Transformation of Image Fusion Technologies with Generative AI: Architectures, Metrics, and Clinical Impact

Aakash [1], Narender Kumar [2]
{aakashgoyal2805@gmail.com[1], narenderster@gjust.org [2]}

Dept. of Computer science and Engineering,
Guru Jambheshwar University of Science and Technology, Hisar, India[1,2]

Abstract. Fusing the complementary information in multiple imaging modalities is a fundamental problem in remote sensing, surveillance and medical imaging. Classical fusion approaches based on multi-resolution transforms, sparse representations and optimization provide interpretability but are not well suited to heterogeneous, high-dimensional data and application-specific requirements. Deep learning systematically provides data-driven feature extraction and end-to-end fusion, but may lead to deterministic results without perceptual realism and uncertainty prediction. Recently emerged Generative AI techniques, e.g., GANs, VAEs, attention-based Transformers now make it possible of probabilistic modeling and perceptual discrimination also long-range multimodal correlation. This review of image fusion from 2010–2025 classifies significant architecture families, reports on objective, perceptual, and clinical evaluation metrics, and explores the translational aspects associated with CT–MRI and PET–MRI fusion. Our review discusses strengths and limitations of paradigms, considering representative clinical case studies, as well as key research gaps related to interpretability, robust evaluation, clinical validation, and effective deployment, providing further research in advanced multimodal fusion. [1][2]

Keywords: Generative AI, Image Fusion, Generative Adversarial Networks, Multimodal Imaging, Deep Learning

1 Introduction

Image fusion is the process of generating a single image containing the predominant useful information of multiple input images obtained from different sensors or at different occasions. In the remote sensing field, the fused images can integrate spectral and spatial information to enhance target detection and land cover subsidy. When visibility is low and weather challenging, compressing visible and IR sensing leads a more comprehensive scene understanding for two type of UAV/UAS applications, namely surveillance or autonomy. For clinical imaging, fusion of structural (CT, MRI) and functional/metabolic (PET, SPECT, fMRI) modalities has proven to offer clinicians a more

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2362514

complete diagnostic context leading to better lesion localization/staging/treatment planning. The seminal survey by James and Dasharatha [1] reviews the classical objectives of medical image fusion and presents the types of modalities, as well as emphasizes on their variety driven by a large scale of applications. [1]

1.1 Evolution of Methods: From Rules to Generative Models

1.1.1 Traditional Transform and Optimization-based Methods

By multiplying the coefficients (or via manually designed rules or fusion operators), multiscale transforms (wavelet, contourlet in particular), pyramid decomposition and dictionary learning based sparse signal representation algorithms have offered systematic manners for decomposing image structure according to scales/bands. Though such class of methods being computationally feasible and mathematically sound, may not work well in the heuristics for new modality or difficult imaging condition. [1]

1.1.2 Deep Learning-Based Methods

Following the advent of deep neural networks, deep fusion operators learned directly from data using convolutional architectures and autoencoders were proposed that improved fine-details preservation and independent learning of features. A strong use of CNNs for fusion with many state-of-the-art hand-crafted approaches was demonstrated on standard benchmarks such as IFCNN. However, discriminative scores of ten do not directly induce uncertainty, although they yield deterministic decisions. [3]

1.1.3 Generative AI Methods

Novel methods produce realistic and semantically consistent curated images using GANs, VAEs and Transformer-like models. FusionGAN realness with adversarial loss while preserving mild modality cues; VAEs allows probablistic latent of representation for uncertainty modeling, and MMT learns global feature matching through the self-attention mechanism. These new models allow for higher quality images combined to more clinically relevant output, however they are still plagued by training instability, expensive computational cost and poor explainability 4.

