquisition protocol, and patient population. As a result, transformer-based fusion models trained on one dataset perform poorly on holdout cohorts with differences to learned joint representations, leading to significant degradation. While such work has already considered domain adaptation losses and style transfer, these mostly are superficial solutions that do not penetrate the core issue of cross-center heterogeneity. Privacy constraints add a further complication: training data cannot always be pooled across institutions, which limits how well fusion models generalize to new clinical sites. Work on telehealth security and medical data privacy—including watermarking-based fusion frameworks [57]—shows that cross-center generalization and secure data handling are not separable problems. Privacy-preserving transformer fusion is early-stage work, and how to bring federated learning, domain adaptation, and security-aware design together into a deployable clinical system remains unresolved.

6.7. Real-world deployment and clinical translation barriers

Most transformer-based fusion work reviewed here has not moved past architectural experimentation and benchmark validation. Regulatory clearance, CE marking, and commercial deployment involve a different set of constraints entirely, and none of the surveyed methods are near that stage. That is not a criticism of the work—it reflects where the field is. DFENet takes around 10 s per image pair (§4.5.5). MATR accumulates overhead as spatial scale increases (§4.5.2). These figures are acceptable for research prototypes but would need to fall considerably before intraoperative use becomes realistic—and that holds for models with strong MI, SSIM, or VIF scores as much as for weaker ones. Clinical validation remains the larger problem. Of the seven models in Table 6, only MACTFusion reports a clinical outcome, and that is confined to brain tumor classification at a single site. Multi-site, multi-scanner evaluation has not been done. Generalization claims, at this point, rest on benchmark performance—which is a limited basis from which to draw conclusions about deployment readiness.

Three factors likely shape this timeline. Regulatory pathways are one: the image-quality metrics anchoring the benchmark comparisons in Section 4 are useful, but a Software as a

Medical Device pathway asks for evidence of clinical benefit alongside them. Technical readiness is another—closing the latency gap seen in models like DFENet and testing across multiple scanners, vendors, and acquisition protocols would address the generalization question already raised in §6.6. Interpretability is the third. A fused image informs rather than replaces clinical judgment, and transformer attention weighting can be difficult to audit (§6.3), so clinical trust will likely track progress on explainability about as closely as it tracks anything else. None of this diminishes the benchmark progress surveyed in Section 4; it just marks the distance still ahead of it.

7. Future Directions and Emerging Trends

The transformer-based multimodal fusion landscape is evolving, and several trends are beginning to emerge that could be transformative. These trends aim to move beyond current limitations and enhance generalizability, as well as clinical applicability, by rethinking training, deployment, and interpretability of transformers. This section outlines some of the most exciting avenues.

Integration of language-guided vision transformers for multimodal medical reasoning is one potential future direction. Models learn richer representations of a patient by learning from imaging unified with clinical text such as radiology reports, enabling prompt-based fusion in which the transformer grounds attention not only in pixel- or feature-level similarity but also in semantic context. SMFusion [58] demonstrates this approach by incorporating medical prior knowledge into the fusion process, encoding text descriptions generated by a biomedical language model and aligning them with image features through a semantic interaction module, offering greater interpretability and aligning more closely with how clinicians connect image findings with text.

Because transformer models are generally resource-intensive, a parallel trend has emerged around efficient ViTs for rapid diagnostics in point-of-care or portable imaging devices. Researchers are designing wavelet-integrated vision transformers and compressed attention heads to reduce parameter count and computational overhead. WaveFusion [56] illustrates this direction, combining an adaptive wavelet transform module with parallel self-attention and convolutional processing of low- and high-frequency components to balance computational cost against fusion quality. Such models are particularly relevant to applications such as diabetic retinopathy grading on fundus images.

Most existing models rely heavily on fully supervised multimodal data, with few-shot and self-supervised fusion methods gaining traction. Such methods align modalities with regularizers based on structural priors and/or contrastive loss functions or even cycle consistency without needing dense annotation. Multi-stage architectures that merge common latent spaces with cross-modal transformers are being investigated to carry out consistent fusion in the presence of a noisy, missing, or unseen modality.

