Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 32(8): 5067–5089.
DOI: 10.3934/era.2024233
Received: 25 June 2024
Revised: 11 August 2024
Accepted: 14 August 2024
Published: 22 August 2024

Research article

MCADFusion: a novel multi-scale convolutional attention decomposition method for enhanced infrared and visible light image fusion

Wangwei Zhang[1], **Menghao Dai**[1], **Bin Zhou**[2,*] **and Changhai Wang**[1]

[1] Software Engineering College, Zhengzhou University of Light Industry, No.136 Science Road, Zhengzhou 450000, China

[2] Electronics and Electrical Engineering College, Zhengzhou University of Science and Technology, No.1 Xueyuan Road, Zhengzhou 450064, China

* **Correspondence:** Email: whelmmail@126.com; Tel: +86-175-391-26677.

Abstract: This paper presents a method called MCADFusion, a feature decomposition technique specifically designed for the fusion of infrared and visible images, incorporating target radiance and detailed texture. MCADFusion employs an innovative two-branch architecture that effectively extracts and decomposes both local and global features from different source images, thereby enhancing the processing of image feature information. The method begins with a multi-scale feature extraction module and a reconstructor module to obtain local and global feature information from rich source images. Subsequently, the local and global features of different source images are decomposed using the the channel attention module (CAM) and the spatial attention module (SAM). Feature fusion is then performed through a two-channel attention merging method. Finally, image reconstruction is achieved using the restormer module. During the training phase, MCADFusion employs a two-stage strategy to optimize the network parameters, resulting in high-quality fused images. Experimental results demonstrate that MCADFusion surpasses existing techniques in both subjective visual evaluation and objective assessment on publicly available TNO and MSRS datasets, underscoring its superiority.

Keywords: image fusion; multi-scale; convolutional attention decomposition; modal specificity; shared features

1. Introduction

Image fusion is a technique used to combine information from multiple images of the same scene, captured by different sensors, from various locations, or at different times [1]. Combining multiple imaging modalities can yield more comprehensive information about the observed world than using a single modality [2]. Fusion targets include digital images [3, 4], multimodal images [5–7], and remote

sensing images [8–10]. This technique can be widely used in several fields such as saliency detection [11, 12], target detection [13, 14], semantic segmentation [15, 16], object tracking [17], autonomous driving, and video surveillance by providing a clearer presentation of the target and scene [18].

Image fusion mainly consists of three key elements: feature extraction, fusion strategy, and image reconstruction. Existing research has focused on one or more of these processes to improve the fusion performance [19]. In multi-sensor image processing, infrared and visible image fusion has received widespread attention for combining the advantages of both. While infrared images utilize the difference in thermal radiation of the target to discriminate the target, visible light images have advantages in texture details [2]. Infrared and visible image fusion technology integrates the thermal radiation information from infrared images with the texture details of visible images, thereby significantly enhancing the clarity and interpretability of the resulting images. This technology holds significant application value in military, security monitoring, and medical fields. In recent years, numerous methods have been explored to address the challenges of fusing infrared and visible light images [20–22]. Among these, infrared and visible light image fusion methods are primarily categorized into traditional fusion methods [23, 24] and deep learning-based fusion methods [25, 26]. In traditional methods, multi-scale transform techniques [27] are commonly used to extract multi-scale features from different source images. These features are subsequently combined using an appropriate fusion strategy and reconstructed with multi-scale transform techniques to generate the final fused image. Another common approach is pixel-level image fusion [28], which enables rapid merging of different source images by directly applying simple fusion rules, such as averaging, maximizing, or weighted averaging, at the pixel level. There are some shortcomings in the traditional fusion methods: 1) Traditional multi-scale transform methods may fail to effectively retain all critical features from different source images, leading to incomplete information in the fused images. 2) Simple pixel-level fusion rules struggle to manage complex image information, leading to lower quality fused images. 3) Fusion rules must be meticulously crafted to suit specific application scenarios, thereby escalating design complexity and diminishing adaptability across different scenarios. Fusion method based on deep learning: In 2017, Liu et al. [29] introduced a convolutional neural network-based fusion method designed specifically for integrating multi-focus images. Li et al. [30] proposed a fusion approach based on dense blocks and auto-encoding architecture, specifically designed for integrating images from various sources. There is also a generative adversarial network (GAN) based fusion method for infrared and visible images [31]. An adversarial model is trained to generate a fused image containing both the thermal radiation information from the infrared image and the texture details from the visible image. Although the application of deep learning techniques to the field of image fusion has greatly improved the fusion performance. There are still the following shortcomings: 1) The same network is used for both infrared and visible light images, which cannot distinguish the unique features of different source images well, and the fusion strategy uses a manually designed scheme. The features of infrared and visible light images are not well preserved. at the same time, and the quality of the fused image is not high. 2) When using complex architectures such as generative adversarial networks (GAN), the training process may become unstable and prone to problems such as pattern collapse or unstable quality of generated images. 3) Some networks lack downsampling operators for extracting multi-scale features, resulting in underutilization of deep features [32]. Moreover, to fully leverage multi-scale features and integrate deep features effectively, network topology improvements and meticulous fusion strategy design are essential.

