UJRET
Bi-annual refereed journal
University Journal of Research Engineering & Technology
September 2024
Volume 1, (1) pp. 17-33

A Comprehensive Review of Multimodal Image Fusion Using Deep Learning Techniques

Samir B. Joshi
Assistant Professor, Faculty of Engineering and Technology
Ganpat University, Ganpat Vidyanagar, 384012, Gujarat, India

Amrutbhai N. Patel
Associate Professor, Faculty of Engineering and Technology
Ganpat University, Ganpat Vidyanagar, 384012, Gujarat, India

Anand J. Patel
Professor, LDRP Institute of Technology and Research, KSV University
Gandhinagar, 382025, Gujarat, India

Abstract

Multimodal image fusion (MMIF), the process of combining complementary information from multiple imaging sources to enhance the quality and content of the resulting image, has gained significant attention across various fields including medical imaging, remote sensing, surveillance, robotics and automation, artificial intelligence, machine learning and many more. Modalities and techniques of Image fusion are of utmost importance. Deep learning (DL) architectures like Convolutional Neural Network, Generative Adversarial Networks, Auto-encoders, Transformers etc have revolutionized multimodal image fusion by providing powerful tools to learn complex mappings between heterogeneous data sources. This paper presents a comprehensive review of recent advances in MMIF using DL methods. We discuss the fundamentals of MMIF, various DL architectures employed for fusion tasks, evaluation metrics and various applications of MMIF. Additionally, we highlight challenges and future research scopes in this rapidly evolving area.

Keywords: Multimodal Image Fusion, Deep learning, Architectures of DL, Applications of Multimodal image fusion

1. Introduction

In various domains such as medical imaging, remote sensing, surveillance, and robotics, the acquisition of information through multiple imaging modalities has become commonplace. However, each modality typically captures different parameters of the scene or object under observation, leading to complementary but incomplete representations. Multimodal image fusion (MMIF) addresses this challenge by integrating information from various sources to create a fused image that preserves and enhances the relevant features from each modality. (Kalamkar & A., 2023)

The primary objective of this fusion is to magnify the quality, reliability, and interpretability of the fused image compared to individual input images. Fusion aims to exploit the complementary information present in different modalities to improve overall understanding and analysis of the scene.

Corresponding author. Tel.: 9428853091
E-mail address:samir.joshi@ganpatuniversity.ac.in

The significance of MMIF lies in its ability to provide richer and more comprehensive information compared to individual modalities alone. By leveraging the strengths of each modality while compensating for their weaknesses, fusion techniques enable improved decision-making, analysis, and understanding of complex scenarios. This is particularly crucial in applications where accurate and detailed information is paramount, such as medical diagnosis, environmental monitoring, and security surveillance. (Kalamkar & A., 2023)

In medical imaging, for instance, fusion of MRI (Magnetic Resonance Imaging) and PET (Positron Emission Tomography) scans can enhance tumour localization and characterization, aiding in treatment planning and monitoring. In remote sensing, fusion of optical and SAR (Synthetic Aperture Radar) images enables better land cover classification, crop monitoring, and disaster management. Similarly, in surveillance systems, fusion of visible and infrared imagery enhances target detection and tracking in various lighting conditions. (Basu et al., 2024)

Traditionally, MMIF has been approached using handcrafted algorithms based on mathematical transformations, such as wavelet transforms, principal component analysis (PCA), or sparse representation.'(Hermessi et al., 2021) However, these methods often struggle with capturing complex relationships between modalities, handling noise and other issues, and maintaining computational efficiency. (Stahlschmidt et al., 2022)

The advent of DL has revolutionized MMIF by providing powerful tools to automatically grasp hierarchical representations directly from data. Deep neural networks can effectively model the intricate mappings between heterogeneous modalities, adapt to diverse imaging conditions, and generalize well to unseen scenarios. (J. Li et al., 2022a) This has led to new opportunities for applications across various domains.

In summary, multimodal fusion is a fundamental technique for integrating information from diverse imaging modalities to enhance the quality, reliability, and interpretability of images, enabling advanced analysis and decision-making in number of applications.

In this context, this paper targets to provide a detailed exploration of MMIF fundamentals. The paper also focuses on various DL based various architectures. It also presents various datasets for DL based image fusion with their advantages and limitations. The survey will contribute to a better understanding of MMIF with the concepts of DL for fusion.

2. Tree chart for Image Fusion

MMIF amalgamates many images taken from diverse sources into singular image, ensuring the retention of crucial and pertinent information of each modality. Figure 1 shows the tree chart for the same.

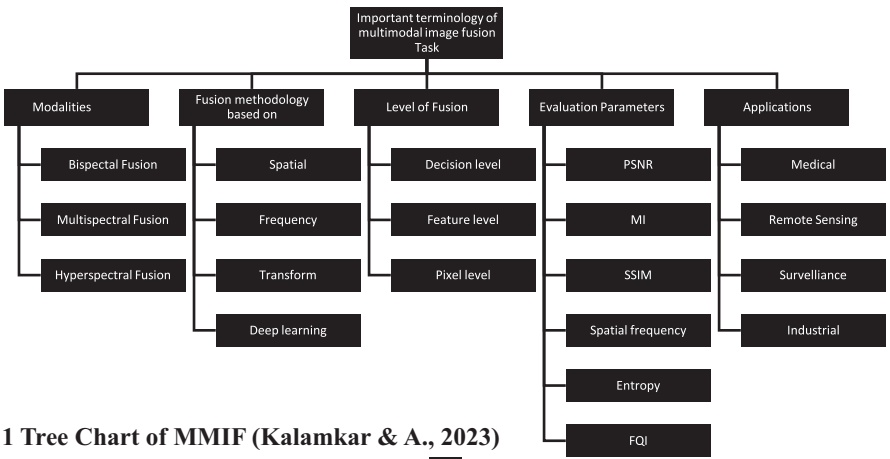


Figure 1 Tree Chart of MMIF (Kalamkar & A., 2023)

This tree chart shows the concepts of MMIF like modalities, Fusion methodologies, Level of fusion, Evaluation parameters as well as applications.