This has significant implications for rare diseases, low-resource hospitals, or studies where only a few modalities overlap. One major area of progression is to go beyond 2D slice-wise fusion into 3D volumetric and 4D temporal transformers, especially for MRI sequences, longitudinal PET scans, or video-based imaging [59]. These models address temporal consistency, anatomical drift, and spatial continuity in time, qualities that are particularly important for cancer surveillance, stroke evolution, or fetal imaging, for instance. Other experimental models use recurrent transformer layers and temporal token linkers, connecting attention over time while fusing across modalities. While these designs are still nascent, they show promise for cross-session, cross-modality, and patient-specific modeling. Federated fusion learning has emerged as a privacy-first alternative to centralized training, enabling transformer-based models to be trained at scale across hospitals and imaging centers.

Thus, transformer weights are updated locally, and only attention statistics (or gradients) are shared, ensuring patient privacy in this setup. PPPML-HMI [60] explores designing attention-aware aggregation rules and modality-specific privacy masks to allow transformers to comply with HIPAA and GDPR requirements, making them suitable for clinical deployment. This is essential to ensure that models can be deployed at scale while maintaining ethical and legal compliance. Eventually, fusion models are also attracting interest in terms of clinical explainability by integrating causal reasoning and medical ontologies into attention pipelines.

This means getting not only a heatmap of where the model attended but also why connecting fused features with corresponding anatomical and/or diagnostic rationale. Initial versions of this work have been applied to the fusion problem, not only to ensure accurate fusion via contrastive attribution, but also to ensure grounded accountability via knowledge from the underlying ontology—in this case, as causal graphs that underpin ontology reasoning. These innovations aim to transition fusion from black-box to interpretable clinical synergism.

8. Conclusion

Transformer-based multimodal image fusion in medical imaging has recently been introduced, enabling the fusion of diagnostic modalities including structural MRI, metabolic PET, and fundus photography. Compared to classical CNN-based approaches, transformer architectures offer specific clinical advantages: cross-attention mechanisms enable direct querying between modalities for tasks such as PET-MRI co-registration in oncology, where CNNs' fixed receptive fields struggle to model long-range spatial dependencies; edge-augmented transformer modules improve boundary delineation in brain tumor segmentation by preserving structural detail across modality transitions; and decision-level transformer fusion supports modular interpretability in tasks such as Alzheimer's disease staging, allowing clinicians to audit the contribution of each modality individually.

This survey examines transformer adaptations for pixel-level, feature-level, and decision-level fusion across clinical tasks ranging from tumor segmentation to cognitive decline prediction as individual clinical purposes in comparison with state-of-the-art architectures, benchmark datasets, and performance metrics, which not only showcase current methods' capabilities but also their limitations. Aside from these comparisons, the survey also covered available open challenges over generalizability, deployment, and trustworthiness and summarized some of their most important practical bottlenecks, including scarce data availability, semantic mismatch, and clinical interpretability.

Transformer architectures are evolving towards more flexible, interpretable, and computationally efficient design. Promising directions for clinical AI include language-guided prompting, lightweight ViT variants, 3D and temporal fusion, and federated learning. Equally important is the transition toward fusion models that produce not only accurate results but also outputs consistent with medical and ethical reasoning. Transformer-based image

fusion thus represents a vital link between unrefined diagnostic data and clinical decision-making.

Acknowledgement

Dr. Sivaneasan contributed to the conceptualization and critical review of this manuscript but was not listed as an author at his own request.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest in this work.

Data Availability Statement

Data sharing is not applicable to this article, as no new data were created or analyzed in this study. The benchmark datasets discussed in this survey (BRATS, ADNI, CHAOS, OASIS, ISLES, TCIA, IXI, and EyePACS) are third-party resources, described in their original publications and listed in Section 5.1.1 and Table 8. Most are publicly accessible; a few, including ADNI and TCIA, require researchers to register with the data provider and sign a data use agreement before access is granted. Access to all datasets is subject to the terms set by their respective providers; readers should consult the original sources cited here for current access procedures.

Author Contribution Statement

Nalini S Jagtap: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Dilip Kumar Jang Bahadur Saini:** Conceptualization, Methodology, Software, Validation, Formal analysis, Resources, Writing – review & editing, Supervision. **Prasun Chakrabarti:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Supervision.