In this paper, we propose a new end-to-end fusion method for multi-scale convolutional attention decomposition to solve the above problems. First, our goal is to separately extract global and local information from different source images, while incorporating similarity constraints to their shared and unique features, so as to improve the controllability and interpretability of feature decomposition, in order to obtain the global and local feature information of different source images. Taking infrared images as an example, we initially employ a multi-scale feature extraction module to extract multi-scale features from infrared images. Subsequently, the restormer module enriches feature maps with global and local information using its multi-attention and convolutional properties. Channel attention and spatial attention modules are then used to extract global and local features from the feature maps. The training process is divided into two stages. In the first stage, we independently fuse the global and local information of infrared and visible images using a fusion strategy. In the second stage, we concatenate the global information of the infrared image with that of the visible image, and similarly concatenate the local information. Finally, we fuse the concatenated global and local information. The final fused image is reconstructed using the multi-attention and convolutional properties of the restormer module. The MCADFusion method shows high complexity in terms of technical design, implementation difficulty, computational resource requirements, and resource optimization, but through fine design and optimization, it successfully achieves efficient and accurate fusion of infrared images and visible light images. Experimental results demonstrate that the MCADFusion method surpasses existing fusion methods in both subjective visual and objective evaluations on public TNO and MSRS datasets. The contributions of this paper can be summarized in four main areas:

- Multi-scale feature extraction combined with attention mechanism: We propose combining a multi-scale feature extraction module with the restormer module to acquire and fuse global and local features from different source images using multi-head attention and convolutional properties. Specifically, the channel attention module extracts global features, while the spatial attention module processes local features, effectively preserving and fusing features from infrared and visible light images.
- Two stage training strategy: To enhance the quality of fused images, we adopt a two-stage training strategy. In the first stage, global and local information from infrared and visible images are fused separately. In the second stage, global and local information from both image types are concatenated and then fused. This strategy better preserves the unique and shared features of different source images, improving the quality and consistency of the fused images.
- Similarity constraints and controllability of feature decomposition: We introduce similarity constraints in the feature extraction and fusion process to enhance the controllability and interpretability of feature decomposition. This approach better distinguishes and preserves unique features of different source images while effectively utilizing shared features in the fusion process to generate high-quality fused images.
- End-to-end training design: The MCADFusion method employs an end-to-end training approach that minimizes human intervention, streamlines the fusion process, and significantly enhances adaptability across various scenes and lighting conditions. Using publicly accessible TNO and MSRS datasets, the approach was rigorously compared to state-of-the-art methods both qualitatively and quantitatively, consistently outperforming them in terms of visual and objective metrics. This technique reliably delivers superior performance in both visual and objective evaluations.