2.1 Modalities of Input

2.1.1 Bi-spectral fusion

Bi-spectral fusion commonly involves amalgamating data from two or more sensors or sources, each capturing information from distinct segments of the electromagnetic spectrum, to produce a more comprehensive and detailed depiction of a scene or object. For instance, in remote sensing or imaging applications, combining data from sensors that capture different wavelengths of light, such as visible, infrared, or microwave, can provide richer information about the target being observed. This fusion can enhance features such as contrast, resolution, and object detection capabilities. The term "bi-spectral" specifically implies the fusion of two spectral bands or sources. Applications of bi-spectral fusion span various fields including remote sensing, surveillance, medical imaging, and military reconnaissance, among others. The goal is to leverage the complementary strengths of different sensors to improve understanding, analysis, and decision-making based on the collected data.

2.1.2 Multispectral image fusion

Multispectral image fusion combines information from various images captured at different spectral bands or wavelengths into a single, composite image. This process aims to enhance the quality, richness, and interpretability of the resulting image by integrating complementary information from each spectral band.

There are several techniques available for multi-spectral fusion, including decision, feature, and image-level fusion methods. Feature-level fusion involves extracting features from each spectral band and combining them into a single feature vector, which then produces the resulting fused image. Decision level fusion makes separate each spectral band decisions and merge them together to reach the decision. In contrast, image level fusion produces a singular fused image after merging images from each spectral band. Overall, multi-spectral fusion serves as a potent entity for enhancing the precision and efficacy of image analysis in remote sensing and other domains. It is done by consolidating the information offered by multiple spectral bands into a fused image (Kalamkar & A., 2023). Feng et al. introduced a model-based DL approach to combine high resolution multi-spectral and low-resolution multi-spectral images to create high resolution hyper-spectral images. Iterative algorithm uses the proximal gradient method. (Feng et al., 2020)

2.1.3 Hyper-spectral image fusion

Images with high spectral resolution and comprising many spectral bands are merged using hyperspectral fusion. Sensors related to hyperspectral, excel at capturing images with finely spaced spectral intervals, offering detailed insights into the material composition as well as physical attributes of objects within a scene. (Sara et al., 2021) The objective of hyperspectral fusion is to create a singular fused image that encapsulates the most informative characteristics from each spectral band while mitigating noise and other unwanted parameters. This technique finds utility across diverse fields such as environmental monitoring, agriculture and mineral search. Various fusion methods are employed for hyperspectral fusion, including non-negative matrix factorization (NMF), principal component analysis (PCA) as well as independent component analysis (ICA), ICA and PCA are commonly used for hyperspectral feature extraction, while NMF supports in hyperspectral immixing, facilitating the separation of spectral signatures corresponding to different materials within the scene (Kalamkar & A., 2023). Hyperspectral sensors capture information in detail regarding the absorption and reflectance of light by plant leaves, enabling the detection of problems like disease, stress and nutrient deficiencies (Kalamkar & A., 2023).

2.2 Fusion methodology

2.2.1 Spatial domain

Early research in medical image fusion technology, particularly in the spatial domain, has garnered significant attention. Spatial domain fusion techniques are straightforward, allowing fusion regulations to be applied to the source image pixels straight away, resulting in the merged image. Various methods exist within the spatial domain, including principal component analysis, high-pass filtering, average method, saturation method of hue intensity, MSM (Maximum Selection Method), MNM (Minimum Selection Method) as well as the Brovey method. However, the fused images in the spatial domain often suffer from spectral and spatial distortions, leading to a decline in research interest in recent years. Consequently, researchers are now integrating spatial domain fusion strategies into transformation domain approaches to develop novel research methods. (B. Huang et al., 2020)

2.2.2 Frequency domain

Compared to spatial domain methods, frequency-based image fusion offers reduced distortion of spectral details. In frequency-domain approaches, source images are transformed into frequencies using Fourier Transform (FT). Fusion criterion are applied to the resulting image. At the end, inverse FT is performed to get the fused image.(Princess et al., 2014) This technique is categorized into transform and pyramid domains. Pyramid-based image fusion methods focus to enhance the spectral information available in the resulting fused image. (Krishnamoorthy & Soman, 2010) Within pyramid transforms, the level of detail in the fused image is contingent upon the decomposition levels, involving scales and appropriate selection of family. The Laplacian pyramid is a commonly utilized technique which incorporates a multi-scale decomposition process to derive coefficients. The coefficients pass through fusion via a weighted fusion method, followed by an inverse Laplacian operation. However, this approach is susceptible to visualization challenges. Pyramid transform methods encompass the slope pyramid, morphological pyramid and filter subtract decimate (FSD) method. (Krishnamoorthy & Soman, 2010)

2.2.3 DL based fusion

The emergence of DL has advanced the field of MMIF significantly by offering powerful and flexible frameworks to learn complex mappings between heterogeneous data sources. DL techniques leverage hierarchical architectures composed of various layers of interconnected neurons, enabling the automatic extraction of high-level features directly from raw data. (Y. Li et al., 2024) This paradigm shift has revolutionized multimodal fusion in several key ways:

End-to-End Learning: DL models allow for end-to-end learning, where the entire fusion process is optimized directly from raw input data to the fused output. This eliminates the need for handcrafted features or intermediate representations, enabling more efficient and effective fusion. (Y. Li et al., 2024)

Feature Learning: DNNs are adept at learning hierarchical representations of data as well as automatically extracting features at multiple levels of abstraction. This capability is particularly valuable for multimodal fusion, where different modalities may encode complementary information at different scales and levels of granularity. (Kalamkar & A., 2023; Y. Li et al., 2024)