References

[1] Zhang, Y., Yin, J., Gu, Y., & Chen, Y. (2025). Multi-level feature attention network for medical image segmentation. *Expert Systems with Applications*, *263*, 125785. https://doi.org/10.1016/j.eswa.2024.125785

[2] Zhang, Y., Sun, K., Liu, Y., & Shen, D. (2023). Transformer-based multimodal fusion for early diagnosis of Alzheimer's disease using structural MRI and PET. In *2023 IEEE 20th International Symposium on Biomedical Imaging*, 1–5. https://doi.org/10.1109/ISBI53787.2023.10230577

[3] Zhang, J., Liu, A., Wang, D., Liu, Y., Wang, Z. J., & Chen, X. (2022). Transformer-based end-to-end anatomical and functional image fusion. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–11. https://doi.org/10.1109/TIM.2022.3200426

[4] Xia, K., & Wang, J. (2023). Recent advances of transformers in medical image analysis: A comprehensive review. *MedComm—Future Medicine*, *2*(1), e38. https://doi.org/10.1002/mef2.38

[5] Wang, Y., Luo, Y., Zu, C., Zhan, B., Jiao, Z., Wu, X., . . . , & Zhou, L. (2024). 3D multi-modality Transformer-GAN for high-quality PET reconstruction. *Medical Image Analysis*, *91*, 102983. https://doi.org/10.1016/j.media.2023.102983

[6] Yan, S., Yang, B., Chen, A., Zhao, X., & Zhang, S. (2025). Multi-scale convolutional attention frequency-enhanced transformer network for medical image segmentation. *Information Fusion*, *119*, 103019. https://doi.org/10.1016/j.inffus.2025.103019

[7] Pathak, D. M., & Tewari, T. K. (2025). M3IFT2.0: Multi-modal medical image fusion framework with vision transformer features and attention integration. *International Journal of Information Technology*, *17*(5), 2905–2918. https://doi.org/10.1007/s41870-025-02488-y

[8] Lyu, J., Chen, X., AlQahtani, S. A., & Hossain, M. S. (2024). Multi-modality MRI fusion with patch complementary pre-training for internet of medical things-based smart healthcare. *Information Fusion*, *107*, 102342. https://doi.org/10.1016/j.inffus.2024.102342

[9] Yao, Z., Xie, W., Chen, J., Zhan, Y., Wu, X., Dai, Y., . . . , & Zhang, G. (2025). IT: An interpretable transformer model for Alzheimer's disease prediction based on PET/MR images. *NeuroImage*, *311*, 121210. https://doi.org/10.1016/j.neuroimage.2025.121210

[10] Liu, J., Li, Y., Sun, X., Wang, X., & Luo, H. (2025). MMIF: Multimodal medical image fusion network based on multi-scale hybrid attention. *Computers, Materials & Continua*, *85*(2), 3551–3568. https://doi.org/10.32604/cmc.2025.066864

[11] Oubelkas, F., Benhaili, Z., Moumoun, L., & Jamali, A. (2025). MOSAIC: A multimodal self-supervised transformer for medical image diagnosis. In *Advances in Intelligent Systems and Digital Applications: Proceedings of ISDA 2025*, 46–55. https://doi.org/10.1007/978-3-031-95330-9_5

[12] Wang, Y., Wang, L., Huang, Z., Zhang, Y., & Han, Y. (2025). Multimodal medical image fusion method based on the swin transformer and self-supervised contrast learning. *Recent Advances in Electrical & Electronic Engineering*, *18*(7), e230724232192. https://doi.org/10.2174/0123520965311272240709053422

[13] Kamara, A. A., He, S., & Joseph Fofanah, A. (2026). FAMAFuse: Functional-anatomical multiscale attention for multimodal image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, *36*(3), 3215–3230. https://doi.org/10.1109/TCSVT.2025.3626562

[14] Gao, X., Shi, F., Shen, D., & Liu, M. (2023). Multimodal transformer network for incomplete image generation and diagnosis of Alzheimer's disease. *Computerized Medical Imaging and Graphics*, *110*, 102303. https://doi.org/10.1016/j.compmedimag.2023.102303

[15] Zeng, P., Zeng, X., Wang, Y., Zhou, L., Zu, C., Wu, X., . . . , & Shen, D. (2025). Multi-modal long-short distance attention-based transformer-GAN for PET reconstruction with auxiliary MRI. *IEEE Transactions on Circuits and Systems for Video Technology*, *35*(8), 7835–7849. https://doi.org/10.1109/TCSVT.2025.3545911

