Decision Analytics Journal 9 (2023) 100327

Contents lists available at ScienceDirect

Decision Analytics Journal

journal homepage: www.elsevier.com/locate/dajour

Multimodal image fusion: A systematic review

Shrida Kalamkar, Geetha Mary A. [*]

VIT, Vellore, India

A R T I C L E I N F O

Keywords:
Multimodal data
Image fusion
Deep learning
Transformers
Wavelet transforms

A B S T R A C T

Multimodal image fusion combines information from multiple modalities to generate a composite image containing complementary information. Multimodal image fusion is challenging due to the heterogeneous nature of data, misalignment and nonlinear relationships between input data, or incomplete data during the fusion process. In recent years, several attention mechanisms have been introduced to enhance the performance of deep learning models. However, little literature is available on multimodal image fusion using attention mechanisms. This paper aims to study and analyze the latest deep-learning approaches, including attention mechanisms for multimodal image fusion. As a result of this study, the graphical taxonomy based on the different image modalities, various fusion strategies, fusion levels, and metrics for fusion tasks has been put forth. The focus has been on various Multimodal image fusion frameworks based on deep-learning techniques as their core methodology. This paper also sheds light on the challenges and future research directions in this field, application domains, and benchmark datasets used for multimodal fusion tasks. This paper contributes to the research on Multimodal image fusion and can help researchers select a suitable methodology for their applications.

1. Introduction

Imagine a scenario where a machine learning model can combine information from multiple modalities such as RGB, thermal, and depth images to detect objects and events with higher accuracy and reliability. Multimodal image fusion can create intelligent systems that identify objects in challenging environments and take appropriate actions in real-time. For example, in medical imaging, multimodal image fusion can help diagnose and treat complex diseases by combining information from different imaging modalities such as computerized tomography (CT) and Magnetic resonance imaging (MRI) scans [1,2]. In remote sensing, the fusion of data from sensors such as radar, lidar, and optical cameras can provide a more comprehensive understanding of the Earth's surface [3]. Although multimodal image fusion is still an emerging technology, recent advances in deep learning and other machine learning approaches have shown promising results (Y. Li, Zhao, Lv, and Li 2021). Combining information from different modalities can improve the accuracy and robustness of various computer vision tasks, such as object recognition, image segmentation, and scene understanding.

Soon, multimodal image fusion is expected to play a critical role in various applications, such as autonomous driving, surveillance systems, and robotics [4]. Integrating information from multiple sources allows these systems to operate in complex and dynamic environments, making them more efficient, reliable, and safe. The progress made in multimodal image fusion technology is accelerating toward a new era of intelligent machines that can work seamlessly with humans, enhancing our capabilities and improving our quality of life [5]. Multimodal image fusion has been used in robotics, remote sensing, surveillance, and medical imaging because of its potential application. Multimodal image fusion can provide a more thorough and accurate image of the scene by merging several features of imaging data, which helps with decision-making and improves end-to-end system performance [6]. Although challenges are still to be overcome, such as dealing with varying illumination conditions and sensor noise, the progress made in multimodal image fusion drives us closer to a future where autonomous systems can operate seamlessly and efficiently without human intervention [7]. While the technology for multimodal image fusion is still nascent, the progress made in deep learning and other non-deep learning approaches is promising. Deep learning methods, such as convolutional neural networks (CNNs), generative adversarial networks (GANs), and Transformer [2,8] based approaches have shown great potential for improving the quality and robustness of multimodal image fusion. Non-deep learning approaches, such as sparse representation-based methods and multi-scale and multi-feature decomposition techniques, have also been used successfully in multimodal image fusion tasks [9]. Multimodal image fusion has become a critical research domain [10,11] in various fields, including medical imaging, remote sensing, robotics, and autonomous driving. Multimodal imaging

[*] Corresponding author.
E-mail addresses: shrida88@gmail.com (S. Kalamkar), geethamary.a@gmail.com (Geetha Mary A.).

https://doi.org/10.1016/j.dajour.2023.100327
Received 10 August 2023; Received in revised form 1 September 2023; Accepted 15 September 2023
Available online 21 September 2023

systems, such as MRI, CT, and PET, provide complementary information that can be fused to provide a more comprehensive understanding of the scene.

Additionally, multimodal imaging data can improve decision-making and enhance system performance. The need for multimodal image fusion technology is increasing rapidly due to its broad applicability in various domains. For example, in medical imaging, multimodal image fusion can help improve diagnosis and treatment planning by combining information from different imaging modalities [12]. In remote sensing, multimodal image fusion can help enhance the quality of image and accuracy by combining information from different sensors [3,13]. Furthermore, using multimodal image fusion technology in robotics and autonomous driving can improve object detection, scene understanding, and decision-making. By combining information from different sensors, such as cameras and lidar, multimodal image fusion can provide a more accurate and comprehensive understanding of the environment [14]. The increasing demand for multimodal image fusion technology drives research and development efforts in academia and industry (Mohanad G. Yaseen, Mohammad Naeemullah, and Ibarhim Adeb Mansoor 2021). The market for multimodal image fusion technology is expected to proliferate in the coming years due to the increasing demand for improved system performance and the growing availability of imaging data [15].

Over the past few years, multimodal image fusion has become an important research domain in computer vision and related fields. Many research papers have been published on this topic, covering different aspects of multimodal image fusion, such as fusion techniques, performance evaluation, and applications. However, there is still a need for a systematic review of the progress made by state-of-the-art deep learning techniques in multimodal image fusion. Existing review papers have covered various aspects of multimodal image fusion, but this task lacks an in-depth analysis of deep learning approaches using attention mechanisms. Therefore, this survey paper aims to thoroughly analyze the progress made by deep learning techniques in multimodal image fusion, including various fusion architectures, loss functions, and evaluation metrics. This paper proposes a graphical taxonomy, which is a valuable addition, aiding readers in visualizing the relationships between image modalities, fusion strategies, levels of fusion, and evaluation metrics.

The contribution of this survey paper is to provide a systematic review of deep learning approaches for multimodal image fusion and to identify the current state-of-the-art techniques in this field. This review will be a valuable resource for researchers and practitioners working on multimodal image fusion and help guide future research. This is the first-ever systematic survey of multimodal image fusion addressing the fusion architecture using attention mechanisms.

The work is contributed as follows:

- The graphical taxonomy of multimodal image fusion has been outlined systematically, covering all factors.
- A thorough survey of deep learning approaches for multimodal image fusion is done.
- The trade-offs in multimodal image fusion are discussed from the perspective of traditional approaches and different deep-learning approaches.
- Datasets with different modalities have been listed, and the needs and issues of multimodal image fusion are explored.
- The current challenges, application domains, and possible future directions in the domain of deep learning applicable to multimodal image fusion are thoroughly put forth.

The roadmap of the paper is outlined in Fig. 1. The taxonomy of the multimodal image fusion task is outlined in Section 2. The topics of traditional and deep learning methods for multimodal image fusion are covered in Sections 3 and 4. Sections 6 and 7 briefly overview application domains and benchmarked datasets. Section 8 outlines the challenges in research and potential approaches for multimodal image fusion. The paper is concluded in Section 9.

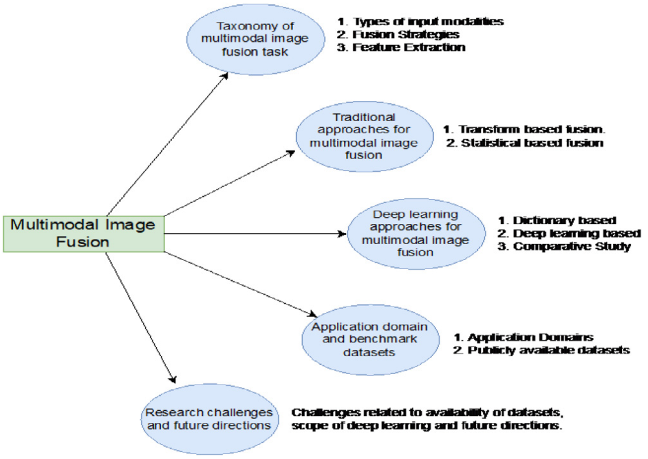
Multimodal Image Fusion
Taxonomy of multimodal image fusion task
1. Types of input modalities
2. Fusion Strategies
3. Feature Extraction
Traditional approaches for multimodal image fusion
1. Transform based fusion.
2. Statistical based fusion
Deep learning approaches for multimodal image fusion
1. Dictionary based
2. Deep learning based
3. Comparative Study
Application domain and benchmark datasets
1. Application Domains
2 Publicly available datasets
Research challenges and future directions
Challenges related to availability of datasets, scope of deep learning and future directions.

Fig. 1. Roadmap of topics discussed in the paper.

2. Taxonomy of multimodal image fusion task

Multimodal image fusion combines multiple images acquired from different imaging modalities into a single fused image that preserves the most useful and relevant information from each input modality. Multimodal image fusion aims to provide a more comprehensive and accurate representation of the underlying object or scene being imaged than can be obtained from any single modality alone. Fig. 2 depicts the taxonomy of multimodal image fusion tasks. The taxonomy is based on various factors to be considered for the fusion of multimodal images.

2.1. Input modalities

2.1.1. Bi-spectral fusion (two modalities)

Bi-spectral fusion is a type of multimodal image fusion that combines two different modalities, typically visible and infrared images, to provide more comprehensive information about the scene or object being observed. Visible images are sensitive to the scene's colors and brightness, while infrared images are sensitive to the scene's temperature. Combining these two modalities allows bi-spectral fusion to provide more information about the scene, especially in low-light or night-time conditions where visible light is limited. This can be useful in various applications, such as surveillance, navigation, and remote sensing.

For Example, Suppose we want to monitor a forest at night to detect potential wildfires. We can use a visible camera to capture images of the forest during the day and an infrared camera to capture images at night when there is no visible light. However, visible images alone may not detect potential fires, as smoke may not be visible during the day. Fire can be difficult to detect at night using only infrared images.

By combining the visible and infrared images using bi-spectral fusion, we can create a more comprehensive image that combines the color and brightness information from the visible

Image with the temperature information from the infrared image. This can help us to detect potential fires more accurately, as the combination of color, brightness, and temperature information can help to distinguish fires from other heat sources and detect smoke even during the day.