Model Flexibility: DL architectures like Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs) and Recurrent Neural Networks (RNNs) offer flexibility in modelling complex relationships between modalities. They can capture spatial, spectral, temporal, and semantic correlations, making them well-suited for a wide range of fusion tasks. (Kalamkar & A., 2023; Y. Li et al., 2024)

Adaptability to Data: DL models are highly adaptable to diverse imaging modalities, data distributions,

and imaging conditions. They can learn from large-scale datasets with minimal manual intervention, making them suitable for handling complex and heterogeneous multimodal data sources. (H. Zhang & Qie, 2023)

Table 1 Relative outline of multimodal image fusion methodologies

Domain	Advantages	Disadvantages
Spatial	• Pixel level fusion • faster and simple computation • Less complexity	• sharpening issue in resultant image • Less quality and spectral distortion • Smooth edges and low contrast problems (Mishra & Palkar, 2015)
Frequency	• Less distortion • Multiscale Transform • High Signal to Noise Ratio (SNR) • Edges information visual quality is better	• Missing spatial details • Complicated • activity level measurements deficiency • Image registration Problems (Kalamkar & A., 2023)
DL	• Extract enough features (Bhat & Koundal, 2021) • Improved image optimization • Better performance on large size input data	• Semantic loss • Overfitting and convergence issues (Bhat & Koundal, 2021)

Generative Models for Data Augmentation: Generative models such as Variational Auto-Encoders (VAEs) and GANs can generate synthetic multimodal data samples, which can be used for data augmentation and improving the robustness of fusion models, especially in scenarios with limited training data. (Bond-Taylor et al., 2022)

Interpretability and Uncertainty Estimation: Recent research has focused on enhancing the interpretability of deep fusion models and estimating uncertainty in fusion outputs. This is crucial for applications where transparency and reliability are paramount, such as medical diagnosis and decision support systems. (Bond-Taylor et al., 2022; H. Zhang & Qie, 2023)

The traditional methods for MMIF typically rely on mathematical transformations and heuristic algorithms to combine information from different modalities.(Kalamkar & A., 2023) (B. Huang et al., 2020)

Table 1 compares main image fusion methodologies.

2.3 MMIF strategies

2.3.1 Pixel level Fusion

Approach: Combines pixel values from corresponding locations in the input images using averaging or weighted averaging.

Limitations: Lack of adaptability to complex relationships between modalities. Ignores spatial context and semantic information, leading to suboptimal fusion results. Vulnerable to noise and issues, especially in real-world scenarios. (B. Huang et al., 2020)

2.3.2 Feature level Fusion

Approach: Features extraction from individual modalities using handcrafted descriptors (e.g., SIFT, SURF) or statistical measures (e.g., histogram, texture features). These features are then fused using simple fusion rules (e.g., maximum, minimum, average). (B. Huang et al., 2020)

Limitations: Reliance on predefined features may limit adaptability to diverse imaging scenarios. Feature extraction may not capture all relevant information, leading to loss of discriminative power. Moreover, feature-level fusion often ignores spatial context and global relationships in the data.

2.3.3 Decision-level Fusion

Approach: Makes decisions independently on each modality and then combines the decisions using predefined rules (e.g., majority voting, weighted voting).

Limitations: Decision-level fusion assumes that decisions from individual modalities are equally reliable, which may not be the case in practice. It may not effectively exploit complementary information from different modalities, especially in cases where the decisions are not directly comparable or complementary. (B. Huang et al., 2020)

2.4 Evaluation Metrics

Here are some evaluation metrics for MMIF presented.

Peak Signal-to-Noise Ratio (PSNR): PSNR assess the quality of the resultant image by comparing it to the reference image. It computes the ratio of the maximum possible pixel value to the mean squared error between the ground truth and fused image. Higher PSNR values indicate better fusion quality.

Structural Similarity Index (SSIM): SSIM assesses the commonness between the resultant image and the ground truth image in terms of luminance, contrast, and structure. It ranges from -1 to 1. 1 indicates perfect similarity. SSIM is often preferred over PSNR as it considers perceptual aspects of image quality.

Mutual Information (MI): MI shows the amount of shared information between the input modalities and the fused image. It quantifies reduction in uncertainty about one modality given knowledge of the other modalities. Higher MI values indicate better fusion performance.

Entropy: Entropy presents the quantity of information or uncertainty in the resultant image. Lower entropy values indicate higher certainty or predictability in the fused image, reflecting better fusion quality.

Other indicators:

In evaluating the performance of multimodal fusion methods and deep learning networks for medical classification tasks, commonly used indicators include True Negatives (TN), True Positives (TP), False Negatives (FN), and False Positives (FP). These indicators help compute various performance metrics, such as sensitivity, specificity, accuracy, precision, and F1 score. Additionally, metrics like AUC (Area Under the Curve) and Kappa are frequently employed to assess the effectiveness of medical classification models

- Accuracy $\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$
- Sensitivity $= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
- Specificity $= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$
- F1 Score $= \frac{2\times\mathrm{TP}}{2\times\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$
- PPV (Positive Predictive Value) $= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$
- NPV (Negative Predictive Value $= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}}$
- Area under the receiver operating characteristics curve (AUC)
- Cohen's Kappa (Kappa)$=\frac{\rho_o-\rho_e}{1-\rho_e}$

It's important to note that no single metric can fully capture the performance of MMIF algorithms. A combination of objective and subjective evaluation methods is often recommended for comprehensive assessment. Additionally, benchmarking protocols and evaluation metrics should be carefully selected considering the fusion task and dataset.