2. Related works

In recent years, a variety of techniques have been developed for image fusion. These include traditional methods that fuse infrared and visible images, deep learning-based fusion methods, and feature-level fusion methods. This section will provide a concise overview of these approaches, alongside discussions on the multi-scale feature extraction module, the transformer module, and the channel and spatial attention modules.

2.1. From traditional methods to deep learning and feature-level fusion

Image fusion techniques have witnessed significant development over the past decades, evolving from initial traditional methods to advanced deep learning techniques and feature-level fusion strategies. Traditional methods for infrared and visible image fusion primarily rely on pixel-level operations, such as averaging, maximum value selection, or multi-scale transformation techniques [30, 31]. While these methods are simple and intuitive, they often lack adaptability to complex scenes. With advancements in computational power and data volume, deep learning-based fusion methods [32–34] have emerged, utilizing sophisticated models like convolutional neural networks (CNNs), generative adversarial networks (GANs), and others. These models effectively learn abstract features from data, enhancing the quality of fused images and expanding application versatility. For example, in [25], Li et al. developed an end-to-end framework for infrared and visible image fusion using an innovative deep learning architecture with CNNs and dense blocks. The framework features a dense encoder network with larger kernel convolutions to efficiently capture features over a wide receptive field. Texture details and feature contrast are then improved using a texture contrast compensation module based on gradient residuals and attention mechanisms. Finally, four convolutional layers reconstruct the fused image.

In addition, fusion methods based on feature level [35] further enhance the processing fineness, by directly manipulating and synthesizing the deep features of an image, these methods are able to capture and fuse the key information in the image more efficiently and achieve more advanced fusion effects. This evolution not only pushes the technological boundaries of image fusion techniques, but also expands the prospects for their application in medical imaging, military reconnaissance, autonomous driving, etc.

2.2. Multi-scale feature extraction module

Inception blocks were first introduced by Szegedy et al. in [36]. The multi-scale feature extraction module primarily utilizes and modifies these inception blocks, which were initially designed to extract multi-scale features within the framework of that thesis. In this module, multi-scale feature information is extracted from different source images using 1×1, 3×3, and 5×5 convolutional filters, as well as max pooling methods. Batch normalization (BN) layers are added after each convolutional layer to speed up the training process and enhance the generalization ability of the model. As shown in Figure 1, the multi-scale feature extraction module is divided into four parallel paths, each extracting features at different scales. The outputs of these four paths are concatenated along the channel dimension to form a feature map containing multi-scale features. This module structure effectively combines the advantages of convolution kernels of different sizes to capture multi-scale information from various

source images. Taking an infrared image and a visible image as examples, the inputs are both single-channel. After extraction through the four parallel paths, the outputs are concatenated along the channel dimension. A 1×1 convolution kernel is then used to adjust the number of channels, resulting in a 64-channel feature map. The results of the infrared and visible images processed by this module are shown in Figure 1.