[16] Peng, P., & Luo, Y. (2025). Multimodal medical image fusion using a progressive parallel strategy based on deep learning. *Electronics*, *14*(11), 2266. https://doi.org/10.3390/electronics14112266

[17] Javed, U., Riaz, M. M., Ghafoor, A., Ali, S. S., & Cheema, T. A. (2014). MRI and PET image fusion using fuzzy logic

and image local features. *The Scientific World Journal*, *2014*(1), 708075. https://doi.org/10.1155/2014/708075

[18] Xia, K., Yin, H., & Wang, J. (2019). A novel improved deep convolutional neural network model for medical image fusion. *Cluster Computing*, *22*(1), 1515–1527. https://doi.org/10.1007/s10586-018-2026-1

[19] Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., & Zhang, L. (2020). IFCNN: A general image fusion framework based on convolutional neural network. *Information Fusion*, *54*, 99–118. https://doi.org/10.1016/j.inffus.2019.07.011

[20] Tan, W., Tiwari, P., Pandey, H. M., Moreira, C., & Jaiswal, A. K. (2025). Multimodal medical image fusion algorithm in the era of big data. *Neural Computing and Applications*, *37*(28), 22995–23015. https://doi.org/10.1007/s00521-020-05173-2

[21] Fu, J., Li, W., Du, J., & Xu, L. (2021). DSAGAN: A generative adversarial network based on dual-stream attention mechanism for anatomical and functional image fusion. *Information Sciences*, *576*, 484–506. https://doi.org/10.1016/j.ins.2021.06.083

[22] Xu, H., & Ma, J. (2021). EMFusion: An unsupervised enhanced medical image fusion network. *Information Fusion*, *76*, 177–186. https://doi.org/10.1016/j.inffus.2021.06.001

[23] Song, J., Zheng, J., Li, P., Lu, X., Zhu, G., & Shen, P. (2021). An effective multimodal image fusion method using MRI and PET for Alzheimer's disease diagnosis. *Frontiers in Digital Health*, *3*, 637386. https://doi.org/10.3389/fdgth.2021.637386

[24] Wang, J., Yu, L., & Tian, S. (2022). MsRAN: A multi-scale residual attention network for multi-model image fusion. *Medical & Biological Engineering & Computing*, *60*(12), 3615–3634. https://doi.org/10.1007/s11517-022-02690-1

[25] Tang, W., He, F., Liu, Y., & Duan, Y. (2022). MATR: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Transactions on Image Processing*, *31*, 5134–5149. https://doi.org/10.1109/TIP.2022.3193288

[26] Li, B., Hwang, J.-N., Liu, Z., Li, C., & Wang, Z. (2022). PET and MRI image fusion based on a dense convolutional network with dual attention. *Computers in Biology and Medicine*, *151*, 106339. https://doi.org/10.1016/j.compbiomed.2022.106339

[27] Lakshmi, A., Rajasekaran, M. P., Jeevitha, S., & Selvendran, S. (2022). An adaptive MRI-PET image fusion model based on deep residual learning and self-adaptive total variation. *Arabian Journal for Science and Engineering*, *47*(8), 10025–10042. https://doi.org/10.1007/s13369-020-05201-2

[28] Panigrahy, C., Seal, A., Gonzalo-Martín, C., Pathak, P., & Jalal, A. S. (2023). Parameter adaptive unit-linking pulse coupled neural network based MRI–PET/SPECT image fusion. *Biomedical Signal Processing and Control*, *83*, 104659. https://doi.org/10.1016/j.bspc.2023.104659

[29] Xie, X., Zhang, X., Ye, S., Xiong, D., Ouyang, L., Yang, B., & Wan, Y. (2023). MRSCFusion: Joint residual swin transformer and multiscale CNN for unsupervised multimodal medical image fusion. *IEEE Transactions on Instrumentation and Measurement*, *72*, 5026917. https://doi.org/10.1109/TIM.2023.3317470

[30] Miao, S., Xu, Q., Li, W., Yang, C., Sheng, B., Liu, F., . . . , & Yu, X. (2024). MMTFN: Multi-modal multi-scale transformer fusion network for Alzheimer's disease diagnosis. *International Journal of Imaging Systems and Technology*, *34*(1), e22970. https://doi.org/10.1002/ima.22970