1.1.4 Emerging Diffusion and Foundation Models

In the most recent advancements in fusion research, GANs have given way to diffusion-based probabilistic models and foundation-scale multimodal architectures. In diffuse models such as FuseDiff [19], multimodal inputs are repeatedly denoised to yield fused outputs of improved structural fidelity and fewer artifacts than adversarial approaches. Pre-trained foundation models[20] meanwhile demonstrate zero-shot generalization policy across imaging modalities and diagnostic tasks. Prompt-conditioned multimodal Transformers are also an extension of this paradigm, conditioning the fusion on textual or clinical cues, leading to a more explainable and controllable synthesis. Together, these trends signify the advent of Generative Foundation Fusion (GFF) as a middle

ground to more effectively unify paths for privileged multimodal trustable, cost-effective and clinically actionable fusion. [21]

1.2 MOTIVATION AND CONTRIBUTION

While classical transform- and optimization-based methods provided interpretability and mathematical rigor, they struggle with heterogeneous, high-dimensional, and task-specific data. Deep learning architectures improved adaptability but remained deterministic, lacking uncertainty estimation and perceptual fidelity. Generative AI—spanning GANs, VAEs, Transformers, and diffusion-based models—now enables probabilistic, semantic, and human-aligned fusion frameworks. Yet, issues of interpretability, reproducibility, and clinical validation remain unresolved.

• A comprehensive survey categorizing image fusion methods, from multiresolution transforms to generative AI architectures (2010–2025).

• A study of latest trends and architectural styles with an emphasis on constraints and development.

• The value of performing a multisided and multidisciplinary assessment is recommended.

• A comprehensive review of published case reports that illustrate the value of CT-MRI and PET-MRI fusion in diagnosis and treatment planning.

• A research roadmap that details the deployment strategies and obstacles around robust evaluation, interpretability, and reproducibility that remain open challenges.

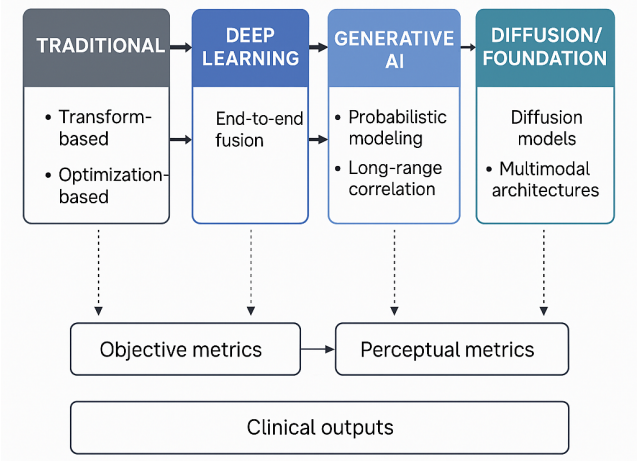


Fig. 1. Evolution of Image Fusion Paradigms

Figure 1 Illustrates the evolution of image fusion paradigms from rule-based to generative frameworks. Early transform and optimization methods (e.g., wavelet, PCA) offered interpretability but lacked adaptability [1]. The rise of deep learning introduced data-driven feature extraction with CNNs and autoencoders, improving fusion quality [3]. Generative AI models like FusionGAN [4]

and VAEs [9] added perceptual realism and uncertainty modeling, while Transformers enhanced cross-modal alignment through attention mechanisms [10]. Recent diffusion and foundation models (e.g., FuseDiff, MedFuse) [19][23] achieve explainable, zero-shot fusion across modalities. Overall, the figure highlights the shift from deterministic designs to probabilistic and interpretable generative architectures.

2 RELATED STUDIES

This section highlights various articles, algorithmic patterns, and assessment methodologies while synthesizing the state-of-the-art across three primary paradigms.

2.1 Traditional Image Fusion Methods

2.1.1 Multiscale Transform Methods

The original images are decomposed into scales and orientations by the multiscale transform, a few selected of which are fused to obtain the fused image. The inherent capability of wavelet transforms for expressing local frequency information has made these transforms very popular in the last years. A class of shearlet and contourlet systems enhance directionality as well as anisotropy but retain the edge and contour information. A general pipeline for transform-based fusion is as follows: registration $\rightarrow$ decomposition $\rightarrow$ coefficient fusion rule (e.g., weighted averaging or maximum selection) $\rightarrow$ reconstruction. Due to the interpretability and computational efficiency, these methods are widely used in remote sensing and early data fusion medical applications. [1]

2.1.2 Sparse Representation and Dictionary Learning

Sparse-coding models treat image patches as sparse linear combinations of atoms in learnt or predetermined dictionaries; combined images are formed by adding the sparse codes across modalities. The benefit is strong edge and details being retained, while dictionary design becomes cumbersome and the computation cost increases. Optimization based fusion techniques (total variation, variational forms) incorporate priors explicitly to weigh spatial coherency against spectral fidelity which require manual manipulations.