Authors Agrawal and Karar [16] proposed a multi-level fusion algorithm based on curvelet transform for the fusion of registered visible and infrared images of the same scene. Zhou et al. [17] propose a novel enhanced spectral fusion network for hyperspectral image classification. Based on the fusion of different spectral strides, the model is divided into two parts: an optimized multi-scale fused spectral attention module (FsSE) and a 3D convolutional neural network (3D CNN).

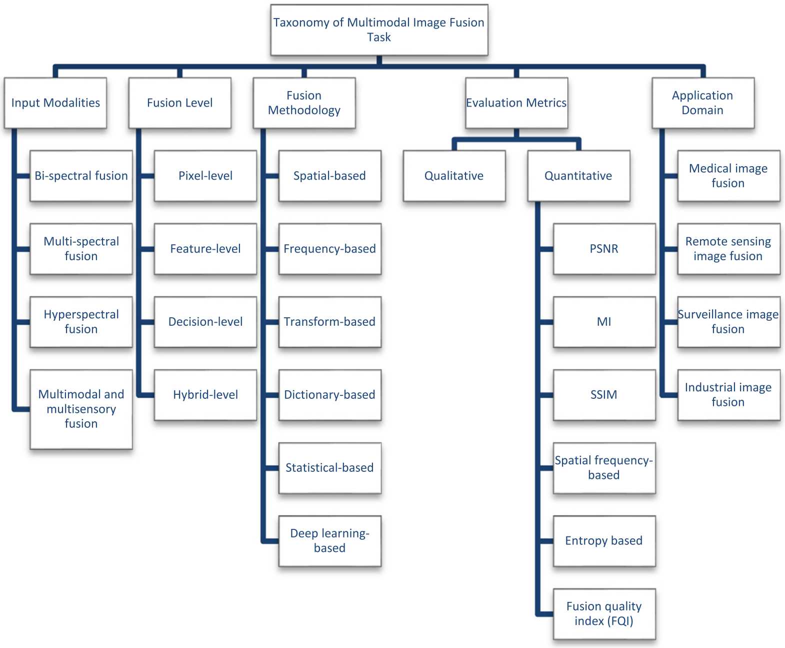

Fig. 2. Taxonomy of multimodal image fusion task.

Bi-spectral fusion can be particularly useful in low-light or nighttime conditions, where visible light is limited and infrared images can provide important information. However, fusion methods such as multispectral or hyperspectral may be more effective in other conditions or applications where other modalities may provide more relevant information.

2.1.2. *Multi-spectral fusion (more than two modalities)*

Multi-spectral fusion is a multimodal image fusion involving combining more than two spectral bands or modalities, typically in the range of 3–10 bands. Multi-spectral fusion aims to produce a fused image that contains the most informative features from each spectral band while eliminating redundant and irrelevant information.

Multi-spectral fusion is commonly used in remote sensing applications, where multiple spectral bands are often used to capture different aspects of the scene or object being observed. For example, in satellite imaging, multi-spectral sensors may capture images in the visible, near-infrared, and thermal infrared bands, each providing unique information about the land cover, vegetation, and surface temperature.

Various fusion methods can perform multi-spectral fusion, including feature, decision, and image-level fusion. Feature-level fusion involves extracting features from each spectral band and combining them into a single feature vector, which is then used to generate the fused image. Decision-level fusion involves making independent decisions based on each spectral band and combining them to make a final decision. Image-level fusion directly combines the images from each spectral band to create a single fused image.

Overall, multi-spectral fusion can be a powerful tool for improving the accuracy and effectiveness of image analysis in remote sensing and other applications by combining the unique information provided by multiple spectral bands into a single, comprehensive image.

Authors Guo et al. [18] proposed a coupled non-negative block-term tensor model to estimate the ideal high spatial resolution hyperspectral images. L1-norm characterizes the sparsity of the image, and total variation (TV) is introduced to describe piecewise smoothness. The proximal alternating optimization (PAO) algorithm and the alternating multiplier method are used to solve the model.

In Feng et al. [19] authors propose a model-based deep learning approach for merging a High resolution Multi-Spectral (HrHS) and Low-resolution Multi-Spectral (LrHS) images to generate a high-resolution hyperspectral (HrHS) image. An iterative algorithm to solve the model by exploiting the proximal gradient method is discussed.

2.1.3. *Hyperspectral fusion (images with high spectral resolution)*

Hyperspectral fusion is a type of multimodal image fusion that combines images with high spectral resolution, typically consisting of hundreds of spectral bands. Hyperspectral sensors can capture images at very fine spectral intervals, providing detailed information about the objects' material composition and physical properties in the scene [20].

Hyperspectral fusion aims to generate a single fused image containing the most informative features from each spectral band, eliminating noise and other artifacts. Hyperspectral fusion can be useful in various applications, including agriculture, environmental monitoring, and mineral exploration. To perform hyperspectral fusion, various fusion techniques can be used, such as principal component analysis (PCA), independent component analysis (ICA), and non-negative matrix factorization (NMF). PCA and ICA are popular techniques for hyperspectral feature extraction, while NMF is used for hyperspectral unmixing, which involves separating the spectral signatures of different materials in the scene.

Hyperspectral fusion can be computationally intensive and require significant processing power and resources.

Typically, hyperspectral fusion can be found in agriculture, where remote sensing data monitors crop health and yield. Hyperspectral sensors can capture detailed information about the reflectance and absorption of light by plant leaves, which can be used to detect stress, disease, and nutrient deficiencies.

Consider a scenario where researchers want to analyze the health of wheat crops in a field. So, we can combine data from a hyperspectral camera and a multispectral camera to generate a single, high-resolution image containing spectral and spatial information about the crops. Further sparse unmixing can extract the spectral signatures of different crop types and soil backgrounds from the hyperspectral data and then fuse the results with the multispectral data using guided filtering. The fused image can improve the classification accuracy and spatial resolution compared to the individual images, enabling the researchers to accurately identify field areas with healthy and stressed crops [21].

Multispectral imaging involves capturing images in a few discrete spectral bands, typically ranging from three to ten bands. These images are often used to generate false-color composites that can highlight specific features or properties of the scene, such as vegetation health or water content.

In contrast, hyperspectral imaging involves capturing images at many more spectral bands, often ranging from dozens to hundreds of bands. These images can provide much more detailed information about the spectral reflectance properties of the scene, enabling more accurate analysis and classification of different materials and features.

2.1.4. Multimodal and multisensory fusion (fusion of images from multiple modalities and sensors)

Multimodal and multisensory fusion refers to combining information from multiple sources, such as different types of sensors or modalities, to improve the accuracy and reliability of the resulting information [8,22].

For example, in image processing, multimodal and multisensory fusion can combine images obtained from sensors, such as visible light and thermal cameras, to produce a more complete and accurate scene representation [23].

The fusion process typically involves several steps, including preprocessing, feature extraction, and fusion. Preprocessing involves preparing the data from each source for fusion, such as aligning the images and correcting for sensor noise or other artifacts. Feature extraction involves identifying relevant features or patterns in the data, such as edges, textures, or object shapes. Finally, fusion involves combining the features from each source to produce a fused representation.

Multimodal and multisensory fusion has applications in many fields, including remote sensing, medical imaging, robotics, and surveillance. It can also be used in human–computer interfaces, such as for gesture recognition or speech recognition, where multiple sensors or modalities are used to improve the accuracy and robustness of the system. A detailed review of work done using multimodal and multisensory fusion is discussed in Table 5.

2.2. Complexity and time factor for fusion

The complexity and time required for image fusion can vary significantly depending on the specific fusion type [24].

The complexity of bi-spectral fusion is relatively low compared to other types of fusion, as only two modalities are involved. Due to the lower complexity, the time required for bi-spectral fusion is also typically shorter than other fusion types.

The complexity of multi-spectral fusion is higher than bi-spectral fusion due to the increased number of fused modalities. However, the time required for multi-spectral fusion may be comparable to or slightly longer than bi-spectral fusion, depending on the specific algorithm used.

Hyperspectral fusion involves the fusion of images with high spectral resolution, typically containing hundreds of narrow spectral bands. Hyperspectral fusion's complexity is higher than bi-spectral and multi-spectral fusion due to the large number of fused spectral bands. The time required for hyperspectral fusion is also typically longer than both bi-spectral and multi-spectral fusion due to the increased complexity of the fusion algorithm.

The complexity and time required for multimodal and multisensory fusion can vary significantly depending on the fused modalities and sensors and the specific fusion algorithm used. Generally, the complexity and time required for multimodal and multisensory fusion are higher than other types of fusion due to the increased number of modalities and sensors being fused.

In summary, the complexity and time required for image fusion depend on the specific type of fusion being performed, with the highest complexity and time required for hyperspectral fusion and multimodal and multisensory fusion, respectively. This is well shown in Fig. 3.

However, the complexity and time required for each type of fusion will depend on the specific fusion algorithm used and the quality and size of the input image.

2.3. Fusion levels

2.3.1. Pixel-level fusion (fusion of individual pixels from input images)

Pixel-level fusion involves fusing individual pixels from multiple input images to create a new output image. This type of fusion is commonly used in applications such as remote sensing, medical imaging, and surveillance.

One of the most common approaches to pixel-level fusion is combining the input images' pixel values using weighted averaging or other mathematical operations. The weights assigned to each pixel can be determined based on various criteria, such as image quality, spatial context, or spectral characteristics [25]. More advanced fusion algorithms, such as those based on machine learning techniques, can also be used for pixel-level fusion. These algorithms can learn to combine the pixel values from the input images in a way that optimizes a specific objective, such as image quality or feature detection.

For instance, convolutional neural networks (CNNs) have been used for fusing visible and infrared images in remote sensing applications. These networks can learn to extract features from the input images and combine them to enhance the fused image's quality and information content.

In the research context, pixel-level fusion has been studied extensively in various domains, such as remote sensing, medical imaging, and computer vision.

One area of research in pixel-level fusion is the development of advanced algorithms that can learn to combine the pixel values of input images to optimize a specific objective.