[31] Xie, X., Zhang, X., Tang, X., Zhao, J., Xiong, D., Ouyang, L., . . . , & Teo, K. L. (2025). MACTFusion: Lightweight cross transformer for adaptive multimodal medical image fusion. *IEEE Journal of Biomedical and Health Informatics*, *29*(5), 3317–3328. https://doi.org/10.1109/JBHI.2024.3391620

[32] Luo, F., Wu, D., Pino, L. R., & Ding, W. (2025). A novel multimodel medical image fusion framework with edge enhancement and cross-scale transformer. *Scientific Reports*, *15*(1), 11657. https://doi.org/10.1038/s41598-025-93616-y

[33] Bi, Y., Abrol, A., Fu, Z., & Calhoun, V. D. (2024). A multimodal vision transformer for interpretable fusion of functional and structural neuroimaging data. *Human Brain Mapping*, *45*(17), e26783. https://doi.org/10.1002/hbm.26783

[34] Li, M., Fang, Y., Shao, J., Jiang, Y., Xu, G., Cui, X., & Wu, X. (2025). Vision transformer-based multimodal fusion network for classification of tumor malignancy on breast ultrasound: A retrospective multicenter study. *International Journal of Medical Informatics*, *196*, 105793. https://doi.org/10.1016/j.ijmedinf.2025.105793

[35] Odusami, M., Maskeliūnas, R., & Damaševičius, R. (2023). Pixel-level fusion approach with vision transformer for early detection of Alzheimer's disease. *Electronics*, *12*(5), 1218. https://doi.org/10.3390/electronics12051218

[36] Li, W., Zhang, Y., Wang, G., Huang, Y., & Li, R. (2023). DFENet: A dual-branch feature enhanced network integrating transformers and convolutional feature learning for multimodal medical image fusion. *Biomedical Signal Processing and Control*, *80*, 104402. https://doi.org/10.1016/j.bspc.2022.104402

[37] Raminedi, S., Shridevi, S., & Won, D. (2024). Multi-modal transformer architecture for medical image analysis and automated report generation. *Scientific Reports*, *14*(1), 19281. https://doi.org/10.1038/s41598-024-69981-5

[38] Zhao, Y., Zheng, Q., Zhu, P., Zhang, X., & Ma, W. (2024). TUFusion: A transformer-based universal fusion algorithm for multimodal images. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(3), 1712–1725. https://doi.org/10.1109/TCSVT.2023.3296745

[39] Wei, X., Qiu, Y., Xu, J., & Zhang, J. (2026). HCF-MaNet: A novel holistic cross-modal fusion mamba network for multi-modal medical image fusion. *IEEE Transactions on Multimedia*, *28*, 4547–4561. https://doi.org/10.1109/TMM.2026.3660130

[40] Rezaee, K., & Zadeh, H. G. (2024). Self-attention transformer unit-based deep learning framework for skin lesions classification in smart healthcare. *Discover Applied Sciences*, *6*(1), 3. https://doi.org/10.1007/s42452-024-05655-1

[41] Dhar, J., Zaidi, N., Haghighat, M., Roy, S., Goyal, P., Alavi, A., & Kumar, V. (2025). Multimodal fusion learning with dual attention for medical imaging. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision*, 4362–4371. https://doi.org/10.1109/WACV61041.2025.00428

[42] Gu, X., Wang, L., Deng, Z., Cao, Y., Huang, X., & Zhu, Y. (2023). *AdaFuse: Adaptive medical image fusion based on spatial-frequential cross attention*. *arXiv*. https://doi.org/10.48550/arXiv.2310.05462

[43] Wei, X., Qiu, Y., Xu, X., Xu, J., Mei, J., & Zhang, J. (2025). ECINFusion: A novel explicit channel-wise interaction network for unified multi-modal medical image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, *35*(5), 4011–4025. https://doi.org/10.1109/TCSVT.2024.3516705

[44] Wang, W., He, J., Liu, H., & Yuan, W. (2024). MDC-RHT: Multi-modal medical image fusion via multi-dimensional

dynamic convolution and residual hybrid transformer. *Sensors*, *24*(13), 4056. https://doi.org/10.3390/s24134056