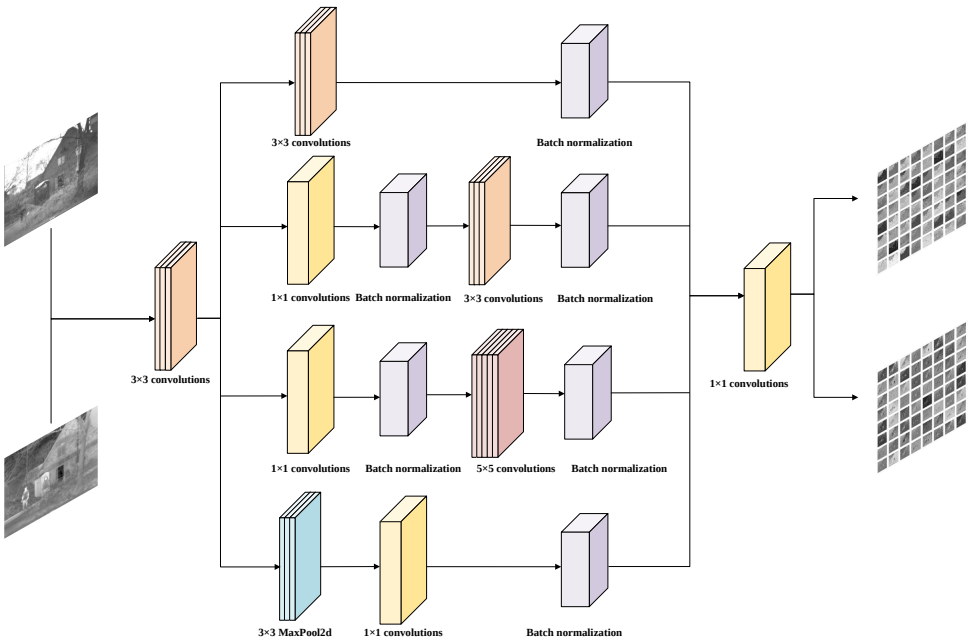

Figure 1. Multi-scale feature extraction module.

2.3. Transformer module in restormer

Transformer was originally proposed by Vaswani et al. [37] for the field of natural language processing, and subsequently, ViT [38] applied it to computer vision. Given the high computational cost of spatial self-attention mechanisms, Wu et al. [39] developed a lightweight LT architecture for mobile NLP tasks. This architecture maintains the performance while reducing the number of model parameters by implementing a long and short ranges attention mechanism and a flattened feedforward network. In addition, restormer [40] improved the transformer module by incorporating a gated deconvolution network and a multi-deconvolution head attention transposition module to support multi-scale local-to-global representation learning for high-resolution images. In the MCADFusion framework, the transformer module is utilized in both the encoder and decoder, with the transformer module depicted in (c) in Figure 2 below. Within the transformer module, input feature maps undergo normalization with LayerNorm (LN) layers to ensure that the inputs of each layer are standardized and not affected by input distribution. In Figure 2(c), Q (Query), K (Key), and V (Value) are employed to compute attentional weights, obtained through matrix multiplication with the input sequence. Subsequently, the dot product operation compares the Q and K vectors, with a softmax activation function applied to provide them with a nonlinear representation. These weighted vectors are then used to compute the final attention representation with the Value vector. Following this, a 1×1 convolution modifies the number of channels. The red connecting arrows denote shortcut connections, which add the input of a layer to its

output to facilitate gradient flow. This is followed by normalization via the LN layer, and subsequently, another 1×1 convolution adjusts the number of channels. The tensor segmentation operation evenly splits the channels, resulting in two feature maps. One of these feature maps undergoes processing with the ReLU activation function, and then the feature maps, after activation, undergo position-wise multiplication to obtain a feature map containing global and local features. Finally, another 1×1 convolution is employed to alter the number of channels and yield the final resultant feature map. The transformer based on primitive blocks cannot guarantee information communication between blocks, which will hinder the extraction of attentional information from the global view [41]. In contrast, our transformer module effectively extracts a feature map containing both global and local features after a multi-scale feature extraction module. Global features improve the efficiency and accuracy of the model by capturing large-scale information, while local features focus on details and filter out irrelevant and redundant information. The notation $8 * (\hat{H} * \hat{W}) * \hat{C}$ and $8 * \hat{C} * (\hat{H} * \hat{W})$ in (c) of Figure 2 represents the shape description of the four-dimensional tensor, with the reshaping primarily for realizing the matrix multiplication operation. The module combines an efficient self-attention mechanism with convolutional operations, which can effectively capture both global and long-range dependencies of the image and accurately extract local features, thus significantly enhancing the image restoration performance; at the same time, its modularized design allows the module to be easily integrated with other models, further extending its performance and application potential. This module is mainly used in the encoding and decoding stages of our method for feature extraction and image reconstruction.