Another area of research in pixel-level fusion is investigating fusion methods that can preserve the spatial and spectral information of the input images. For example, spatial–spectral fusion methods have been proposed for hyperspectral imaging to fuse high-resolution spectral information with high-resolution spatial information. Such methods can preserve the spectral and spatial characteristics of the input images and produce fused images that contain more information than each input image.

2.3.2. Feature-level fusion (fusion of features extracted from input images)

Feature-level fusion is a technique for combining information from multiple input images by fusing features that are extracted from these images. The main idea behind feature-level fusion is to extract relevant information from each input image and combine these features to create a new output image with enhanced features or improved quality.

Various techniques such as edge detection, texture analysis, or object recognition extract features from input images. These features are combined to create a new set representing the information in all input images. The new set of features can be used to create a fused image that contains more information and is more robust than each input image [22].

One of the key advantages of feature-level fusion is that it can be used to fuse images with different modalities, such as visible and infrared images, or images with different spatial resolutions. This is because the features extracted from each input image can be combined in a way that preserves the unique characteristics of each image while still providing a complete representation of the underlying scene.

Feature-level fusion can also reduce the impact of noise and other artifacts in the input images. By combining features from multiple images, the fused image can be less affected by noise or other anomalies in any input image.

In literature, feature-level fusion has been studied extensively, and various algorithms have been proposed to improve its performance. One approach is to use machine learning techniques to learn the optimal way to combine features. Another approach is to use statistical techniques such as principal component analysis or independent component analysis to identify the most informative features for fusion [26].

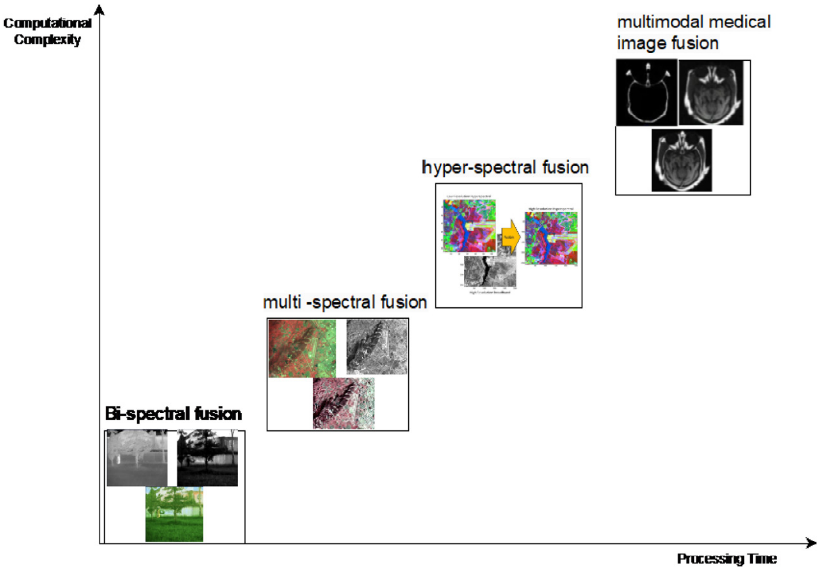

Fig. 3. Complexity vs. Processing time for different modalities of image.

2.3.3. Decision-level fusion (fusion of decision-level information or classification results from input images)

Decision-level fusion is a technique used in image processing and computer vision to combine decisions or classification results from multiple input images to produce a final output decision or classification. Decision-level fusion aims to improve the accuracy and robustness of the classification or decision-making process by incorporating information from multiple sources.

In decision-level fusion, each input image is classified independently using a classifier such as a neural network or support vector machine. The resulting classification decisions are combined using a fusion rule to produce a final decision or classification. The fusion rule can be based on various methods such as weighted averaging, majority voting, or Dempster–Shafer theory [6,27].

For example, in a surveillance application, decision-level fusion can identify a person or object by combining classification decisions from multiple cameras with different viewpoints. Each camera independently produces a classification decision based on the images captured, and these decisions are combined using a fusion rule to produce a final classification decision.

Another example of decision-level fusion is in medical image analysis. In medical image analysis, multiple imaging modalities such as MRI, CT, and ultrasound may be used to diagnose a disease or condition. Each imaging modality produces a classification decision based on the features extracted from the image, and these decisions are combined using a fusion rule to produce a final diagnosis [1].

One advantage of decision-level fusion is that it can be used with any feature extraction or image processing technique as long as a classifier can be trained on the extracted features. In addition, decision-level fusion is relatively simple to implement and can be applied to various applications.

However, decision-level fusion can be sensitive to the quality and reliability of the classification decisions from each input image. The fused decision may also be noisy or unreliable if the classification decisions are noisy or unreliable. Therefore, careful consideration must be given to selecting and training the classifiers used in decision-level fusion.

2.3.4. Hybrid-level fusion (combination of multiple levels of fusion)

Hybrid-level fusion is a technique used in image processing and computer vision that combines multiple fusion levels to produce a final output image. Hybrid-level fusion combines the advantages of different fusion levels to achieve improved quality, accuracy, and robustness performance.

Hybrid-level fusion typically combines pixel-level, feature-level, and decision-level fusion techniques. The input images are first preprocessed to extract pixel-level features such as color, texture, and shape. These features are then used to extract higher-level features such as edges, corners, and regions of interest. The higher-level features are combined using decision-level fusion techniques to produce a final output image [28].

For example, hybrid-level fusion can combine information from multiple input images with different lighting conditions, viewpoints, and facial expressions in a face recognition system. The pixel-level features such as color and texture are extracted from each input image, and these features are combined using feature-level fusion techniques such as Principal Component Analysis (PCA) or Independent Component Analysis (ICA) to extract higher-level features such as facial landmarks and facial expressions [28,29]. The higher-level features are combined using decision-level fusion techniques such as weighted averaging or fuzzy logic to produce a final recognition decision.

Hybrid-level fusion offers several advantages over individual levels of fusion. By combining multiple fusion levels, hybrid-level fusion can overcome the limitations and drawbacks of individual fusion levels, such as the lack of contextual information in pixel-level fusion or the sensitivity to noisy or unreliable classification decisions in decision-level fusion. Hybrid-level fusion can also provide greater flexibility in designing and optimizing fusion algorithms by allowing for the selection and combination of different fusion techniques at each level.

However, hybrid-level fusion is more complex and computationally expensive than individual fusion levels, requiring implementing and integrating multiple fusion techniques at each level. Therefore, careful consideration must be given to hybrid-level fusion's computational and memory requirements and the trade-offs between performance and complexity. Table 1 provides a summary of all four strategies, focusing on the advantages, disadvantages, and applicability of these strategies.

3. Traditional approaches for multimodal image fusion

Traditional approaches for multimodal image fusion refer to a set of techniques and methods developed over the years for integrating information from multiple sources of images. These approaches are typically based on mathematical or statistical models and involve extracting features from the input images and fusing them to generate a single output image.

Table 1
Comparative summary of multimodal image fusion strategies.

Fusion level	Description	Advantages	Disadvantages	Suitable applications
Pixel-level fusion	Combines input images at the pixel level by using mathematical operations such as averaging or maximum selection	Simple and easy to implement, preserves spatial information, good for low-level image processing tasks such as noise reduction or contrast enhancement	Sensitive to misregistration, can be affected by outliers or noise, limited by the spatial resolution and quality of the input images	Image denoising, image sharpening, image enhancement, low-level feature extraction
Feature-level fusion	Extracts feature from input images such as edges, textures, or shapes and combine them using mathematical operations or statistical models	Robust to misregistration, preserves spatial and spectral information, suitable for medium-level image processing tasks such as object recognition or tracking	Requires careful feature selection and extraction, may not capture all relevant information, can be computationally expensive depending on the number and complexity of features	Object recognition, image segmentation, object tracking, texture analysis
Decision-level fusion	Combines classification results from multiple input images using mathematical operations such as majority voting or weighted averaging	Robust to misregistration and noise, suitable for high-level image processing tasks such as object classification or scene recognition	Sensitive to the quality and reliability of the classification results from each input image, may not capture all relevant information, can be affected by the number and diversity of input images	Object classification, scene recognition, face recognition, medical diagnosis
Hybrid-level fusion	Combines multiple levels of fusion, such as pixel-level, feature-level, and decision-level fusion, to achieve improved performance in terms of quality, accuracy, and robustness	Combines the advantages of different levels of fusion, provides greater flexibility in designing and optimizing fusion algorithms, suitable for complex and challenging image processing tasks such as remote sensing or medical imaging	More complex and computationally expensive than individual levels of fusion, requires careful consideration of the computational and memory requirements, may require the integration of multiple fusion techniques at each level depending on the task	Remote sensing, medical imaging, surveillance, robotics, autonomous vehicles, quality control in manufacturing and inspection, multimedia applications such as video and image editing

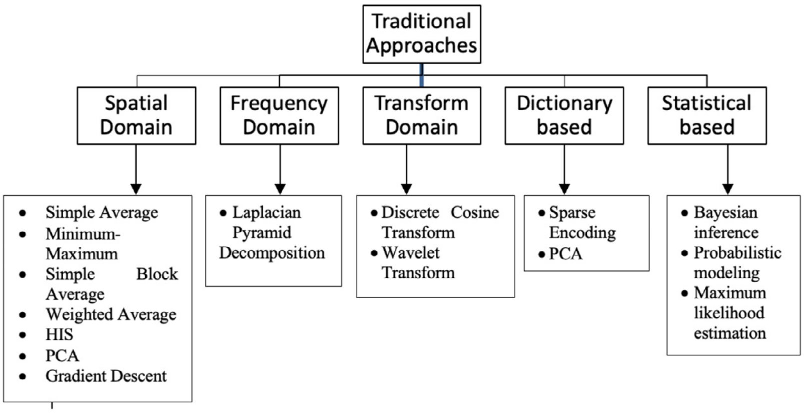

Fig. 4. Typical Algorithms used in traditional image fusion approaches.

Transform-based fusion, dictionary-based fusion, and statistical-based fusion are examples of traditional approaches that have been widely used in the literature. These traditional approaches have been used in various medical imaging, remote sensing, surveillance, and industrial image processing applications. While they have shown promising results, they also have limitations, such as high computational complexity and limited robustness to noise and artifacts. As a result, recent research has focused on developing deep learning-based approaches that can learn the fusion process directly from the input data, overcoming some of the limitations of traditional approaches.