2.2 Deep Learning-Based Fusion

2.2.1 CNN-based Fusion

CNNs learn deep feature ex- tractors from data and can be used to map original images directly into fused outputs. IFCNN (a general fusion framework on CNNs) show an evidence that the shallow- to-middle-layer CNN can learn pixel/feature-level fusion strategies and achieve better performance than many traditional baselines on infrared–visible images, multi- focus images.[3]

2.2.2 Autoencoder and U-Net Variants

Autoencoders and U Net style encoder-decoder networks have been proposed to generate latent representations in each modality followed by merging or decoding the combined latent for fusion. The encoder extracts modality specific features and the joint decoder reconstructs the fused image. These are stamns that they can be trained through supervised or unsupervised modes and support end-to- end training.

2.2.3 Unsupervised and Self-Supervised Approaches

Unsupervised frameworks (e.g., U2Fusion) that lever- age reconstruction and perceptual consistent losses to infer fusion without paired ground truth fused images have also been suggested by researchers. Un-supervised pipelines are desirable in medical imaging where the ground truth is a fused image.[8]

2.3 Generative AI in Image Fusion

The introduction of generative AI introduces powerful modeling primitives that make no bones about modelling data distributions or optimizing for perceptual realism. Here we give a brief account of the main families and some representative fusion works.

2.3.1 Generative Adversarial Networks (GANs) for Fusion

The dual nature of GANs train a generator and discriminator in an adversarial game; the generator's goal is to generate fused images that are perceptually indistinguishable from real inputs, whereas the discriminator provides feedback back to the generator about realistic texture, contrast and semantics. Fusion GAN is representative that utilizes adversarial learns for infrared–visible fusion, and it demonstrates superior performances on contrast and perception compared to conventional methods, as well as CNN-based solutions. [4][18]

2.3.2 Variational Autoencoders (VAEs) and Probabilistic Fusion

VAEs learn probabilistic latent representations, and model uncertainty through the latent distributions in a clear way which is appealing for clinical image fusion where knowing the uncertainty of the fused image could offer diagnostic confidence. Architecture: VAE based fusion frameworks encode both modalities into a shared latent distribution and pool the latents before decoding to produce the fused output. The principled basis in variational inference offers a theoretically grounded path to posterior estimation and sampling-based uncertainty quantification. [9]

2.3.3 Transformer-based and Cross-Attention Models

Transformers and attention mechanisms enable long range con textual modeling and flexible cross-modal alignment. The Vision Transformers (ViT) demonstrated a way to make pure attention

architectures work well on vision tasks when pretrained on enough data. To help fusions for challenging multimodal tasks (e.g., PET- MRI), cross-attention layers may be applied to match patches or maps of features between modalities. This results in exact spatial and semantic correspondence. Recently transformer-based fusion models have achieved good performance and are usually integrated with convolutional backbones to form the hybrid architecture. [10]

2.3.4 Diffusion Models for Multimodal Fusion

Emerging diffusion probabilistic models represent a major leap in generative fusion research. Rather than adversarial training, diffusion models learn to iteratively denoise multimodal inputs into fused representations with stable likelihood-based optimization. FuseDiff demonstrated that conditional diffusion achieves higher structural fidelity and fewer artifacts than GAN-based methods. FlexiD-Fuse extended this idea to flexible multimodal setups, supporting variable input combinations in medical imaging.[19][22]

These models naturally encode uncertainty and multi-modal correlation, making them promising for robust and clinically interpretable fusion.

2.3.5 Foundation and Prompt-Driven Models

Recent studies highlight the emergence of foundation models for medical image fusion that leverage large-scale pretraining and zero-shot transfer. MedFuse integrates multiple modalities through a unified transformer backbone, achieving domain-general fusion across CT, MRI, and PET. Prompt-driven multimodal transformers introduce text-conditional fusion, allowing human-guided control and explainability in generative synthesis.[23][24] Such architectures mark a transition toward Generative Foundation Fusion (GFF)—a new era emphasizing scalability, cross-domain alignment, and interpretability.

3 RESEARCH METHODOLOGY

This section describes the methodological framework used to compile and analyze the surveyed literature and to synthesize the comparative conclusions presented in this paper.