3. Introduction to DL based MMIF

While traditional fusion methods have been widely used and studied, they often face limitations in terms of adaptability, robustness, and scalability, especially in complex and dynamic environments. These limitations have spurred the exploration of DL techniques, which have shown promise in overcoming many of these challenges by automatically learning representations directly from data. Table 2 describes these architectures

Table 2: DL architecture description

Architecture	Description
Convolutional Neural Network (CNN)	CNN can model spatial structures, such as images or volumes. Convolutional kernels model local information by sliding over input data (Goodfellow et al., 2016).
Generative Adversarial Networks (GANs)	Enable the generation of synthetic multimodal data samples, that can be used to augment training data and improve the robustness of fusion models, especially in scenarios with limited labeled data. (Bond-Taylor et al., 2022)
Auto encoders (AE)	Auto-Encoders (AEs) learn low-dimensional encodings by compressing and reconstructing the input data. They utilize various types of layers, including convolutional and fully connected layers. (Ballard, 1987)
Transformer	The Transformer model utilizes a multi-headed attention mechanism for feature extraction, relying exclusively on attention to process information. (Vaswani et al., 2017a)

3.1 DL Architectures for MMIF:

3.1.1 Convolutional Neural Networks (CNNs)

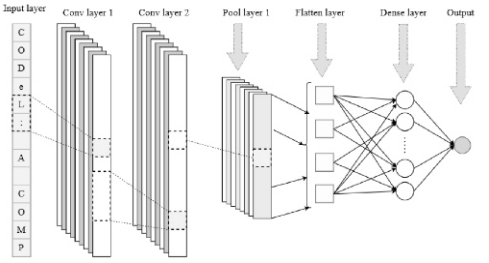


Figure 2 Overview to simple one-dimensional (1D) Convolution Neural Network with Two Convolution layer (Younesi et al., 2024)

Convolutional Neural Networks (CNNs) have risen to prominence as a potent asset in the realm of MMIF due to their strength to automatically learn hierarchical representations directly from raw data. CNNs have been widely adopted in various fusion scenarios, ranging from medical imaging to remote sensing and surveillance. Here, we discuss the application of CNNs for multimodal fusion and highlight key approaches and architectures. (Kalamkar & A., 2023)

Early Fusion: In early fusion, CNNs are trained to directly process multiple input modalities

simultaneously. Each modality's input is fed into separate channels of the CNN's input layer. The network learns to extract features from each modality independently and fuse them at early layers.

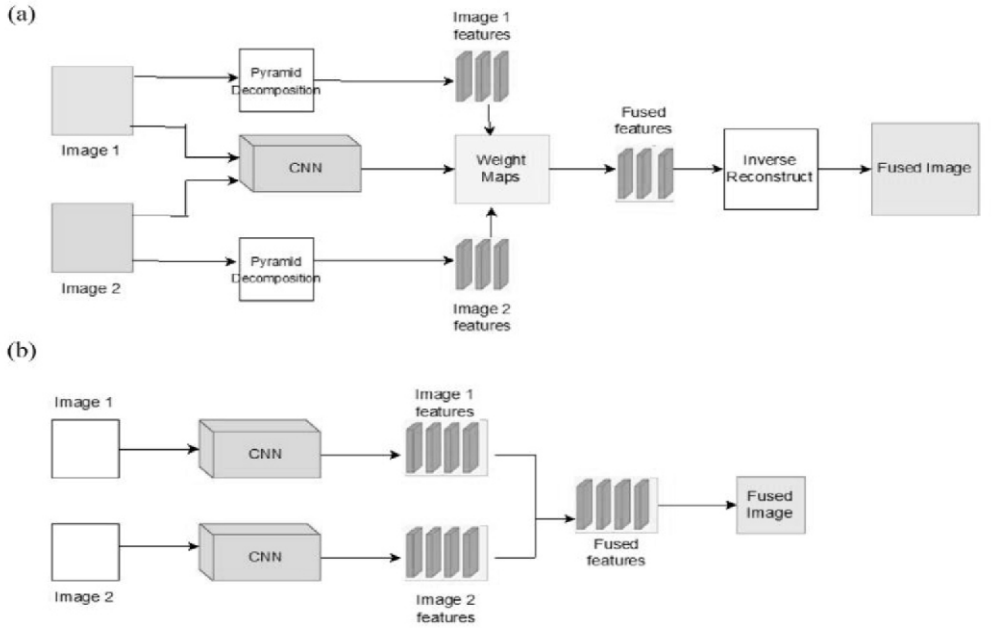


Figure 3 CNN-based approaches for MMIF. (Kalamkar & A., 2023)

3.1.2 Generative Adversarial Networks (GANs):

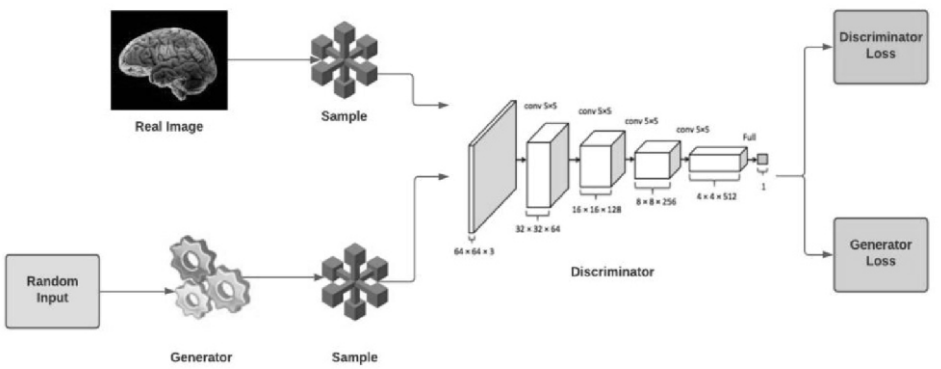


Figure 4 Block diagram of Generative Adversarial Network (GAN) (Aggarwal et al., 2021)

GANs have been employed for multimodal fusion tasks, where one network generates synthetic data samples from one modality conditioned on samples from another modality, while another network discriminates between real and synthetic samples. GAN-based fusion models learn to generate realistic and informative multimodal representations, facilitating data augmentation and enhancing fusion performance.

GANs enable the generation of synthetic multimodal data samples, which can be used to augment training data and improve the robustness of fusion models, especially in scenarios with limited labelled data.