This paper focuses more on advancements in deep learning-based fusion techniques, so just an overview of traditional methods has been put forth in this paper. More references available in Agrawal and Karar [16], Rao et al. [30], Nejati et al. [31], Sun et al. [32], Bhujle [33], He et al. [34] and Kaur and Singh [35] (see Fig. 4).

4. Deep learning-based multimodal image fusion (CNN, Auto-encoders, GAN, transformers)

The motivation for introducing deep learning into image fusion is to overcome the limitations of traditional methods [31,35].

Deep learning has become a popular technique for multimodal image fusion due to its ability to automatically learn complex mappings between different modalities and efficiently handle large amounts of data.

4.1. Multimodal image fusion using CNN

CNNs are widely used in image processing tasks, including multimodal image fusion [36]. CNNs can automatically learn spatial features from the input images, which can be used to create a fused representation by extracting the relevant information from the source image [37–39]. Fig. 5 shows a general architecture for multimodal image fusion using CNN.

Typical steps involved in using CNNs for multimodal image fusion:

- Data preparation: The input images must be pre-processed and prepared for input into the CNN. This may involve resizing the images to a common size, normalizing the pixel values, and separating the input images into different channels.

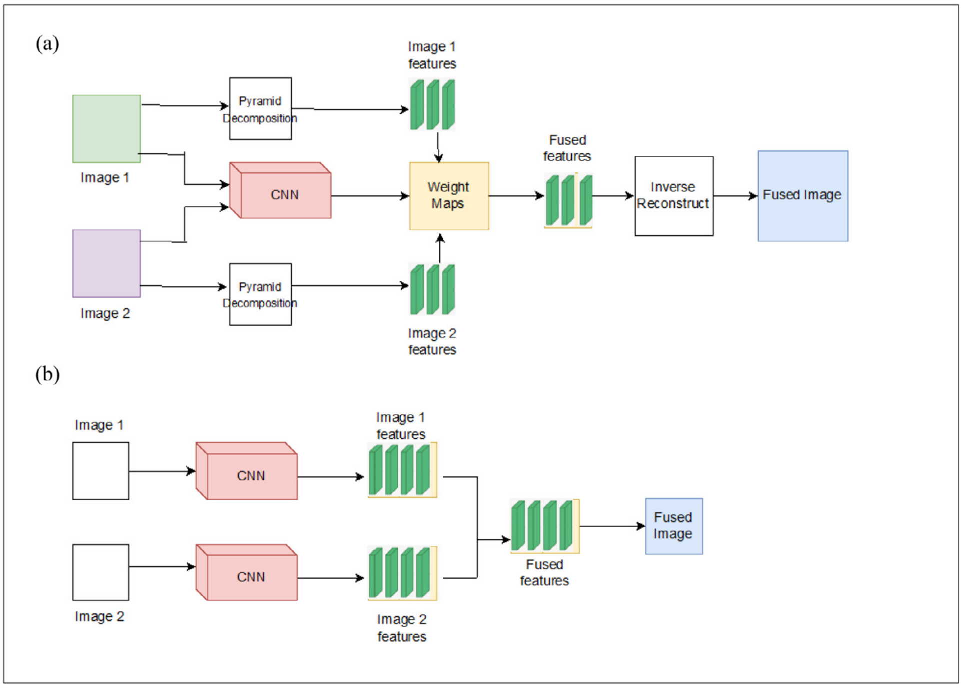

Fig. 5. Different CNN-based approaches for multimodal image fusion.

- CNN architecture selection: Next, an appropriate CNN architecture needs to be selected for the specific fusion task. Common CNN architectures for multimodal image fusion include the VGG, ResNet, and Inception models.
- Training the CNN: The selected CNN architecture must be trained on a dataset of paired input images and their corresponding fused images. The loss function is used during training to ensure that the fused image preserves the most relevant information from the input images.
- Fusing the input images: Once the CNN has been trained, it can fuse new pairs of input images. The input images are fed through the CNN, and the resulting feature maps from each input modality are combined to create a fused representation.
- Evaluation: Finally, the performance of the fused images needs to be evaluated using metrics such as peak signal-to-noise ratio (PSNR), SSIM, and visual inspection. The CNN can be fine-tuned based on the evaluation results to improve the quality of the fused images.

More details of CNN-based multimodal image fusion architectures are provided in Table 2.

4.2. Multimodal image fusion using Auto-encoders

Auto-encoders are another type of deep learning model that can be used for multimodal image fusion. Auto-encoders can be used to perform feature-level fusion, where the input images are encoded into a lower-dimensional feature space and then decoded to create a fused representation [35,40,41]. Fig. 6 shows the general architecture of Multimodal Image Fusion using Stacked Auto-encoders. Here are the general steps involved in using auto-encoders for multimodal image fusion:

- Data preparation: The input images must be pre-processed and prepared for input into the auto-encoder. This may involve resizing the images to a common size, normalizing the pixel values, and separating the input images into different channels.
- Auto-encoder architecture selection: An appropriate auto-encoder architecture must be selected for the specific fusion task. The encoder network typically comprises convolutional and pooling layers, while the decoder network uses deconvolutional layers to reconstruct the input images.
- Training the auto-encoder: The auto-encoder is trained on a dataset of paired input images and their corresponding fused images. The loss function used during training typically combines mean squared error (MSE) and structural similarity index (SSIM) to ensure that the fused image preserves the most relevant information from the input images.
- Fusing the input images: Once the auto-encoder has been trained, it can fuse new pairs of input images. The encoder network encodes the input images into a lower-dimensional feature space. The resulting feature maps from each input modality are combined to create a fused representation. The fused image is then decoded using the decoder network to create the final fused image.

4.3. Multimodal image fusion using generative adversarial networks

Generative adversarial networks (GANs) have been used for image synthesis. They can also be used for multimodal image fusion, where the fused image is generated by combining the features from multiple input images [42–44]. Here are the general steps involved in using GANs for multimodal image fusion:

- Data preparation: The input images must be pre-processed and prepared for input into the GAN. This may involve resizing the images to a common size, normalizing the pixel values, and separating the input images into different channels.
- GAN architecture selection: An appropriate GAN architecture must be selected for the specific fusion task. The generator network typically comprises convolutional and deconvolutional layers, while the discriminator network distinguishes between real and fake images.

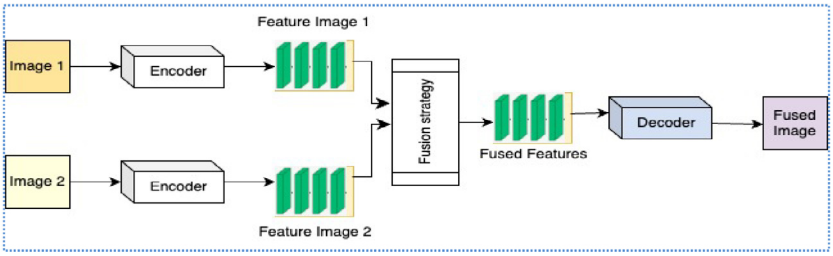

Fig. 6. Auto-encoder-based multimodal image fusion.

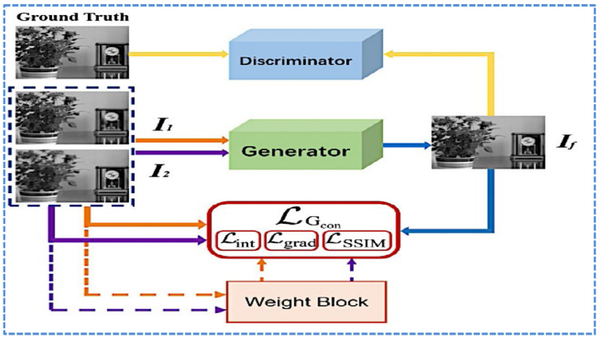

Fig. 7. Generative adversarial network-based multi-focus image fusion [11].

- Training the GAN: The GAN is trained on a dataset of paired input images and their corresponding fused images. The generator network is trained to produce a fused image similar to the ground truth fused image, while the discriminator network is trained to distinguish between real and fake images. The loss function used during training is typically a combination of adversarial and content loss to ensure that the fused image is realistic and preserves the most relevant information from the input images.
- Fusing the input images: Once the GAN has been trained, it can fuse new pairs of input images. The input images are fed through the generator network, and the resulting feature maps from each input modality are combined to create a fused representation. Fig. 7 shows a general GAN-based multi-focus image fusion.

4.4. Multimodal image fusion using transformers

Transformers are a type of deep learning model widely used in natural language processing (NLP), but they can also be applied to image fusion tasks [43,45; Tang et al. 2023; 2022]. Attention mechanisms can be used in multimodal image fusion to enable the model to focus on the most relevant features from each input modality. Different attention mechanisms can be used for fusion tasks.

1. Self-attention: Self-attention mechanisms can compute attention scores for each feature map within a single modality. This allows the model to focus on the most informative regions within each modality and can improve the overall quality of the fused image.
2. Cross-attention: Cross-attention mechanisms can compute attention scores between different modalities. This allows the model to identify the most informative features from each modality and combine them in the fused image. Cross-attention can be especially useful when fusing images with different resolutions or sensor types.
3. Multi-level attention: Multi-level attention mechanisms can be used to compute attention scores at multiple levels of abstraction. For example, the model could compute self-attention scores at the pixel level and higher levels of abstraction, such as object or scene level. This allows the model to capture fine-grained details and global context information in the fused image.
4. Channel-wise attention: Channel-wise attention mechanisms can compute attention scores for each channel within a single modality or between different modalities. This allows the model to identify the most informative channels and weigh them accordingly when fusing the input images.

Fig. 8 gives a generalized Transformer architecture for multimodal image fusion tasks.

Transformer-based multimodal image fusion represents a cutting-edge approach to computer vision. Researchers continue to explore and develop new techniques and architectures to improve the quality and efficiency of image fusion using Transformers.

In Zhang et al. [43], the authors have proposed a novel multimodal image fusion framework with a transformer-based conditional generative adversarial network (CGAN). It integrates the advantages of different fusion methods and feature representation to improve training efficiency and global cross-domain information interaction.