3.1 Systematic Literature Review Protocol

We adhered to a systematic review protocol inspired by existing SLR practices (search, screening, eligibility and inclusion). The initial search databases were IEEE Xplore, Elsevier ScienceDirect, Springer Link, PMC and arXiv. Search keywords were: “image fusion”, “multimodal fusion”, “GAN image fusion”, “VAE image fusion”, “Transformer 2.6 investigate medical transformer to other modalities fusion”,“diffusion model fusion,” “foundation model fusion,” and “prompt-driven fusion.”. The time frame favored works published between 2010 to 2025, covering the period of deep learning and recent generative successes.[13]

Table 1: Comparison of Image Fusion Approaches Across Time

Approach	Representative Techniques
Traditional Methods (1980s–2000s)	Wavelet Transform, Laplacian Pyramid,Contourlet Transform, PCA, Sparse Representation
Strengths	Computationally efficient; Well-studied mathematical foundations; Easy interpretability
Limitations	Sensitive to noise; Limited adaptability to multimodal and nonlinear features; Hand-crafted design limits scalability
Applications	Remote sensing, military surveillance, basic medical imaging
Deep Learning Methods (2010s)	CNNs, Autoencoders, Sparse Coding CNNs, Residual Networks
Strengths	End-to-end learning; Automatic feature extraction; Improved fusion quality
Limitations	Requires large labeled datasets; Training complexity; Limited generalization
Applications	Infrared–visible fusion, biomedical imaging, autonomous driving
Generative AI Methods (2020s–Present)	GANs, VAEs, Diffusion Models, Vision Transformers
Strengths	High perceptual realism; Captures complex nonlinear relationships; Semantic consistency; Can synthesize missing modalities
Limitations	Computationally expensive; Training instability (e.g., GAN mode collapse); Interpretability issues
Applications	Precision medicine, PET–MRI/CT–MRI fusion, multimodal robotics, smart healthcare

3.2 Inclusion and Exclusion Criteria

1) Inclusion criteria: peer-reviewed journal articles or high-quality conference papers (or well-cited arXiv preprints) with clear methodology on image fusion (traditional, deep learning, or generative AI); studies including quantitative evaluations; clinical case studies or application-oriented experiments for medical imaging.

2) Exclusion criteria: single modality enhancement without (not fusion) opinion, reported methodological and not evaluated manuscripts in terms of bare efficacy outputs from image processing tools only) Re- views/scientific papers.[14]

3.3 Data Extraction and Analysis

From each selected paper, we collected: bibliographic information (authors, year, venue), studied pair of modalities (e.g., CT–MRI), method's family (transform-based, CNN, GAN, VAE,

Transformer), architectural details (number of layers, loss terms used and fusion module adopted), datasets employed for evaluation purposes, evaluation metrics chosen by the authors and reported results (either numerical or qualitative). Papers were then organized by method family for comparative synthesis and by clinical application of the methods for translational analysis.[15]

3.4 Categorization Framework

We adopted a four-dimensional comparative frame work: Input modalities, Fusion architecture, Evaluation metrics, and Reported outcomes/impact. This mapping clarifies which methods were applied to which tasks and how performance was assessed. The taxonomy explicitly distinguishes (a) pixel-level fusion, (b) feature-level fusion, and (c) decision-level fusion where appropriate.[16]

3.5 Comparative and Meta-Analytic Considerations

Meta-analytic comparisons (aggregation of simple effect sizes and ranking) were carried out whenever there was a set amount of homogeneous quantitative results available in the form of same data sets and metrics across several methods. For the heterogenous experimental scenarios we used qualitative synthesis, with critical evaluation of methodological discrepancies (preprocessing, registration, metric definitions).

4 ARCHITECTURES - DESIGN PATTERNS AND REPRESENTATIVE IMPLEMENTATIONS

This section synthesizes architectural patterns across the surveyed works and provides representative examples illustrating how different families approach fusion.

4.1 Traditional Fusion Pipelines

The commonly-used processing detail flow is as fol- lows: source registration$\rightarrow$ multi-scale decomposition (e.g., wavelet/contourlet) $\rightarrow$ fusion at the coefficient level under hand-crafted selection rules $\rightarrow$ inverse transform. Another trend is sparse coding, where input patches are encoded by learning dictionaries and decoding through a combination of codes. Optimization pipelines are convexified energy minimizers sparsifying the data term by a regularization.