[45] Ding, Z., Liu, Y., Liu, S., He, K., & Zhou, D. (2025). KD^{3}mt: Knowledge distillation-driven dynamic mixer transformer for medical image fusion. *The Visual Computer*, *41*(6), 3679–3693. https://doi.org/10.1007/s00371-024-03627-5

[46] Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., . . . , & Davatzikos, C. (2017). Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, *4*(1), 170117. https://doi.org/10.1038/sdata.2017.117

[47] Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., . . . , & Weiner, M. W. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, *27*(4), 685–691. https://doi.org/10.1002/jmri.21049

[48] Kavur, A. E., Gezer, N. S., Barış, M., Aslan, S., Conze, P.-H., Groza, V., . . . , & Selver, M. A. (2021). CHAOS Challenge—combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, *69*, 101950. https://doi.org/10.1016/j.media.2020.101950

[49] Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, *19*(9), 1498–1507. https://doi.org/10.1162/jocn.2007.19.9.1498

[50] Hernandez Petzsche, M. R., de la Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., . . . , & Kirschke, J. S. (2022). ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific Data*, *9*(1), 762. https://doi.org/10.1038/s41597-022-01875-5

[51] Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., . . . , & Prior, F. (2013). The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, *26*(6), 1045–1057. https://doi.org/10.1007/s10278-013-9622-7

[52] Cuadros, J., & Bresnick, G. (2009). EyePACS: An adaptable telemedicine system for diabetic retinopathy screening. *Journal of Diabetes Science and Technology*, *3*(3), 509–516. https://doi.org/10.1177/193229680900300315

[53] Kahol, A., & Bhatnagar, G. (2024). Deep learning-based multimodal medical image fusion. In A. K. Singh & S. Berretti (Eds.), *Data fusion techniques and applications for smart healthcare* (pp. 251–279). Academic Press. https://doi.org/10.1016/B978-0-44-313233-9.00017-5

[54] Ogbuanya, C. E., Obayi, A., Larabi-Marie-Sainte, S., Saad, A. O., & Berriche, L. (2025). A hybrid optimization approach for accelerated multimodal medical image fusion. *PLOS One*, *20*(7), e0324973. https://doi.org/10.1371/journal.pone.0324973

[55] Liu, X., Wang, Z., Gao, H., Li, X., Wang, L., & Miao, Q. (2024). HATF: Multi-modal feature learning for infrared and visible image fusion via hybrid attention transformer. *Remote Sensing*, *16*(5), 803. https://doi.org/10.3390/rs16050803

[56] Wang, Q., Li, Z., Zhang, S., Chi, N., & Dai, Q. (2025). WaveFusion: A novel wavelet vision transformer with saliency-guided enhancement for multimodal image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, *35*(8), 7526–7542. https://doi.org/10.1109/TCSVT.2025.3549459

[57] Singh, K. N., Singh, O. P., Singh, A. K., & Agrawal, A. K. (2024). WatMIF: Multimodal medical image fusion-based watermarking for telehealth applications. *Cognitive Computation*, *16*(4), 1947–1963. https://doi.org/10.1007/s12559-022-10040-4

[58] Xiang, H., Zhang, H., Cheng, Y., Quan, X., & Huang, W. (2026). SMFusion: Semantic-preserving fusion of multimodal medical images for enhanced clinical diagnosis. *IEEE Journal of Biomedical and Health Informatics*, *30*(7), 6072–6085. https://doi.org/10.1109/JBHI.2025.3649749

[59] Wang, Y.-R., Qu, L., Sheybani, N. D., Luo, X., Wang, J., Hawk, K. E., . . . , & Daldrup-Link, H. E. (2023). AI transformers for radiation dose reduction in serial whole-body PET scans. *Radiology: Artificial Intelligence*, *5*(3), e220246. https://doi.org/10.1148/ryai.220246

[60] Zhou, J., Zhou, L., Wang, D., Xu, X., Li, H., Chu, Y., . . . , & Gao, X. (2024). Personalized and privacy-preserving federated heterogeneous medical image analysis with PPPML-HMI. *Computers in Biology and Medicine*, *169*, 107861. https://doi.org/10.1016/j.compbiomed.2023.107861

How to Cite: Jagtap, N. S., Saini, D. K. J. B., & Chakrabarti, P. (2026). Transformer-Based Multimodal Image Fusion in Healthcare: A Comprehensive Survey of Architectures, Fusion Strategies, and Clinical Applications. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS62028331