Ma et al. (2023) proposes a selectable Transformer and Gist CNN network (STGC-Net). It designs a subspace-similar recombination module (SSR-Module) based on non-negative matrix factorization (NMF) and the self-attention mechanism for feature decomposition. This can alleviate the redundant information of multimodal data and extract their singular and standard features.

Currently, Transformer architectures are not much explored for multimodal image fusion. Here is an overview of how Transformers can be applied to multimodal image fusion:

1. **Understanding Multimodal Image Fusion:**
   Multimodal image fusion involves combining information from multiple sources or modalities (e.g., visible light, infrared, depth) to create a single, more informative image. Each modality typically provides unique information, and the goal is to fuse this information to enhance the overall image quality or extract specific features.
2. **Transformers in Image Fusion:**
   Transformers are robust neural network architectures that have shown remarkable success in various tasks due to their ability to capture complex dependencies in data. In image fusion, Transformers can be used to learn the relationships and dependencies between the different modalities and create a fused representation.
3. **Input Representation:**
   The input to a Transformer-based multimodal image fusion system includes multiple source images from different modalities. Each source image is usually passed through a convolutional neural network (CNN) to extract image features. These features are then combined and passed as input to the Transformer model.
4. **Attention Mechanism:**
   Transformers leverage attention mechanisms to weigh the importance of different parts of the input features when making predictions. In the context of image fusion, the attention mechanism can highlight regions or features from each modality that are most relevant for fusion.

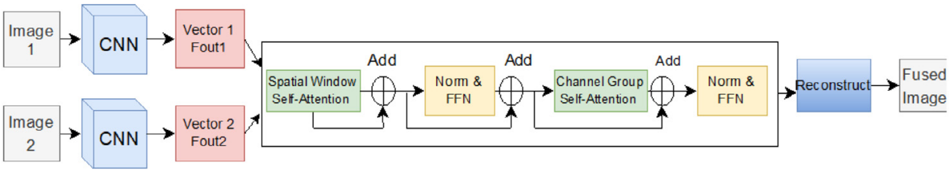

Fig. 8. Attention mechanism based Transformer architecture for image fusion [11].

5. **Transformer Architecture:**
   The Transformer architecture consists of multiple layers of self-attention and feedforward networks. These layers learn to model complex interactions between modalities, making them suitable for capturing the relationships between different information sources.
6. **Loss Function and Training:**
   The model is trained to minimize a loss function that measures the difference between the fused image and a ground truth or target image. Standard loss functions include mean squared error (MSE), perceptual loss, or other domain-specific loss functions.
7. **Evaluation:**
   The quality of the fused image can be evaluated using subjective (human perception) and objective (quantitative) metrics.
8. **Challenges:**
   Implementing Transformer-based image fusion models can be computationally expensive and may require significant computational resources. The choice of architecture, hyperparameters, and loss functions can also impact the model's performance.

Transformer architectures can improve the performance and quality of fused images in multimodal image fusion tasks through several mechanisms.

4.5. Comparative study

Table 2 shows the comparison of selected deep-learning approaches. The deep learning technique has achieved the highest performance for multimodal image fusion on the datasets. Though it is not verified from a single research article, deep learning has a great scope to excel in anomaly detection tasks.

5. Evaluation metrics

When evaluating multimodal image fusion, you can use subjective and objective metrics to comprehensively assess the quality of the fused images.

5.1. Qualitative metrics (subjective evaluation)

Qualitative metrics are used to evaluate the visual quality of the fused image by comparing it with the original input images.

The simplest way to evaluate the quality of a fused image is to visually inspect it and compare it with the original input images (H. Kaur, Koundal, and Kadyan 2021b). Visual inspection is subjective, but it can provide a quick and intuitive way to assess the quality of the fusion result. Contrast and brightness are important metrics for evaluating the quality of a fused image. A good fusion result should have similar contrast and brightness levels as the input images. Sharpness is another important metric for evaluating the quality of a fused image. A good fusion result should have sharp edges and details. A good fusion result should preserve the most relevant information from each input modality. The fused image should contain all the relevant information from both input images without losing any important details. Artifacts such as noise, blur, or ringing can occur in the fused image due to fusion. A good fusion result should have minimal artifacts. Color consistency is important for fused images that contain color information. A good fusion result should have consistent color across the image.

Overall, qualitative metrics are useful for evaluating the visual quality of the fused image and providing feedback on the performance of the fusion algorithm. However, they are subjective and may not always reflect the accuracy or reliability of the fusion result. Therefore, it is important to use quantitative metrics to evaluate the performance of multimodal image fusion algorithms.

5.2. Quantitative metrics (objective evaluation)

Quantitative metrics are used to objectively evaluate the performance of a multimodal image fusion algorithm. Here are some of the common quantitative metrics used for multimodal image fusion:

1. Mutual information (MI): MI is a statistical measure that quantifies the amount of shared information between the fused image and the input images. A good fusion result should have high mutual information with the input images.
2. Peak Signal-to-Noise Ratio (PSNR): PSNR is a metric that measures the quality of a fused image by comparing it with the original input images in terms of signal-to-noise ratio. A good fusion result should have a high PSNR value.
3. Structural Similarity Index Measure (SSIM): SSIM is a metric that measures the similarity between the fused image and the input images in terms of structural information. A good fusion result should have a high SSIM value.
4. Feature-based metrics: Feature-based metrics use features extracted from the input and fused images to evaluate the quality of the fusion result. Common feature-based metrics include edge-based metrics, texture-based metrics, and color-based metrics.
5. Entropy: Entropy is a metric that measures the amount of uncertainty or randomness in the fused image. A good fusion result should have a low entropy, indicating that the image contains more meaningful information.
6. Receiver Operating Characteristic (ROC) curve: The ROC curve is a graphical plot that evaluates the performance of a binary classifier by plotting the true positive rate against the false positive rate. ROC curve is commonly used in medical imaging to evaluate the performance of multimodal image fusion algorithms for disease detection.

Mathematical equations for the above metrics can be found in Kalamkar and A (2022).

Overall, quantitative metrics provide an objective way to evaluate the performance of a multimodal image fusion algorithm. However, different metrics may be more suitable for different applications, and the choice of metrics depends on the specific requirements of the task. It is important to use a combination of qualitative and quantitative metrics to evaluate the performance of multimodal image fusion algorithms comprehensively.

6. Application domain

Multimodal image fusion has a wide range of application domains across various fields.

Table 2
Survey of deep learning based multimodal image fusion approaches.

Author	Deep learning technique	Fusion level	Image modality	Fusion strategy	Evaluation metrics
Li et al. [37]	CNN	Feature	CT and MRI, MRI and SPECT images	Batch Processing	EOG, RMSE, PSNR, SF, SSIM, MI, MEAN, SD, GRAD, *QAB/F*
Wang et al. [38]	CNN	Pixel	MR–CT, MR-T1–MR-T2, MR–PET, and MR–SPECT images	Trained Siamese convolutional network is used to fuse the pixel activity information of source images to generate weight map. Later a contrast pyramid is implemented to decompose the source image.	QTE, QAB/F, MI, VIF
Marwah Mohammad Almasri et al.	CNN	Pixel	CT and MRI, CT and PET, CT and SPECT	Modified discrete wavelet transform (MDWT) is used as fusion strategy. The fused image is classified into malignant or benign using a convolutional neural network (CNN). The Hybrid Optimization Dynamic (HOD) Algorithm is utilized for enhancing the classification accuracy of the CNN algorithm.	QAB/F, AG, SD, MI, EN, SF, FF
Guo et al. [39]	(CSR)	Feature	CT–MRI MRI–SPECT	Fusion strategy is based on convolutional sparse representation (CSR) and mutual information correlation. The source image is decomposed into one high-frequency and one low-frequency sub-band by non-subsampled shearlet transform. For the high-frequency sub-band, CSR is used for high-frequency coefficient fusion. For the low-frequency sub-band, different fusion strategies are used for different regions by mutual information correlation analysis.	EN, MI, QAB/F, VIF
Tawfik et al. [40]	SAE	Feature	MRI-T1 and MRI-T2, MRI–SPECT	Non-subsampled contourlet transform (NSCT) is used for image decomposition firstly. sparse auto-encoder (SSAE) is implemented for feature extraction to obtain a sparse and deep representation from high-frequency coefficients. Then, the spatial frequencies are computed for the obtained features to be used for high-frequency coefficient fusion. After that, a maximum-based fusion rule is applied to fuse the low-frequency sub-band coefficients. The final integrated image is acquired by applying the inverse NSCT.	EN, MI, QAB/F, SD
Kaur and Singh [35]	CNN	Feature	CT–MRI	NSCT is used for image decomposition. Later Inception (Xception) is used for feature extraction of the source images. Optimal features are selected using multi-objective differential evolution The final integrated image is acquired by applying the inverse NSCT.	edge strength, fusion symmetry, entropy, and fusion factor
Kaur and Singh [41]	DBN	Feature	CT and MRI-T1 image	Feature extraction is done using Deep belief networks(DBN). DBN is used to distinguish the informative and non-informative blocks. Fuzzy based fusion rules are then applied on informative blocks to compute a partially fused image. Finally, DBN and fuzzy fusion rules are again implemented on both fused images .	edge strength, fusion symmetry, entropy, and fusion factor
Kaur and Singh [41]	GAN	Feature	CT–MRI PET–SPECT	Based on residual attention mechanism of GAN. After source images are concatenated to a matrix, the matrix is put into two blocks at the same time to extract information such as size, shape, spatial location and texture details. The obtained features are put into the merge block to reconstruct the image. The obtained reconstructed image and source images are respectively put into two discriminators for correction to obtain the final fused image.	SSIM, MI, VIF, QAB/F
Nair et al. [42]	GAN	Feature	MRI–PET MRI–SPECT	The encoding network is combined with a convolutional neural network layer and a dense block called Generative Adversarial Network (GAN), in contrast to conventional convolutional networks.	MI, FF, QAB/F, ISQE, SF,CC

(*continued on next page*)

Table 2 (*continued*).