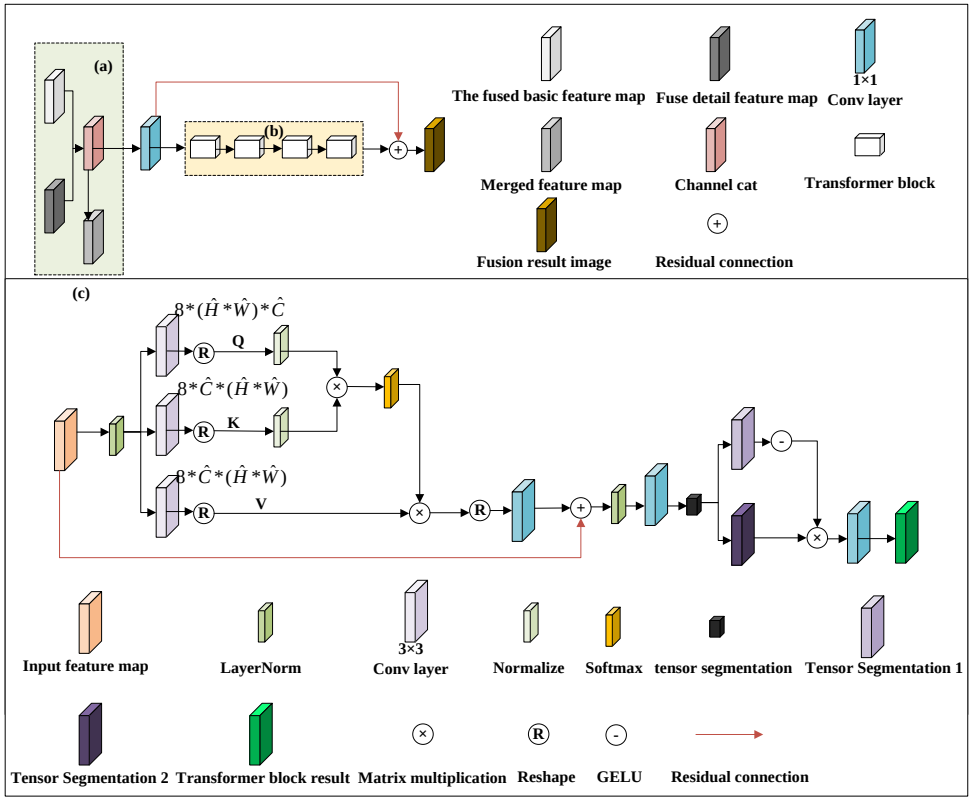

Figure 2. This figure shows the decoder stage, where figure (a) shows the fusion strategy, figure (b) shows the restormer module and figure (c) shows the transformer module.

2.4. Channel and spatial attention module

In [33], Li et al. proposed a network based on nested connections and a spatial/channel attention model. Inspired by this, in the encoder stage of MCAD, we use the spatial attention module (SAM) and the channel attention module (CAM) to extract local and global information of different source images, respectively. Specifically, the channel attention mechanism is mainly used to further extract global features of infrared or visible images, called basic features; the spatial attention mechanism is mainly used to further extract local features of infrared or visible images, called detailed features. Among the CAMs is the shared MLP (a Multilayer Perceptron), which consists of multiple fully connected (dense) layers, each applying a linear transformation followed by a nonlinear activation function. In this process, different regions of the feature map share the same set of weights (parameters), ensuring parameter efficiency and consistency. This sharing mechanism facilitates the learning of uniform transformations that can be applied to different parts of the input data. The specific implementations of the channel attention module (CAM) and spatial attention module (SAM) are shown in Figures 3 and 4, respectively.