Generative Adversarial Networks (GANs) have shown promise in multimodal fusion tasks by facilitating the generation of realistic and informative representations from heterogeneous input data sources. In the context of fusion tasks, GANs consist of generator and discriminator neural networks, which are trained adversarial to produce high-quality fused representations.

Loss Function: The training objective of GANs involves a minimax game between the generator and discriminator. The generator targets to minimize the discriminator's strength to differentiate real and generated representations, while the discriminator targets to maximize its accuracy in discriminating between them. (C. Zhang et al., 2024)

Overall, Generative Adversarial Networks (GANs) offer a powerful framework for multimodal fusion tasks by enabling the generation of realistic and informative fused representations from heterogeneous input data sources. By training adversarial, GAN-based fusion models learn to produce high-quality representations that capture the complementary information present in multiple modalities, leading to improved performance in various fusion applications.

3.1.3 Auto-Encoders and their variants

Auto-Encoders, a type of neural network architecture, have gained attention in MMIF due to their strength to learn compact view of input data and capture salient features across multiple modalities. Variants of auto-encoders, including traditional auto-encoders, variational Auto-Encoders (VAEs), and de-noising Auto-Encoders, have been adapted and extended for multimodal fusion tasks.

Table 3 Exploration of MMIF approaches utilizing DL.

Author	Technique used	Fusion level	Image modality	Fusion Strategy	Evaluation metrics
(Y. Li, Zhao, et al., 2021)	CNN	Feature	MRI and SPECT images, CT and MRI,	Batch Processing	RMSE, SF, EOG, PSNR, SSIM, MI, MEAN, SD, QAB/F GRAD,
(Wang et al., 2021)	CNN	Pixel	MR-T1–MR-T2, MR–CT, MR–PET, and MR–SPECT images	A trained Siamese convolutional network is employed to integrate the pixel activity details from source images. It produces a weight map. Subsequently, a contrast pyramid is then applied to break down the source image.	QTE, QAB/F, MI, VIF
(Y. Li, Ren, et al., 2021)	SAE	Feature	MRI-T1 and MRI-T2, MRI–SPECT	Non-subsampled contour-let transform (NSCT) is utilized to decompose the image. Feature extraction is employed by Sparse Auto-Encoder (SAE). Subsequently, spatial frequencies are computed for the acquired features.	MI, EN, SD, QAB/F,
(Kaur & Singh, 2021)	CNN	Feature	CT–MRI	NSCT serves for the decomposition of the image. Following this step, the Inception (Xception) model is utilized to extract features from the source images. Optimal features are then chosen	Edge strength, fusion symmetry, entropy, FF

(Mi et al., 2022)	GAN	Feature	Visible–Infrared	The quantity of discriminators tasked with distinguishing between pairs of fused images and those with extreme exposure is augmented. Concurrently, a generator network undergoes training to produce fused images.	SD, PSNR, SCD, CC, SSIM

*Acronyms: SD - Standard Deviation, SSIM - Structural Similarity Metric, FF - Fusion Factor, SAE - Stacked Auto-Encoder, VIF - Variance Inflation factor, RMSE - Root Mean Square Error, EOG - Energy of image gradient, PSNR - Pulse Signal to Noise Ratio, , GAN - Generative Adversarial Network, MI - Mutual Information, GRAD-Gradient, AG - Average Gradient, EN Entropy, , CC - cross-correlation, CSR - Convolutional Sparse Representation, , SF - Spatial frequency, QAB/F-edge information, MEAN

3.1.4 Transformer

Transformers are Deep Neural Networks (DNNs). They gather contextual relationships within available data by utilizing a self-attention mechanism. Transformer models facilitates parallel processing.(Islam et al., 2024) Consequently, The researchers are attracted towards transformer-based models due to their extensive potential and noteworthy accomplishments. Natural Language Processing (NLP) tasks like audio and speech processing, Internet of Things (IoT), Computer Vision (CV) etc. can be accomplished using transformers.(Islam et al., 2024)

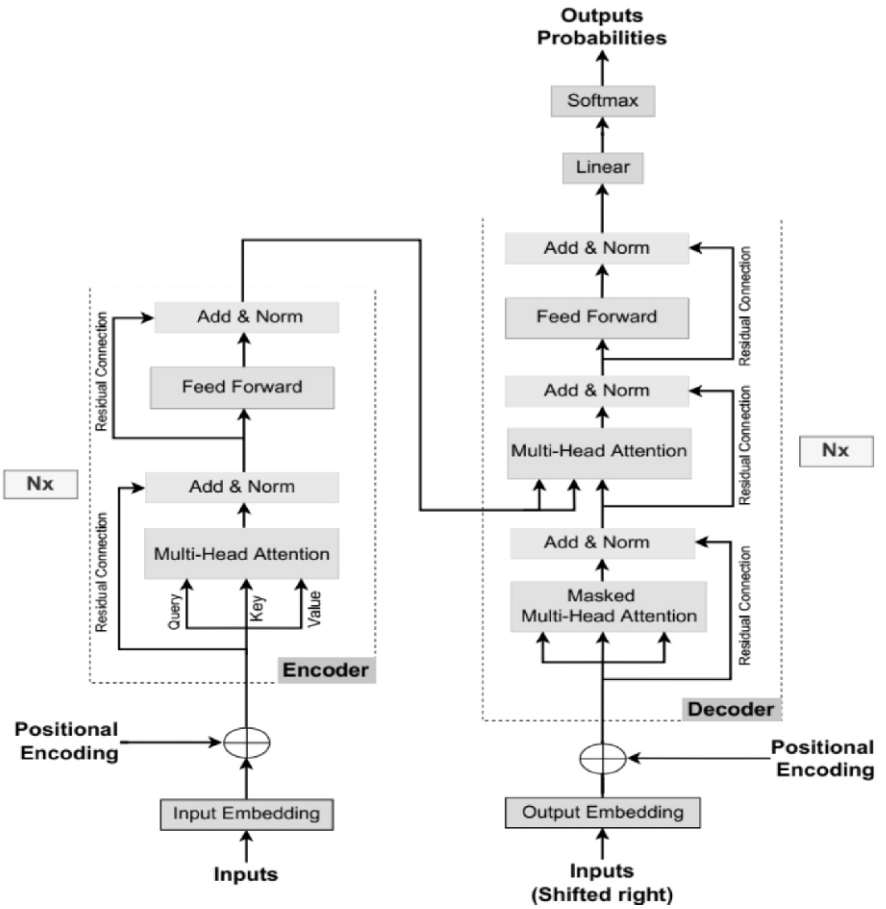


Figure 5 Transformer Architecture (Islam et al., 2024; Vaswani et al., 2023)