Zhang et al. [43]	GAN	Feature	MRI–PET	The goal of discriminator is to differentiate the output image from the original image. When the discriminator is not capable of differentiating the output image from the actual image, the generator has learnt the data. The goal of the GAN proposed in this model is to selectively retain the information present in the input images, namely, MRI and PET.	SSIM, MI
Mi et al. [44]	Generative Adversarial Network	Feature	Visible–Infrared	number of discriminators differentiating between fused image and extreme exposure image pairs is increased. While, a generator network is trained to generate fused images. Through the adversarial relationship between generator and discriminators, the fused image will contain more information from extreme exposure image pairs	SD, PSNR, SCD, CC, SSIM
Huang et al. [11]	GAN	Feature	Visible–Infrared	generator aims to generate a fused image with major infrared intensities together with additional visible gradients, and the discriminator aims to force the fused image to have more details existing in visible images.	EN, SD, SSIM, CC, SF and VIF
Jing Zhang et al.	Transformer	Feature	Visible–Infrared	transformer-based fusion network for capturing both local and global information of source modalities	SSIM, MI, PSNR, QAB/F
Xiangzeng Liu et al.	Transformer	Feature	Visible–Infrared	multi-modal features are extracted from the input images by a CNN. Then, these features are fused by the focal transformer blocks that can be trained through an adaptive fusion strategy according to the characteristics of different features. Finally, the fused features and saliency information of the infrared image are considered to obtain the fused image.	SSIM, MI, VIF, QAB/F
Kaur and Singh [46]	Transformer	Feature	SPECT–MRI PET–MRI	An adaptive convolution is used for adaptively modulating the convolutional kernel based on the global complementary context. For long-range dependencies, an adaptive Transformer is employed to enhance the global semantic extraction capability. It is multiscale fashion so that useful multimodal information can be adequately acquired from the perspective of different scales. Moreover, an objective function composed of a structural loss and a region mutual information loss is devised to construct constraints for information preservation at both the structural-level and the feature-level.	MI, EN, SSIM

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

1. Medical imaging: Multimodal image fusion is commonly used to combine information from multiple modalities, such as MRI, CT, Positron Emission Tomography (PET), and ultrasound, to improve diagnostic accuracy and treatment planning. By fusing information from multiple imaging modalities, multimodal image fusion can provide a more comprehensive and accurate patient condition assessment, aiding in medical diagnosis and treatment planning [46–48].

   A real-world example of multimodal image fusion is in the field of medical imaging, particularly in the fusion of MRI and PET scans. Here's how it works:

   1. **MRI**: MRI provides high-resolution images of the body's internal structures, including soft tissues like organs, muscles, and the brain. It excels in offering detailed anatomical information.
   2. **PET**: PET scans, on the other hand, provide functional information by highlighting areas with high metabolic activity. This is particularly useful in cancer diagnosis and staging, as cancer cells tend to have higher metabolic rates than normal cells.

   Now, consider a patient with a suspected brain tumor:

   - An MRI scan would provide detailed information about the brain's anatomy, showing the exact location and size of the tumor.
   - A PET scan, on the other hand, would show areas with high metabolic activity, which might indicate the presence of cancer cells.

By themselves, these images are valuable, but they offer different types of information. Multimodal image fusion in this scenario involves combining the MRI and PET images into a single fused image. The result is an image that shows both the detailed anatomical structure (from MRI) and the areas of high metabolic activity (from PET) in one view. This fused image can help doctors precisely locate and characterize the tumor. It can be particularly useful in planning surgery or radiation therapy because it provides a more comprehensive understanding of the tumor's size, location, and its relation to surrounding structures. Multimodal image fusion enhances the diagnostic and treatment planning capabilities in the field of medical imaging.

In neurological disorders, multimodal image fusion can combine information from MRI and PET scans to improve the accuracy of diagnosis and treatment. By fusing the anatomical information from the MRI scan with the metabolic information from the PET scan, doctors can more accurately identify the location and severity of the neurological disorder, which can aid in treatment planning.

2. Remote sensing: Multimodal image fusion is used in remote sensing applications to combine information from different sensors, such as optical, radar, and LiDAR, to improve the quality and accuracy of the images and enable better interpretation of the data.
   One application of multimodal image fusion in remote sensing is land use and land cover classification. Land use and land cover classification involves identifying and mapping different land cover types, such as forests, agricultural land, and urban areas. By fusing information from different sensors, such as optical and radar data, the accuracy of land use and land cover classification can be improved, as each sensor provides complementary information that can help to overcome the limitations of individual sensors [3,13,19,21].
   Another application of multimodal image fusion in remote sensing is change detection. Change detection involves identifying changes in the Earth's surface over time, such as deforestation or urbanization. By fusing information from different sensors, such as optical and LiDAR data, changes in the Earth's surface can be detected more accurately and precisely, as each sensor provides complementary information that can help to distinguish between different types of changes.

3. Surveillance and security: Multimodal image fusion is used in surveillance and security applications to combine information from multiple sources, such as cameras, sensors, and thermal imaging, to enhance situational awareness and improve the detection and recognition of objects and events. A few examples are as follows:

   (a) Object recognition: Multimodal image fusion can identify and recognize objects of interest in surveillance footage, such as vehicles or people. For example, combining visible and thermal images can improve object recognition accuracy in low-light or night-time conditions [49].
   (b) Tracking: Multiple sensors, such as cameras, radar, and acoustic sensors, can track objects of interest. Multimodal image fusion can combine the information from different sensors to improve tracking accuracy and robustness. This can be useful in scenarios such as border patrol or tracking of vehicles or individuals.
   (c) Anomaly detection: Multimodal image fusion can detect anomalies in surveillance imagery by combining information from multiple sensors. For example, a combination of visible and infrared imagery can be used to detect the presence of a person in low-light conditions or to detect potential fires.
   (d) Perimeter security: Multimodal image fusion can be used to enhance the security of a perimeter by combining information from multiple sensors, such as cameras, radar, and motion detectors. The fusion of information can provide a more comprehensive view of the area and improve the accuracy of intrusion detection.
   (e) Situational awareness: Multimodal image fusion can provide a more complete situational awareness by combining information from different sources, such as cameras, radar, and acoustic sensors. This can be useful in border security or critical infrastructure protection scenarios.

4. Robotics and autonomous systems: Multimodal image fusion is used in robotics and autonomous systems to combine information from multiple sensors, such as cameras, LiDAR, and radar, to improve perception and enable better decision-making.
   Multimodal image fusion can improve the accuracy of obstacle detection and avoidance in robotics and autonomous systems. For example, combining depth information from a 3D camera with visible images can provide a more accurate representation of the environment and enable better obstacle avoidance.
   Multimodal image fusion can improve navigation accuracy in autonomous systems by combining information from sensors such as cameras, lidar, and GPS. This can help robots and autonomous vehicles navigate complex environments and avoid obstacles.
   Multimodal image fusion can improve the accuracy of localization and mapping in robotics and autonomous systems. For example, combining data from a camera and lidar can enable more accurate mapping of the environment and localization of the robot or vehicle.
   Multimodal image fusion can improve the interaction between robots and humans. For example, combining data from cameras and microphones can enable the robot to recognize and respond to human gestures and speech.

5. Industrial Inspection: Multimodal image fusion can be used for quality control in manufacturing and production processes [50]. For example, combining visible and infrared images can provide a more comprehensive view of the product and enable more accurate detection of defects.
   Multimodal image fusion can be used to improve the capabilities of robots and automated systems in industrial settings. For example, combining data from cameras and sensors can enable robots to navigate and manipulate objects more accurately.
   Multimodal image fusion can be used for process monitoring in industrial settings. For example, combining data from multiple sensors can provide a more comprehensive view of the production process and enable better control and optimization.
   Multimodal image fusion can be used for remote monitoring and control of industrial processes. For example, combining data from cameras and sensors can enable remote monitoring of equipment and processes and enable remote control and intervention when necessary. Table 3 summarizes some of the applications of multimodal image fusion in different fields.

7. Datasets

Image fusion is an emerging area of research. Many efforts are made to collect multimodal images for the fusion task. There are quite a few datasets available. Table 4 provides an overview of publicly available datasets for image fusion tasks. The datasets are compared based on modality type, image size, and ground truth.

8. Challenges and future research areas

Multimodal image fusion using deep learning has made significant progress in recent years Huang et al. [8], but several challenges and future research directions still need to be addressed. Here are some of the challenges and future research directions in multimodal image fusion using deep learning:

(1) **Interpretability**: Deep learning-based methods can be challenging, making it difficult to understand why a particular image fusion result was produced. Developing interpretable deep learning methods to explain the fusion results is an essential research direction.
(2) **Robustness**: Deep learning-based methods can be sensitive to changes in the input data, such as lighting conditions or occlusions. Developing deep learning-based methods robust to these changes is an important challenge.
(3) **Generalization**: Deep learning-based methods can be prone to overfitting, where the model performs well on the training data but poorly on new data. Developing methods that match new data well is an essential research direction.

Table 3
Applications of multimodal image fusion in various fields.

Field	Applications
Medical diagnostics	Cancer diagnosis, neurological disorder diagnosis, cardiovascular disease diagnosis, bone imaging, fetal imaging
Remote sensing	Land use and land cover classification, change detection, target detection, terrain analysis, disaster response
Robotics	Object recognition, obstacle avoidance, navigation
Surveillance and security	Object tracking, anomaly detection
Industrial inspection	Defect detection, quality control
Multimedia	Image and video enhancement, image and video restoration
Biometrics	Face recognition, fingerprint recognition
Astronomy	Image processing and analysis

(4) **Multimodal data fusion**: Most existing methods focus on fusing two input modalities. Developing deep learning-based methods for fusing multiple input modalities is an important research direction.

(5) **Real-time processing**: Real-time processing of multimodal image fusion is still a challenging problem, particularly for deep learning-based methods. Developing efficient deep learning-based methods that can operate in real-time is an important research direction [5]. Image fusion algorithms can be computationally intensive, especially when dealing with large input images or complex fusion methods. Real-time image fusion requires the algorithm to perform the fusion task within a strict time constraint, typically a few milliseconds or less.

In real-world applications, the input images can come from various sources and modalities, each with unique characteristics and challenges. The image

fusion algorithm needs to be robust enough to handle these variations and produce accurate fused images in real-time.