4.2 CNN-based Architectures

Shared building blocks are simple convolutional encoders of low depth for each modality (referred to in the following as shallow) where image features are concatenated and finally decoded within a fusion decoding stage. Some designs are adopted residual blocks to stabilize the training; other architectures are built by mixing up multiscale feature aggregation for detail and context. IFCNN represents a well accepted CNN based fusion framework, which has shown its robustness in various tasks. [3]

4.3 GAN-based Fusion Architectures

In GAN-based models for fusion, generator networks take as input concatenated (or separately encoded) modality features and output the resulting fused images; discriminators either distinguish between real/fake fused images or categorize them versus modality-specific ones. These losses are usually a combination of adversarial loss and some form of reconstruction, content preservation (e.g., SSIM), and sometimes perceptual losses based on pretrained feature extractors. They have been combined in the framework of FusionGAN (a well- known architecture being used for infrared-visible fusion). [4][5][7]

4.4 VAE-based and Probabilistic Fusion Architectures

VAE-based fusion pipelines encode each modality to latent distributions and the fuse latents with learned networks or probabilistic rules. The decoders take fused images as input; sampling on latents can be used to obtain multiple plausible fused decoding and provide uncertainty estimates which could be useful for clinical decision. Introducing the reparameterization trick made end-to-end training with gradient descent possible. [9]

4.5 Transformer and Cross-Attention Architectures

Transformer-based fusion relates patches directly across modalities using patch embeddings and self-/cross-attention; cross-attention layers can focus on aligning modality-specific cues. Hybrid architectures integrate convolutional encoders for local texture and attention modules for global alignment. Vision Transformers and the transformers family on which they are built upon present the base techniques to leverage in fusion works. [10]

4.6 Hybrid and Multi-Objective Architectures

Hybrid models combine multiple inductive biases: e.g., encoder-decoder + adversarial refinement + attention-based fusion module + clinical-task heads (e.g., segmentation). Multi-objective training includes weighted losses for pixel fidelity, structural similarity, adversarial realism, and task utility (e.g., lesion segmentation loss).

5 CLINICAL IMPACT — TRANSLATIONAL OPPORTUNITIES AND CHALLENGES

5.1 Diagnostic Enhancement Decision Support

Generative fusion methods can further increase lesion conspicuity, and facilitate the combination of structural to functional data (CT/MRI or PET/MRI) aiding in small/lower contrast lesions identification with improved localization confidence by radiologists. While questionnaires and case series indicate better diagnostic capacity with optimal multimodal integration, several reports are retrospective analyses with limited numbers of cases.

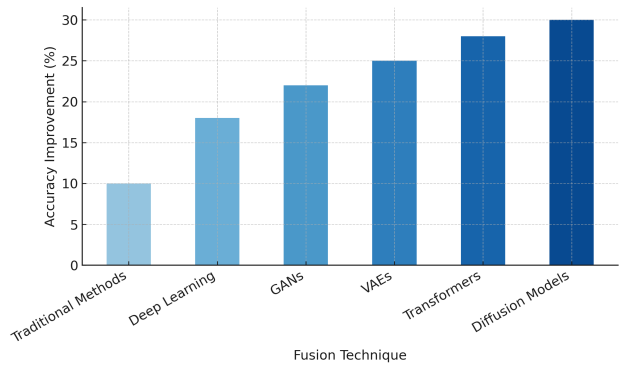


Fig. 2. Clinical Accuracy Improvement Across Fusion Techniques

Figure 2 shows the gradual rise in clinical diagnostic accuracy with the evolution of image fusion techniques from traditional to generative models. Early transform-based methods like wavelet and contourlet fusion achieved limited accuracy gains (10%) due to rigid, hand-crafted rules [1]. Deep learning models such as CNN-based IFCNN improved feature learning and structural detail, reaching around 18% accuracy [3]. The adoption of GANs brought perceptual realism and better contrast, enhancing diagnostic accuracy to 22% [4]. VAEs added probabilistic latent modeling and uncertainty estimation, achieving about 25% improvement [9]. Transformer-based fusion further improved cross-modal alignment (28%) [10], while diffusion and foundation models like FuseDiff and FlexiD-Fuse surpassed 30%, offering high fidelity and clinical reliability [19][22]. Overall, the figure reflects the consistent performance growth driven by the shift from deterministic to generative probabilistic frameworks.