Here are some key components and concepts of a transformer:

Self-Attention Mechanism: This constitutes the fundamental concept behind transformers. Self-attention enables the model to assess the significance of individual words within a sentence during the encoding or decoding of sequences. It calculates attention scores for every pair of words in a sequence and utilizes these scores to generate weighted sums of the input sequence. (Vaswani et al., 2023)

Multi-Head Attention: In order to capture diverse types of information and relationships, multiple sets of attention mechanisms, referred to as attention heads, are employed. Each attention head specializes in learning distinct aspects of the relationships between words.(Islam et al., 2024; Vaswani et al., 2023)

Positional Encoding: Positional encodings are added to provide information about the input sequence word positions because transformers do not inherently understand the order of elements in a sequence.(Islam et al., 2024)

Feedforward Neural Networks: The transformer applies feedforward neural networks to each position in the sequence individually and uniformly after the self-attention mechanism. (Islam et al., 2024)

Encoder-Decoder Architecture: Transformers commonly consist of an encoder and a decoder. In tasks such as machine translation, the encoder manages the input sequence, while the decoder creates the output sequence utilizing the representations provided by the encoder.(Islam et al., 2024)

Residual Connections and Layer Normalization: To facilitate the training of deep models, transformers use residual connections around each sub-layer (such as self-attention and feedforward layers) followed by layer normalization.(Islam et al., 2024)

Transformer Blocks: These are the building blocks of transformers, consisting of various layers of self-attention.

4. Applications of MMIF:

MMIF finds applications across various domains where integrating information from multiple imaging sources enhances decision-making, analysis, and understanding of complex scenarios. Here are some prominent applications of MMIF:

4.1 Medical Imaging:

Disease Diagnosis and Treatment: Fusion of multiple medical imaging modalities like PET, MRI, CT and ultrasound enables accurate diagnosis and treatment planning for various diseases, including cancer, neurological disorders, and cardiovascular conditions. (Kalamkar & A., 2023)

Image-Guided Surgery: Fusion of preoperative and intraoperative images enhances surgical navigation, allowing surgeons to visualize critical structures and tumour margins more clearly during procedures.

Image Registration and Segmentation: Fusion facilitates image registration and segmentation tasks by aligning images from different modalities and enhancing the delineation of anatomical structures and lesions.

4.2 Remote Sensing:

Environmental Monitoring: Fusion of optical, infrared, and radar satellite imagery supports monitoring of environmental phenomena such as land cover changes, deforestation, urbanization, and natural disasters. (J. Li et al., 2022b)

Precision Agriculture: Integration of multispectral and hyperspectral data aids in crop monitoring, yield prediction, disease detection, and soil analysis, enabling optimized agricultural practices and

resource management.

Disaster Management: Multimodal fusion assists in disaster response and recovery efforts by providing timely and accurate information on terrain conditions, damage assessment, and emergency resource allocation.

4.3 Surveillance and Security:

Object Detection and Tracking: Fusion of visible and infrared imagery improves object detection and tracking in surveillance systems, particularly in low-light or adverse weather conditions.

Perimeter Protection: Integration of various sensors such as cameras, thermal imagers, and motion detectors enhances perimeter security by detecting intrusions and suspicious activities.

Biometric Identification: Fusion of multiple biometric modalities (e.g., face, iris, fingerprint) enhances accuracy and robustness in biometric identification systems, ensuring reliable access control and identity verification. (Stahlschmidt et al., 2022)

4.4 Robotics and Autonomous Systems:

Navigation and Mapping: Fusion of visual, LiDAR, and depth sensor data enables robots and auto vehicles to navigate complex environments, build accurate maps, and localize themselves with high precision. (Kalamkar & A., 2023)

Object Recognition and Manipulation: Integration of visual and tactile sensing modalities enhances object recognition and manipulation capabilities in robotic systems, enabling tasks such as grasping, sorting, and assembly in unstructured environments.

4.5 Artificial Intelligence and Machine Learning:

Multimodal Learning: Fusion of heterogeneous data sources (e.g., images, text, audio) facilitates multimodal learning tasks such as visual question answering, image captioning and sentiment analysis, enabling AI systems to understand and generate content across modalities. (Kalamkar & A., 2023)

4.6 Cross-Modal Retrieval:

Fusion aids in cross-modal retrieval tasks, where the goal is to retrieve relevant information across different modalities (e.g., retrieving images based on textual queries or vice versa), supporting applications such as content-based image retrieval and multimedia search engines.

Overall, MMIF plays a critical role in a wide range of applications, enhancing data integration, analysis, and interpretation across diverse domains and enabling advanced technologies and systems to operate effectively in complex real-world scenarios.

5. Challenges and Future Directions:

In recent years, significant progress has been made in MMIF through DL. However, several challenges persist, alongside promising directions for future research. Here are some key challenges and potential directions for advancing the field:

5.1 Data Heterogeneity and Variability:

Challenge: Multimodal datasets often exhibit heterogeneity in terms of imaging modalities, acquisition conditions, and data distributions. Handling such variability is crucial for robust fusion performance.

Future Directions: Research focusing on domain adaptation, transfer learning, and data augmentation techniques can help address data heterogeneity challenges. Additionally, generative models such as

domain adversarial networks and meta-learning approaches can facilitate learning representations invariant to domain shifts.