Real-time image fusion systems often have to meet strict quality constraints, such as preserving the most relevant information from each input modality while ensuring the fused image is visually appealing and easily interpreted.

They must be optimized for specific hardware platforms to ensure they can run efficiently and reliably in real-time.

Despite these challenges, researchers are developing real-time image fusion algorithms to meet real-world applications' demands (Akbar et al. 2018; 14).

(6) **Lightweight fusion Algorithm**: A challenge is to develop computationally efficient image fusion algorithms that can be implemented on resource-constrained devices such as mobile phones, embedded systems, and drones. Here are some ways to achieve lightweight multimodal image fusion:

Feature extraction: lightweight feature extraction methods such as local binary patterns (LBP), histograms of oriented gradients (HOG), or Haar wavelets can extract features from each input modality. These methods can be implemented efficiently and require fewer computational resources than deep learning.

Dimensionality reduction: Another approach is to use dimensionality reduction techniques such as principal component analysis (PCA) or linear discriminant analysis (LDA) to reduce the dimensionality of the feature space. This can reduce the computational cost of the fusion algorithm while preserving the most relevant information.

Sparse representation: Techniques such as compressive sensing or sparse coding can represent the input images sparsely, reducing the amount of data that needs to be processed and stored.

Model optimization: Model optimization techniques such as pruning or quantization can be used to reduce the size and complexity of the fusion model without sacrificing performance.

Hardware acceleration: Hardware acceleration techniques such as graphics processing units (GPUs) or field-programmable gate arrays (FPGAs) can be used to speed up the computation of the image fusion algorithm.

Overall, developing lightweight multimodal image fusion algorithms is crucial for applications that require real-time processing or operate on resource-constrained devices. Optimizing the algorithm's computational efficiency makes achieving accurate and reliable image fusion results possible without overburdening the device's resources.

Overall, multimodal image fusion using deep learning is a promising research area with many challenges and future research directions. Addressing these challenges and advancing the field can lead to more accurate and reliable image fusion results with a wide range of practical applications.

9. Conclusion

A proliferation of deep learning is changing how real-world problems are solved, and multimodal image fusion is no exception. For the past few years, deep learning has been employed for multimodal image fusion, and this paper is an attempt to analyze and summarize deep learning techniques for multimodal image fusion. It would profoundly contribute to investigating deep learning for the multimodal image fusion domain.

This research domain is very promising since it will act as a foundation stone in many future computer vision-based projects like medical diagnosis, remote sensing, object detection, etc. Considering the deep learning aspect, there is much scope for improvement in multimodal image fusion approaches by implementing transformer-based architecture models of deep learning. Multimodal fusion for the real-time environment is still in its infancy and needs to be explored.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

The authors would like to thank the editors and anonymous reviewers for their constructive comments and suggestions, which immensely helped improve the overall quality of the manuscript.

Table 4
Benchmarked Datasets for multimodal image fusion.

Fusion modality	Dataset	Description	Ground truth	Link
Hyper-spectral	Chikusei Dataset	Airborne hyperspectral data taken over Chikusei city in Japan. 128 bands in the spectral range from 363 nm to 1018 nm. The scene consists of 2517 × 2335 pixels and the ground sampling distance was 2.5 m.	Y	http://park.itc.u-tokyo.ac.jp/sal/hyperdata
	Harvard	50 indoor and outdoor scenes and 25 images under artificial illumination; 31 bands from 420 nm to 720 nm at 10 nm steps	Y	http://vision.seas.harvard.edu/hyperspec/download.html
	Indian Pines	145 × 145 pixels and 224 spectral reflectance bands in the wavelength range 0.4–2.5 10ˆ(−6) m.	Y	https://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes
	Cuprite	There are 224 channels, ranging from 370 nm to 2480 nm.	Y	https://aviris.jpl.nasa.gov/
	PAVIA	The number of spectral bands is 102 for Pavia Centre and 103 for Pavia University. Pavia Centre is a 1096*1096 pixels image, and Pavia University is 610*610 pixels, but some of the samples in both images contain no information and have to be discarded before the analysis. The geometric resolution is 1.3 m	Y	https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes#Pavia_Centre_and_University
	AVIRIS	AVIRIS sensor; acquired over Moffett Field in California, USA; HS image (17 m GSD); 224 bands from 400 nm to 2500 nm	Y	https://aviris.jpl.nasa.gov/data/free_data.html
	HYDICE Washington	HYDICE sensor; HS image acquired over the National Mall, Washington, USA; HS (2.5 m GSD); 210 bands between 400 nm and 2500 nm	N	https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html
Multi-spectral	CAVE	The images are of a wide variety of real-world materials and objects. 31 band, 512 × 512 pixel	N	https://www.cs.columbia.edu/CAVE/databases/multispectral/
	FLAME 2	FIRE DETECTION AND MODELING: AERIAL MULTI-SPECTRAL IMAGE DATASET, The main dataset contains seven raw, unlabeled RGB and IR video pairs, a set of labeled original resolution RGB/IR frame pairs, and a set of 254p x 254p RGB/IR frame pairs. Both sets of frame pairs are derived from the 7 raw video pairs.	Y	https://ieee-dataport.org/open-access/flame-2-fire-detection-and-modeling-aerial-multi-spectral-image-dataset
	Low-Resolution Multispectral Eds - High-Resolution Panchromatic Sem Images For Close-Range Pansharpening Testing.	represents the X-ray Energy Dispersive (EDS)/ Scanning Electron Microscopy (SEM) images of a shungite-mineral particle. both the noise-treated and original images are present in the current dataset. The image dimensions are 256 by 192 for the EDS maps and 1024 by 768 for the SEM images,	Y	https://ieee-dataport.org/documents/low-resolution-multispectral-eds-high-resolution-panchromatic-sem-images-close-range
	KAIST Multispectral Dataset	consists of 95,328 color-thermal pairs taken from a vehicle. Each image contains RGB color image and infrared image. The dataset captured various conventional traffic scenes, including campus, street and countryside during the day and night respectively	Y	https://soonminhwang.github.io/rgbt-ped-detection/
Multimodal	ATLAS whole brain dataset	Whole Brain ATLAS brain dataset different modalities of medical images which include CT scan, MRI scan, PET scan	N	http://www.med.harvard.edu/aanlib/home.html
	Cancer Imaging Archive (TCIA)	DICOM images of CT and MRI scans	N	https://www.cancerimagingarchive.net/about-the-cancer-imaging-archive-tcia/
	Image angle-Udacity (IA-UDACITY)	The Udacity Autonomous Vehicle Dataset	Y	https://ieee-dataport.org/documents/imageangle-udacity-ia-udacity#files
	C3I Synthetic Face Depth Dataset	3D virtual human models and 2D rendered RGB and GT depth images. The male and female sub-folders contain 56 and 44 subjects, respectively. For the three types of backgrounds — simple, textured, and complex	N	https://ieee-dataport.org/documents/c3i-synthetic-face-depth-dataset
	DIAST	This is a collection of paired thermal and visible ear images. Images in this dataset were acquired in different illumination conditions ranging between 2 and 10700 lux. There are total 2200 images of which 1100 are thermal images while the other 1100 are their corresponding visible images. Images consisted of left and right ear images of 55 subjects. Images were capture in 5 illumination conditioned for every subjects.	N	https://ieee-dataport.org/open-access/diast-variability-illuminated-thermal-and-visible-ear-image-dataset

(continued on next page)

Table 4 (*continued*).

	Road Scene	This datset has 221 aligned Vis and IR image pairs containing rich scenes such as roads, vehicles, pedestrians and so on. These images are highly representative scenes from the FLIR video. We preprocess the background thermal noise in the original IR images, accurately align the Vis and IR image pairs, and cut out the exact registration regions to form this dataset.	Y	https://github.com/jiayi-ma/RoadScene
	LLVIP dataset	A Visible–infrared Paired Dataset for Low-light Vision	Y	https://bupt-ai-cz.github.io/LLVIP/
	MOFA	The MOFA dataset contains 1062 images of 118 groups, in which 450 are indoor and 612 are outdoor.	N	https://www.sciencedirect.com/science/article/pii/S1566253523001008#b36
	TNO	contains multispectral (intensified visual, near-infrared, and longwave infrared or thermal) nighttime imagery of different military relevant scenarios, registered with different multiband camera systems.	N	https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029/2

References

[1] A. Dogra, B. Goyal, S. Agrawal, Medical image fusion: A brief introduction, Biomed. Pharmacol. J. 11 (3) (2018) 1209–1214, http://dx.doi.org/10.13005/bpj/1482.

[2] B. Huang, F. Yang, M. Yin, X. Mo, C. Zhong, A review of multimodal medical image fusion techniques, in: Computational and Mathematical Methods in Medicine, Vol. 2020, 2020a, http://dx.doi.org/10.1155/2020/8279342.

[3] X. Liu, Q. Liu, Y. Wang, Remote sensing image fusion based on two-stream fusion network, Inf. Fusion 55 (2020) 1–15, http://dx.doi.org/10.1016/j.inffus.2019.07.010.

[4] Y. Cui, et al., Deep Learning for Image and Point Cloud Fusion in Autonomous Driving: A Review, 2020, http://dx.doi.org/10.1109/TITS.2020.3023541.

[5] J. Gao, P. Li, Z. Chen, J. Zhang, A survey on deep learning for multimodal data fusion, Neural Comput. 32 (5) (2020) 829–864, http://dx.doi.org/10.1162/neco_a_01273, MIT Press Journals May 01.

[6] G. Xiao, D.P. Bavirisetti, G. Liu, X. Zhang, Image Fusion, Springer Nature, Singapore, 2020, [Online]. Available: https://books.google.co.in/books?id=GGn6DwAAQBAJ.

[7] H. Kumar, P. Pimparkar, Data Fusion for the Internet of Things, Vol. 8, No. 3, 2018, pp. 278–282, http://dx.doi.org/10.29322/IJSRP.8.3.2018.p7541.

[8] B. Huang, F. Yang, M. Yin, X. Mo, C. Zhong, A review of multimodal medical image fusion techniques, in: Computational and Mathematical Methods in Medicine, Hindawi Limited, 2020b, http://dx.doi.org/10.1155/2020/8279342.