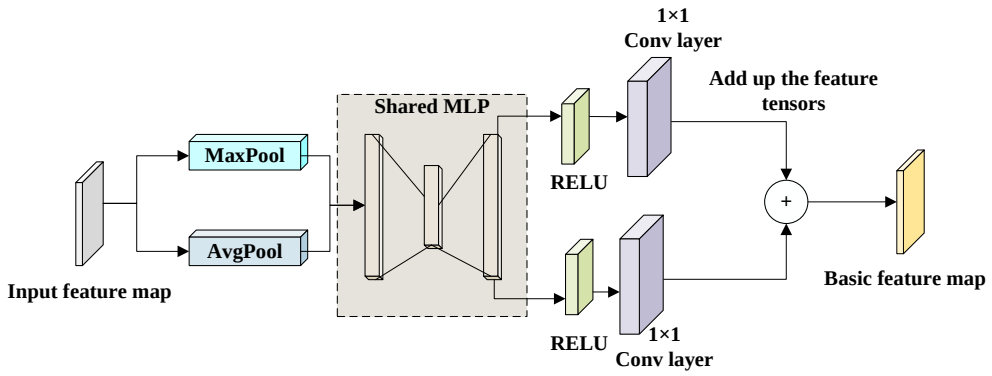

Figure 3. Channel attention module (CAM).

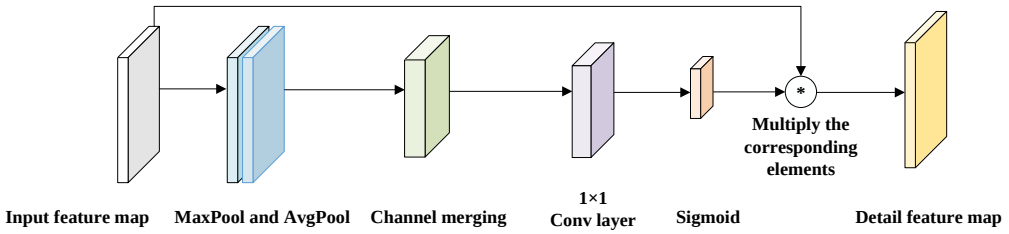

Figure 4. Spatial attention module (SAM).

2.5. Comparison with existing methods

The MCADFusion model demonstrates significant advantages over existing image fusion techniques. Compared with traditional infrared and visible image fusion methods, MCADFusion not only extracts and fuses local and long-range features from different source images more efficiently, but its

structural design also makes the model intuitive and easy to understand. Additionally, the multi-scale convolutional attention fusion strategy adopted by MCADFusion, combined with the channel attention (CAM) and spatial attention (SAM) modules, enables it to excel in preserving key global information and enhancing local texture details when handling high-complexity image fusion tasks. Compared with conventional convolutional neural network (CNN)-based fusion methods, our model not only offers more fine-grained feature processing capabilities but also improves the quality of feature extraction by using an innovative loss function that effectively suppresses redundant information.

3. Methodology

In this section, we first briefly describe the workflow of MCADFusion and its sub-modules. Then, we provide a detailed description of the implementation process for the encoder, fusion layer, and decoder. For ease of discussion, we refer to the low-frequency global features as base features and the high-frequency local features as detailed features.