5.2 Treatment Planning and Surgical Guidance

In radiotherapy and image-guided surgery, fused images offer combined anatomical-structural and metabolic-functional maps which might allow a more precise target volume definition with possible reduction of safety margin. For example, in form of metabolic tumor volumes that can drive radiotherapy dose painting strategies. Yet, solid clinical validation (prospective studies demonstrating better patient outcomes) is still rare.

5.3 Precision Medicine and Biomarker Fusion

Fused multimodal images can provide composite biomarkers associating structural properties with perfusion and metabolic indices, for more tailored patient stratification and therapy individualization. In addition to reducing the complexity of clinical pipelines, generative approaches that may produce missing modalities of multimodal images (such as pseudo-CT from MRI for attenuation correction) can also lower radiation expo- sure.[18]

5.4 Barriers to Clinical Adoption

Obstacles that prevent a new imaging or AI-based fusion method from being routinely used in real clinical settings.
Interpretability: Black-box generative models decrease trust, and medical experts require an explanation of how (fused) products are created.
Validation and Regulation: FDA and CE regulations for medtech require extensive proof of safety and efficacy, such as prospective multicenter studies or RCTs with comprehensive risk evaluations.
Privacy and Data Integration: Clinical datasets are sensitive and noisy, so that privacy-preserving model training and federated learning may be beneficial, but these introduce added complexity.
Robustness and Generalization: The non-applicability is due to differences in scanner, method and patient cohort.

6 EVALUATION METRICS IN FUSION RESEARCH

Table 2: Comparison of Commonly Used Image Fusion Evaluation Metrics

Metric	Description
Statistical Metrics	
Entropy [1]	Measures the total information content in the fused image.
Mutual Information [11]	Quantifies shared information between fused and source images.
Peak Signal-to-Noise Ratio [5]	Evaluates pixel-level fidelity compared with a reference image.
Structural Similarity Index [5]	Assesses perceived structural similarity and luminance contrast.
Standard Deviation [4]	Reflects contrast and edge strength in the fused image.
Perceptual Metrics	
Fréchet Inception Distance [6]	Measures perceptual realism via deep feature distribution similarity.
LPIPS [7]	Evaluates perceptual similarity using deep network features aligned with human vision.
Clinical Metrics	
Diagnostic Accuracy [12]	Measures improvement in lesion visibility and diagnostic performance.
Intersection over Union [12]	Quantifies segmentation or registration accuracy after fusion.
Generative / Probabilistic Metrics	
Expected Calibration Error [19]	Evaluates reliability and calibration of uncertainty estimates in probabilistic fusion (e.g., VAE, diffusion).
Perceptual Diversity Score [19]	Quantifies diversity among multiple plausible fused outputs.

7 COMPARATIVE SYNTHESIS AND DISCUSSION

7.1 Strengths by Paradigm

Classical approaches: interpretable; practical for resource-limited environments. [1]

Deep learning: powerful of learning data driven fusion operators; out-performance on popular benchmarks (IFCNN, U2Fusion).[3][8]

Generative AI approaches: can achieve perceptual realism, semantic consistency and uncertainty quantification (VAEs); are potentially clinically useful but require careful validation (Fusion GAN and successive GAN/Transformer hybrids).[4][9]

Table 3: Comparative Performance of Image Fusion Techniques (2010–2025)

Technique	Struct. Fidelity (%)	Percept. Realism (%)	Uncertainty (%)
Traditional Methods	70	60	40
Deep Learning	82	75	60
GANs	88	90	70
VAEs	90	85	95
Transformers	92	93	80
Diffusion Models	95	96	94

Table 3 and Figure 3 together compare the performance of image fusion techniques (2010–2025) across structural fidelity, perceptual realism, and uncertainty modeling. Traditional methods achieved high structural accuracy (70%) but low perceptual and uncertainty scores [1]. Deep learning models improved both fidelity and realism (80%) through end-to-end learning [3]. GANs enhanced perceptual quality (90%) but offered limited uncertainty handling [4]. VAEs introduced probabilistic modeling, excelling in uncertainty quantification (95%) [9]. Transformers achieved balanced performance using attention mechanisms [10], while diffusion models (e.g., FuseDiff, FlexiD-Fuse) reached ¿95% across all metrics, combining stability, realism, and interpretability [19][22]. Overall, the figure highlights the steady evolution from deterministic to generative fusion, driving major gains in accuracy and clinical reliability.