[9] D.P. Bavirisetti, R. Dhuli, Multi-focus image fusion using multi-scale image decomposition and saliency detection, Ain Shams Eng. J. 9 (4) (2018) 1103–1117, http://dx.doi.org/10.1016/j.asej.2016.06.011.

[10] K. He, et al., Transformers in medical image analysis: A review, 2022, [Online]. Available: http://arxiv.org/abs/2202.12165.

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

[12] W. Tan, P. Tiwari, H.M. Pandey, C. Moreira, A.K. Jaiswal, Multimodal medical image fusion algorithm in the era of big data, Neural Comput. Appl. (2020) http://dx.doi.org/10.1007/s00521-020-05173-2.

[13] Y. Yang, C. Han, X. Kang, D. Han, An Overview on Pixel-Level Image Fusion in Remote Sensing *, 2007.

[14] C. Radu, et al., Integration of real-time image fusion in the robotic-assisted treatment of hepatocellular carcinoma, Biology (Basel) 9 (11) (2020) 1–13, http://dx.doi.org/10.3390/biology9110397.

[15] Image fusion market, 2023, https://www.theinsightpartners.com/reports/multimodal-image-fusion-software-market. (Accessed 06 May 2023).

[16] D. Agrawal, V. Karar, Bispectral image fusion using multi-resolution transform for enhanced target detection in low ambient light conditions, 2019.

[17] J. Zhou, S. Zeng, Z. Xiao, J. Zhou, H. Li, Z. Kang, An enhanced spectral fusion 3D CNN model for hyperspectral image classification, Remote Sens. (Basel) 14 (21) (2022) http://dx.doi.org/10.3390/rs14215334.

[18] H. Guo, W. Bao, K. Qu, X. Ma, M. Cao, Multispectral and hyperspectral image fusion based on regularized coupled non-negative block-term tensor decomposition, Remote Sens. (Basel) 14 (21) (2022) http://dx.doi.org/10.3390/rs14215306.

[19] X. Feng, L. He, Q. Cheng, X. Long, Y. Yuan, Hyperspectral and multispectral remote sensing image fusion based on endmember spatial information, Remote Sens. (Basel) 12 (6) (2020) http://dx.doi.org/10.3390/rs12061009.

[20] D. Sara, A.K. Mandava, A. Kumar, S. Duela, A. Jude, Hyperspectral and Multispectral Image Fusion Techniques for High-Resolution Applications: A Review. http://dx.doi.org/10.1007/s12145-021-00621-6/Published.

[21] G. Vivone, Multispectral and hyperspectral image fusion in remote sensing: A survey, in: 10.1016/J.Inffus.2022.08.032, Inf. Fusion 89 (2023) 405–417, Elsevier B.V.

[22] F. Abdullah Al-Wassai, N. Kalyankar, A.A. Al-Zaky, A. Professor, Multisensor Images Fusion Based on Feature-Level.

[23] M.A. Bakr, S. Lee, Distributed multisensor data fusion under unknown correlation and data inconsistency, Sensors (Switzerland) 17 (11) (2017) http://dx.doi.org/10.3390/s17112472.

[24] M.A. Saleh, A.A. Ali, K. Ahmed, A.M. Sarhan, A brief analysis of multimodal medical image fusion techniques, Electronics (Basel) 12 (1) (2023) http://dx.doi.org/10.3390/electronics12010097.

[25] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, Pixel-level image fusion: A survey of state of the art, Inf. Fusion 33 (2017) 100–112, http://dx.doi.org/10.1016/j.inffus.2016.05.004.

[26] M. Ehatisham-Ul-Haq, et al., Robust human activity recognition using multimodal feature-level fusion, IEEE Access 7 (2019) 60736–60751, http://dx.doi.org/10.1109/ACCESS.2019.2913393.

[27] H. Kaur, D. Koundal, V. Kadyan, Image fusion techniques: A survey, Arch. Comput. Methods Eng. 28 (7) (2021) 4425–4447, http://dx.doi.org/10.1007/s11831-021-09540-7.

[28] S. Yu, M. He, R. Nie, C. Wang, X. Wang, An unsupervised hybrid model based on CNN and ViT for multimodal medical image fusion, in: Proceedings - 2021 2nd International Conference on Electronics, Communications and Information Technology, CECIT 2021, Institute of Electrical and Electronics Engineers Inc, 2021, pp. 235–240, http://dx.doi.org/10.1109/CECIT53797.2021.00048.

[29] J. Amritha Varshini, S. Aravinth, Hybrid level fusion schemes for multimodal biometric authentication system based on matcher performance, in: J.M.R. S, B. R, S.F. Smys, S. Tavares (Eds.), Computational Vision and Bio-Inspired Computing, Springer Singapore, Singapore, 2021, pp. 431–447.

[30] B.S.N. Rao, N.V.K. Raju, M. Dhanush, P.N.S.M. Harshith, M.J. Mehdi, MRI and spect medical image fusion using wavelet transform, in: 7th International Conference on Communication and Electronics Systems, ICCES 2022 - Proceedings, Institute of Electrical and Electronics Engineers Inc, 2022, pp. 1690–1696, http://dx.doi.org/10.1109/ICCES54183.2022.9835857.

[31] M. Nejati, S. Samavi, S. Shirani, Multi-focus image fusion using dictionary-based sparse representation, Inf. Fusion 25 (2015) 72–84, http://dx.doi.org/10.1016/j.inffus.2014.10.004.

[32] J. Sun, Q. Han, L. Kou, L. Zhang, K. Zhang, Z. Jin, Multi-focus image fusion algorithm based on Laplacian pyramids, J. Opt. Soc. Amer. A 35 (3) (2018) 480–490, http://dx.doi.org/10.1364/JOSAA.35.000480.

[33] H. Bhujle, Weighted-average fusion method for multiband images, in: 2016 International Conference on Signal Processing and Communications, SPCOM, 2016, pp. 1–5, http://dx.doi.org/10.1109/SPCOM.2016.7746635.

[34] C. He, Q. Liu, H. Li, H. Wang, Multimodal medical image fusion based on IHS and PCA, Procedia Eng. 7 (2010) 280–285, http://dx.doi.org/10.1016/j.proeng.2010.11.045.

[35] M. Kaur, D. Singh, Multi-modality medical image fusion technique using multi-objective differential evolution based deep neural networks, J. Amb. Intell. Humaniz. Comput. 12 (2) (2021) 2483–2493, http://dx.doi.org/10.1007/s12652-020-02386-0.

[36] Y. Li, J. Zhao, Z. Lv, Z. Pan, Multimodal medical supervised image fusion method by CNN, Front. Neurosci. 15 (2021b) http://dx.doi.org/10.3389/fnins.2021.638976.

[37] Y. Li, J. Zhao, Z. Lv, J. Li, Medical image fusion method by deep learning, Int. J. Cogn. Comput. Eng. 2 (2020) (2021a) 21–29, http://dx.doi.org/10.1016/j.ijcce.2020.12.004.

[38] K. Wang, M. Zheng, H. Wei, G. Qi, Y. Li, Multi-modality medical image fusion using convolutional neural network and contrast pyramid, Sensors (Switzerland) 20 (8) (2020) http://dx.doi.org/10.3390/s20082169.

[39] P. Guo, G. Xie, R. Li, H. Hu, Multimodal medical image fusion with convolution sparse representation and mutual information correlation in NSST domain, Complex Intell. Syst. 9 (1) (2023) 317–328, http://dx.doi.org/10.1007/s40747-022-00792-9.

[40] N. Tawfik, H.A. Elnemr, M. Fakhr, M.I. Dessouky, F.E.A. El-Samie, Multimodal medical image fusion using stacked auto-encoder in NSCT domain, J. Digit. Imag. 35 (5) (2022) 1308–1325, http://dx.doi.org/10.1007/s10278-021-00554-y.

[41] M. Kaur, D. Singh, Fusion of medical images using deep belief networks, Cluster Comput. 23 (2) (2020) 1439–1453, http://dx.doi.org/10.1007/s10586-019-02999-x.

[42] R.R. Nair, T. Singh, R. Sankar, K. Gunndu, Multimodal medical image fusion using LMF-GAN - A maximum parameter infusion technique, J. Intell. Fuzzy Systems 41 (5) (2021) 5375–5386, http://dx.doi.org/10.3233/JIFS-189860.

[43] J. Zhang, et al., Transformer based conditional GAN for multimodal image fusion, IEEE Trans. Multimedia (2023) 1–14, http://dx.doi.org/10.1109/TMM.2023.3243659.

[44] J. Mi, L. Wang, Y. Liu, J. Zhang, KDE-GAN: A multimodal medical image-fusion model based on knowledge distillation and explainable AI modules, Comput. Biol. Med. 151 (2022) 106273, http://dx.doi.org/10.1016/j.compbiomed.2022.106273.

[45] Y. Wang, J. Peng, J. Zhang, R. Yi, Y. Wang, C. Wang, Multimodal industrial anomaly detection via hybrid fusion, 2023, [Online]. Available: http://arxiv.org/abs/2303.00601.

[46] M. Kaur, D. Singh, Fusion of medical images using deep belief networks, Cluster Comput. 23 (2019) 1439–1453.

[47] R.C. King, E. Villeneuve, R.J. White, R.S. Sherratt, W. Holderbaum, W.S. Harwin, Application of data fusion techniques and technologies for wearable health monitoring, Med. Eng. Phys. 42 (2017) 1–12, http://dx.doi.org/10.1016/j.medengphy.2016.12.011.

[48] R. Prashanth, S. Dutta Roy, Early detection of Parkinson's disease through patient questionnaire and predictive modelling, Int. J. Med. Inform. 119 (2018) 75–87, http://dx.doi.org/10.1016/j.ijmedinf.2018.09.008.

[49] M. Person, M. Jensen, A.O. Smith, H. Gutierrez, Multimodal fusion object detection system for autonomous vehicles, J. Dyn. Syst. Meas. Control Trans. ASME 141 (7) (2019) http://dx.doi.org/10.1115/1.4043222.

[50] Y. Xie, C. Liu, L. Huang, H. Duan, Ball screw fault diagnosis based on wavelet convolution transfer learning, Sensors 22 (16) (2022) http://dx.doi.org/10.3390/s22166270.