3.1. General

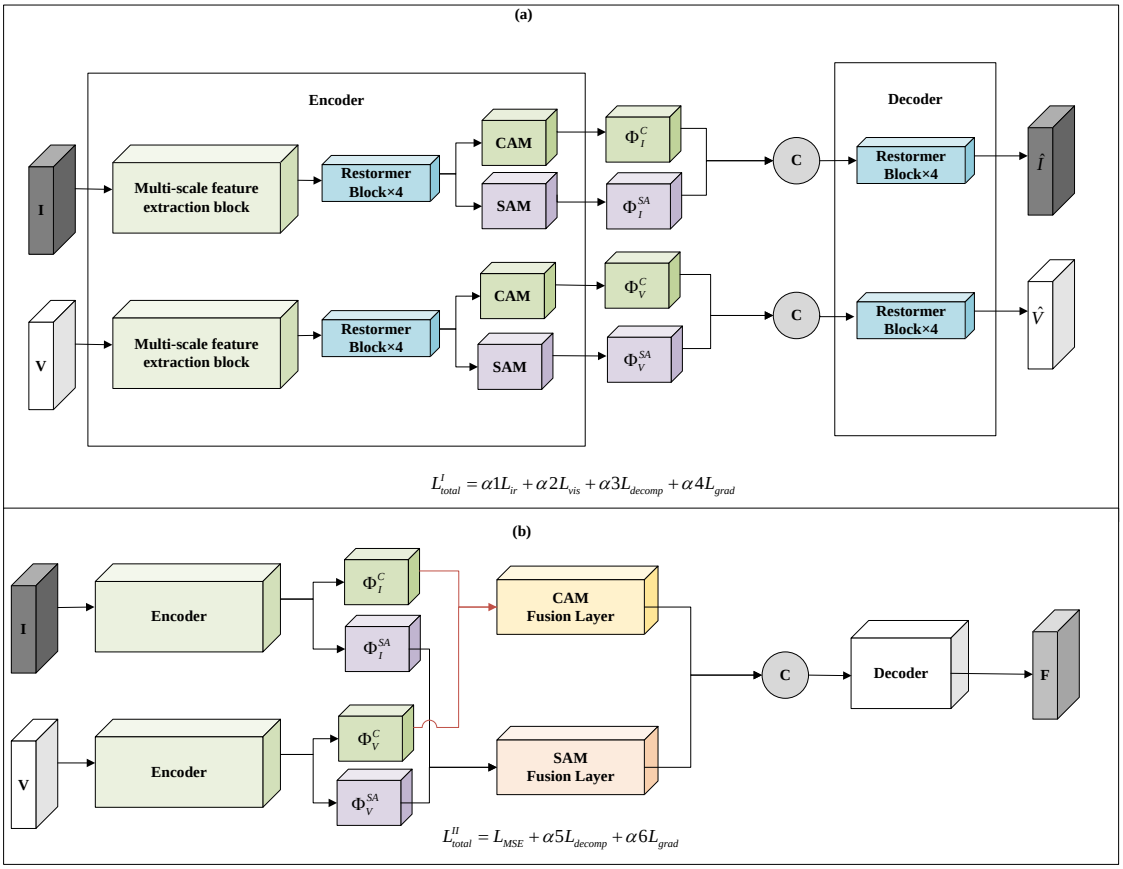

Figure 5. Framework of our MCADFusion method: (a) Training Phase I, which aims to train the AE structure for base/detail feature decomposition and reconstruction of the source image. (b) Training Phase II, which aims to obtain the final fused image.

The MCADFusion method begins by decomposing the base and detail features of different source images using a decoder. These features are then fused through a fusion layer, and finally, the image is reconstructed using a decoder. In the first stage, the input consists of separate channels for infrared and visible light images These images pass through a multi-scale feature extraction module and the restormer module to produce shallow feature maps. These maps are then processed through the channel attention mechanism and spatial attention mechanism modules to extract basic and detailed features for both infrared and visible light images. For example, the base and detail features of the infrared image are input into the restormer module after channel merging, resulting in the recovered infrared image. In the second stage, the same inputs are used as in the first stage. To better integrate the base and detail features from both the infrared and visible light images, the training parameters from the first stage are utilized. The shallow base or detail features extracted from the infrared image via the multi-scale feature extraction and restormer modules, along with the corresponding features from the visible light image, are channel-merged. They are then passed through the channel attention module and apatial attention module, respectively, to obtain the base and detail features for the infrared or visible light images. These features are subsequently input into the restormer module after channel merging, ultimately producing the fused image of the infrared and visible light images. In both the first and second stages, the feature extraction module is referred to as the encoder stage, while the stage where the images are reconstructed after channels merging is called the decoder stage. The detailed workflow is illustrated in Figure 5.

3.2. Encoders

The encoder stage consists of four main modules: the shared multi-scale feature extraction (MSF) module, the restormer sharing (SFE) module, the CAM module, and the SAM module. The MSF module extracts multi-scale features from different source images. Following this, the SFE module utilizes a transformer module to extract feature information that includes both global and local features. The two-branch structure employs the CAM and SAM modules to simultaneously extract base and detailed feature information from different source images. These CAM and SAM modules ensure synchronized extraction of these features.