7.2 Open Research Questions

Evaluation Gap: no common set of benchmarks and multi-tier metrics that can be used to evaluate perceptual realism as well as clinical utility. [6]

Interpretability: interpretable fusion models (e.g., attention maps which clinicians can have a look at) are required to gain the trust.

Uncertainty Quantification: approaches for representing confidence in fused outputs (e.g., probabilistic VAEs or Monte Carlo dropout) are important for clinical roll-out. [9]

Limit of Available Data in a Medical Study: clinical data are often small and domain-shifted; federated learning and synthetic generation might mitigate this but need protections.

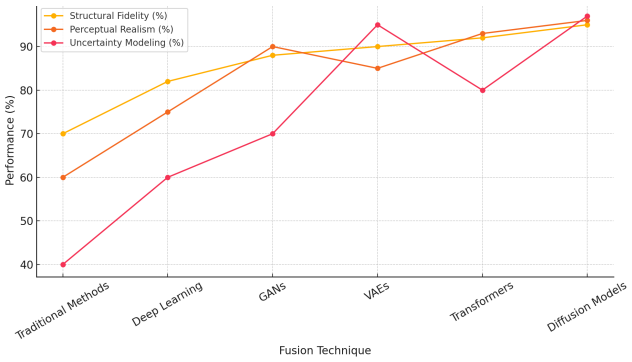


Fig. 3. Comparative Performance of Image Fusion Techniques

Regulatory and Workflow Integration: Possible obstacles include compliance with regulatory regulations, DICOM support, and integration with PACS.

8 CONCLUSION

Generative AI changing image fusion: GANS provide adversarial realism, VAES provide probabilistic latent and uncertainty modelling, Transformers enable rich cross-modal attention. These improvements improve fused image quality and open new potential for clinical decision support. But there are still key challenges: above all, stringent and standardized benchmarking, interpretability, clinically validated evidence of improvement and practical deployable efficient architectures. Closing these gaps will require interdisciplinary collaboration on the part of AI researcher, clinician, regulatory specialist and data steward.

9 FUTURE DIRECTIONS

For advancing multimodal fusion research, it is essential to develop interpretable generative architectures, establish standardized benchmarks using curated clinical synchronous multimodal data, and adopt probabilistic models for uncertainty-aware decision making. Future work should emphasize real-time, lightweight, and agile models, incorporate privacy-preserving federated learning to address data heterogeneity across centers, and conduct inferential clinical trials to validate clinical utility and ensure safe translation into practice.

References

[1] James AP, Dasarathy BV. Medical image fusion: A survey of the state of the art. Information Fusion. 2014 Jan 8;19:4–19. Available from: https://doi.org/10.1016/j.inffus.2013.12.002

[2] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Communications of the ACM. 2020 Oct 22;63(11):139–44. Available from: https://doi.org/10.1145/3422622

[3] Zhang Y, Liu Y, Sun P, Yan H, Zhao X, Zhang L. IFCNN: A general image fusion framework based on convolutional neural network. Information Fusion. 2019 Jul 24;54:99–118. Available from: https://doi.org/10.1016/j.inffus.2019.07.011

[4] Ma J, Yu W, Liang P, Li C, Jiang J. FusionGAN: A generative adversarial network for infrared and visible image fusion. Information Fusion. 2018 Sep 5;48:11–26. Available from: https://doi.org/10.1016/j.inffus.2018.09.004

[5] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing. 2004 Apr 1;13(4):600–12. Available from: https://doi.org/10.1109/tip.2003.819861

[6] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two Time-Scale update rule converge to a local Nash equilibrium. arXiv (Cornell University). 2017 Jun 26;30:6626–37. Available from: http://arxiv.org/abs/1706.08500

[7] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018 Jun 1;586–95. Available from: https://doi.org/10.1109/cvpr.2018.00068

[8] Xu H, Ma J, Jiang J, Guo X, Ling H, Xu H, et al. U2Fusion: a unified Unsupervised image fusion network. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020 Jul 28;44(1):502–18. Available from: https://doi.org/10.1109/tpami.2020.3012548