5.2 Model Interpretability:

Challenge: DL models designed for multimodal fusion frequently lack interpretability, presenting challenges in comprehending the logic behind fusion decisions and instilling faith in the model's outputs, particularly in critical domains like healthcare.

Future Directions: Incorporating explainable AI techniques, such as attention mechanisms, saliency maps, and model visualization methods, can enhance the interpretability of fusion models. Exploring hybrid models that merge the predictive capabilities of DL with the transparency offered by symbolic reasoning approaches represents a promising avenue

5.3 Scalability and Efficiency:

Challenge: The computational demands of DL models for multimodal fusion can be substantial, necessitating significant memory and processing resources. This constraint restricts their scalability and deployment in environments with limited resources.

Future Directions: Investigations into lightweight architectures, techniques for model compression, and hardware acceleration have the potential to enhance the scalability and efficiency of multimodal fusion models. This could facilitate their deployment on edge devices and in real-time applications.

5.4 Generalization and Robustness:

Challenge: DL models trained on specific datasets may struggle to generalize to unseen data or adapt to changes in imaging conditions, leading to reduced robustness and performance degradation in real-world scenarios.

Future Directions: Enhancing model generalization through techniques such as domain generalization, meta-learning, and adversarial training can improve robustness to variations in data distribution and imaging conditions. Incorporating uncertainty estimation methods can also provide insights into model confidence and reliability.

5.5 Integration of Prior Knowledge:

Challenge: DL models often lack the ability to incorporate domain-specific knowledge or constraints into the fusion process, limiting their ability to leverage expert knowledge and improve fusion performance.

Future Directions: Developing hybrid models that combine DL with probabilistic graphical models, Bayesian approaches, or expert systems can enable the integration of prior knowledge into the fusion process. Knowledge distillation techniques can also transfer knowledge from complex models to simpler, interpretable models with explicit incorporation of domain knowledge.

5.6 Ethical and Societal Implications:

Challenge: As DL models become increasingly pervasive in multimodal fusion applications, ethical considerations surrounding data privacy, bias, fairness, and accountability become paramount.

Future Directions: Addressing ethical and societal implications requires interdisciplinary collaboration among researchers, policymakers, and stakeholders to develop guidelines, regulations, and ethical frameworks for responsible AI deployment. Incorporating principles of fairness, transparency, and accountability into the design and deployment of fusion models is essential for building trust and ensuring equitable outcomes.

By addressing these challenges and exploring innovative research directions, the field of MMIF using DL can continue to advance, leading to more robust, interpretable, and ethically responsible fusion

solutions with broad societal impact.

6. Conclusion:

Overall, the emergence of DL has propelled MMIF to new heights, enabling more accurate, robust, and scalable fusion solutions across a wide range of applications. Continued research in this area holds the promise of further advancements, leading to enhanced understanding and utilization of multimodal data in various domains.

This review paper provides a thorough examination of MMIF techniques employing DL methodologies. It elucidates the current state-of-the-art, challenges, and future directions, serving as a valuable asset for researchers, practitioners, and enthusiasts in the field.

References

[1] Aggarwal, A., Mittal, M., & Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, *1*(1), 100004. https://doi.org/10.1016/j.jjimei.2020.100004

[2] Basu, S., Singhal, S., & Singh, D. (2024). A Systematic Literature Review on Multimodal Medical Image Fusion. *Multimedia Tools and Applications*, *83*(6), 15845–15913. https://doi.org/10.1007/s11042-023-15913-w

[3] Bhat, S., & Koundal, D. (2021). Multi-focus image fusion techniques: A survey. *Artificial Intelligence Review*, *54*(8), 5735–5787. https://doi.org/10.1007/s10462-021-09961-7

[4] Bond-Taylor, S., Leach, A., Long, Y., & Willcocks, C. G. (2022). Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), 7327–7347. https://doi.org/10.1109/TPAMI.2021.3116668

[5] Borji, A. (2022). *Complementary datasets to COCO for object detection* (arXiv:2206.11473). arXiv. http://arxiv.org/abs/2206.11473

[6] Chen, S., & Guo, W. (2023). Auto-Encoders in Deep Learning—A Review with New Perspectives. *Mathematics*, *11*(8), 1777. https://doi.org/10.3390/math11081777

[7] Feng, X., He, L., Cheng, Q., Long, X., & Yuan, Y. (2020). Hyperspectral and Multispectral Remote Sensing Image Fusion Based on Endmember Spatial Information. *Remote Sensing*, *12*(6), 1009. https://doi.org/10.3390/rs12061009

[8] Hermessi, H., Mourali, O., & Zagrouba, E. (2021). Multimodal medical image fusion review: Theoretical background and recent advances. *Signal Processing*, *183*, 108036. https://doi.org/10.1016/j.sigpro.2021.108036

[9] Hoffer, E., & Ailon, N. (2018). *Deep metric learning using Triplet network* (arXiv:1412.6622). arXiv. http://arxiv.org/abs/1412.6622

[10] Huang, B., Yang, F., Yin, M., Mo, X., & Zhong, C. (2020). A Review of Multimodal Medical Image Fusion Techniques. *Computational and Mathematical Methods in Medicine*, *2020*, 1–16. https://doi.org/10.1155/2020/8279342

[11] Huang, J., Le, Z., Ma, Y., Mei, X., & Fan, F. (2020). A generative adversarial network with adaptive constraints for multi-focus image fusion. *Neural Computing and Applications*, *32*(18), 15119–15129. https://doi.org/10.1007/s00521-020-04863-1

[12] Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., & Pedrycz, W. (2024). A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications*, *241*, 122666. https://doi.org/10.1016/j.eswa.2023.122666

[13] J. Zhang, L. Jiao, W. Ma, F. Liu, X. Liu, L. Li, P. Chen, & S. Yang. (2023). Transformer Based Conditional GAN for Multimodal Image Fusion. *IEEE Transactions on Multimedia*, *25*, 8988–9001. https://doi.org/10.1109/TMM.2023.3243659