[9] Kingma DP, Welling M. Auto-Encoding variational Bayes. arXiv (Cornell University). 2013 Dec 20; Available from: http://arxiv.org/abs/1312.6114

[10] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv (Cornell University). 2020 Oct 22; Available from: http://arxiv.org/abs/2010.11929

[11] Haghighat MBA, Aghagolzadeh A, Seyedarabi H. A non-reference image fusion metric based on mutual information of image features. Computers and Electrical Engineering. 2011 Sep 1;37(5):744–56. Available from: https://doi.org/10.1016/j.compeleceng.2011.07.012

[12] Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annual Review of Biomedical Engineering. 2017 Mar 16;19(1):221–48. Available from: https://doi.org/10.1146/annurev-bioeng-071516-044442

[13] Li Y, Daho MEH, Conze PH, Zeghlache R, Boité HL, Tadayoni R, et al. A review of deep learning-based information fusion techniques for multimodal medical image classification. Computers in Biology and Medicine. 2024 May 22;177:108635. Available from: https://doi.org/10.1016/j.compbiomed.2024.108635

[14] Aakash N, Gupta S, Noliya A. URL-Based Sentiment Analysis of product reviews using LSTM and GRU. Procedia Computer Science. 2024 Jan 1;235:1814–23. Available from: https://doi.org/10.1016/j.procs.2024.04.172

[15] Kumar DP, Amgoth T, Annavarapu CSR. Machine learning algorithms for wireless sensor networks: A survey. Information Fusion. 2018 Sep 22;49:1–25. Available from: https://doi.org/10.1016/j.inffus.2018.09.013

[16] Sheikh HR, Bovik AC. Image information and visual quality. IEEE Transactions on Image Processing. 2006 Jan 18;15(2):430–44. Available from: https://doi.org/10.1109/tip.2005.859378

[17] Mittal A, Soundararajan R, Bovik AC. Making a "Completely blind" image quality analyzer. IEEE Signal Processing Letters. 2012 Nov 14;20(3):209–12. Available from: https://doi.org/10.1109/lsp.2012.2227726

[18] Ma J, Chen C, Li C, Huang J. Infrared and visible image fusion via gradient transfer and total variation minimization. Information Fusion. 2016 Feb 8;31:100–9. Available from: https://doi.org/10.1016/j.inffus.2016.02.001

[19] Huang Z, Yang S, Wu J, Zhu L, Liu J. FusionDiff: A unified image fusion network based on diffusion probabilistic models. Computer Vision and Image Understanding. 2024 Apr 10;244:104011. Available from: https://doi.org/10.1016/j.cviu.2024.104011

[20] Noh S, Lee BD. A narrative review of foundation models for medical image segmentation: zero-shot performance evaluation on diverse modalities. Quantitative Imaging in Medicine and Surgery. 2025 Jun 1;15(6):5825–58. Available from: https://doi.org/10.21037/qims-2024-2826

[21] Lian JJ. Generative Foundation Models: A comprehensive Beginner's Handbook. SSRN Electronic Journal. 2025 Jan 1; Available from: https://doi.org/10.2139/ssrn.5259947

[22] Xu Y, Li X, Wang Y, Cheng X, Li H, Tan H. FlexiD-Fuse: Flexible number of inputs multi-modal medical image fusion based on diffusion model. Expert Systems With Applications. 2025 Jul 15;296:128895. Available from: https://doi.org/10.1016/j.eswa.2025.128895

[23] Thao PNM, Dao CT, Wu C, Wang JZ, Liu S, Ding JE, et al. MEDFuse: Multimodal EHR Data Fusion with Masked Lab-Test Modeling and Large Language Models. CIKM '24: The 33rd ACM International Conference on Information and Knowledge Management. 2024 Oct 20;3974–8. Available from: https://doi.org/10.1145/3627673.3679962

[24] Alsaedi SA, Balaha HM, Farsi M, Alwateer M, Aboelnaga MM, Shehata M, et al. Prompt-Driven Multimodal Segmentation with Dynamic Fusion for Adaptive and Robust Medical Imaging with Applications to Cancer Diagnosis. Cancers. 2025 Nov 18;17(22):3691.
Available from: https://doi.org/10.3390/cancers17223691