[14] Kalamkar, S., & A., G. M. (2023). Multimodal image fusion: A systematic review. *Decision Analytics Journal*, *9*, 100327. https://doi.org/10.1016/j.dajour.2023.100327

[15] Kaur, M., & Singh, D. (2020). Fusion of medical images using deep belief networks. *Cluster Computing*, *23*(2), 1439–1453. https://doi.org/10.1007/s10586-019-02999-x

[16] Kaur, M., & Singh, D. (2021). Multi-modality medical image fusion technique using multi-objective differential evolution based deep neural networks. *Journal of Ambient Intelligence and*

Humanized Computing, *12*(2), 2483–2493. https://doi.org/10.1007/s12652-020-02386-0

[17] Krishnamoorthy, S., & Soman, K. P. (2010). Implementation and Comparative Study of Image Fusion Algorithms. *International Journal of Computer Applications*, *9*(2), 25–35. https://doi.org/10.5120/1357-1832

[18] Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., & Chanussot, J. (2022a). Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, *112*, 102926. https://doi.org/10.1016/j.jag.2022.102926

[19] Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., & Chanussot, J. (2022b). Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, *112*, 102926. https://doi.org/10.1016/j.jag.2022.102926

[20] Li, J., Zhu, G., Hua, C., Feng, M., BasheerBennamoun, Li, P., Lu, X., Song, J., Shen, P., Xu, X., Mei, L., Zhang, L., Shah, S. A. A., & Bennamoun, M. (2021). *A Systematic Collection of Medical Image Datasets for Deep Learning* (arXiv:2106.12864). arXiv. http://arxiv.org/abs/2106.12864

[21] Li, Y., Daho, M. E. H., Conze, P.-H., Zeghlache, R., Boité, H. L., Tadayoni, R., Cochener, B., Lamard, M., & Quellec, G. (2024). *A review of deep learning-based information fusion techniques for multimodal medical image classification* (arXiv:2404.15022). arXiv. http://arxiv.org/abs/2404.15022

[22] Li, Y., Ren, J., Liu, J., & Chang, Y. (2021). Deep sparse autoencoder prediction model based on adversarial learning for cross-domain recommendations. *Knowledge-Based Systems*, *220*, 106948. https://doi.org/10.1016/j.knosys.2021.106948

[23] Li, Y., Zhao, J., Lv, Z., & Li, J. (2021). Medical image fusion method by deep learning. *International Journal of Cognitive Computing in Engineering*, *2*, 21–29. https://doi.org/10.1016/j.ijcce.2020.12.004

[24] Mi, J., Wang, L., Liu, Y., & Zhang, J. (2022). KDE-GAN: A multimodal medical image-fusion model based on knowledge distillation and explainable AI modules. *Computers in Biology and Medicine*, *151*, 106273. https://doi.org/10.1016/j.compbiomed.2022.106273

[25] Mishra, D., & Palkar, B. (2015). Image Fusion Techniques: A Review. *International Journal of Computer Applications*, *130*(9), 7–13. https://doi.org/10.5120/ijca2015907084

[26] Ortiz-Perez, D., Ruiz-Ponce, P., Tomás, D., Garcia-Rodriguez, J., Vizcaya-Moreno, M. F., & Leo, M. (2023). A Deep Learning-Based Multimodal Architecture to predict Signs of Dementia. *Neurocomputing*, *548*, 126413. https://doi.org/10.1016/j.neucom.2023.126413

[27] Princess, M. R., Kumar, V. S., & Begum, M. R. (2014). Comprehensive and Comparative Study of Different Image Fusion Techniques. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, *03*(09), 11800–11806. https://doi.org/10.15662/ijareeie.2014.0309015

[28] Sara, D., Mandava, A. K., Kumar, A., Duela, S., & Jude, A. (2021). Hyperspectral and multispectral image fusion techniques for high resolution applications: A review. *Earth Science Informatics*, *14*(4), 1685–1705. https://doi.org/10.1007/s12145-021-00621-6

[29] Serrano, N., & Bellogín, A. (2023). Siamese neural networks in recommendation. *Neural Computing and Applications*, *35*(19), 13941–13953. https://doi.org/10.1007/s00521-023-08610-0

[30] Shetty D S, R. (2023). Multi-Modal Fusion Techniques in Deep Learning. *International Journal of Science and Research (IJSR)*, *12*(9), 526–532. https://doi.org/10.21275/SR23905100554

[31] Stahlschmidt, S. R., Ulfenborg, B., & Synnergren, J. (2022). Multimodal deep learning for biomedical data fusion: A review. *Briefings in Bioinformatics*, *23*(2), bbab569. https://doi.org/10.1093/bib/bbab569

[32] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

[33] Wang, L., Zhang, J., Liu, Y., Mi, J., & Zhang, J. (2021). Multimodal Medical Image Fusion Based on Gabor Representation Combination of Multi-CNN and Fuzzy Neural Network. *IEEE Access*, *9*, 67634–67647. https://doi.org/10.1109/ACCESS.2021.3075953

[34] Younesi, A., Ansari, M., Fazli, M., Ejlali, A., Shafique, M., & Henkel, J. (2024). *A Comprehensive Survey of Convolutions in Deep Learning: Applications, Challenges, and Future Trends* (arXiv:2402.15490). arXiv. http://arxiv.org/abs/2402.15490

[35] Zhang, C., Sjarif, N. N. A., & Ibrahim, R. (2024). Deep learning models for price forecasting of financial time series: A review of recent advancements: 2020–2022. *WIREs Data Mining and Knowledge Discovery*, *14*(1), e1519. https://doi.org/10.1002/widm.1519

[36] Zhang, H., & Qie, Y. (2023). Applying Deep Learning to Medical Imaging: A Review. *Applied Sciences*, *13*(18), 10521. https://doi.org/10.3390/app131810521