For clear understanding, we denote the input IR and visible images as $I \in R^{H*W*1}$ and $V \in R^{H*W*3}$. where the four modules MSF, SFE, CAM, and SAM in the encoder stage are represented by M(.), SF(.), CC(.), and SA(.) representations. The purpose of shared feature encoder based on multi-scale feature extraction (MSF) is to extract multi-scale features $\{\Phi_I^M, \Phi_V^M\}$ from different source images $\{I, V\}$ of the input, i.e.,

$$\Phi_I^M = M(I), \quad \Phi_V^M = M(V) \tag{3.1}$$

The multi-scale feature extraction module is used in MSF to improve the robustness of the model, to enhance the model's adaptability to scale changes, and to help the model to distinguish between foreground and background. The purpose of SFE is to extract features containing both base and detailed features $\{\Phi_I^{SF}, \Phi_V^{SF}\}$ from the extracted multi-scale features $\{\Phi_I^M, \Phi_V^M\}$, i.e.,

$$\Phi_I^{SF} = SF(\Phi_I^M), \Phi_V^{SF} = SF(\Phi_V^M) \tag{3.2}$$

The restormer block is chosen in SFE because it can be used to extract the global features of the

source image from the high-resolution input feature maps by applying its multi-head attention across feature dimensions. Therefore, the use of restormer block can be realized to extract the cross-modal shallow feature maps. The purpose of CAM is to extract the base features from containing both base and detailed features $\{\Phi_I^{SF}, \Phi_V^{SF}\}$, i.e.,

$$\Phi_I^C = C(\Phi_I^{SF}), \Phi_V^C = C(\Phi_V^{SF}) \tag{3.3}$$

where $\{\Phi_I^C, \Phi_V^C\}$ denote the base features of infrared and visible image respectively. The use of channel attention mechanism (CAM) mainly improves the performance, efficiency, flexibility, adaptability to new data, and robustness of deep learning models by emphasizing important features and suppressing minor features. Spatial attention detail feature extraction (SAM) is just the opposite of CAM, which mainly extracts the detail features from containing both base and detailed features $\{\Phi_I^{SF}, \Phi_V^{SF}\}$, i.e.,

$$\Phi_I^{SA} = SA(\Phi_I^C), \Phi_V^{SA} = SA(\Phi_V^C) \tag{3.4}$$

where $\{\Phi_I^{SF}, \Phi_V^{SF}\}$ denotes the detail features of infrared and visible images, respectively. The use of the spatial attention mechanism (SAM) effectively enhances the ability of the deep learning model in processing image details, adapting to new environments, improving computational efficiency, and resisting interference by focusing on the key parts of the image, thus achieving better performance in the task of extracting detailed features.

3.3. Fusion layer

The function of the base/detail fusion layer is mainly to fuse the base/detail feature information from different source images separately. Since the generalization bias in base/detail fusion is similar to the base/detail feature extraction obtained through encoder decomposition, we adopt CAM and SAM blocks as the base/detail fusion layer. The base/detail features $\{\Phi_I^C, \Phi_V^C\}$, $\{\Phi_I^{SA}, \Phi_V^{SA}\}$ are extracted from the encoder for infrared and visible light, where:

$$\Phi^C = F_C(\Phi_I^C, \Phi_V^C), \Phi^{ST} = F_{SA}(\Phi_I^{SA}, \Phi_V^{SA}) \tag{3.5}$$

$\{F_C, F_{SA}\}$ is base/detail fusion layer respectively.

3.4. Decoder

In the decoder D(.), in which the features obtained from decomposition are concatenated in the channel dimension as input to the decoder. In training phase I, the decoder outputs the original image; while in training phase II, the output is the fused image. It is formulated as:

$$\begin{aligned} Stage\ I &: \Phi^C = F_C(\Phi_I^C, \Phi_V^C) \\ Stage\ II &: \Phi^C = F_C(\Phi_I^C, \Phi_V^C) \end{aligned} \tag{3.6}$$

In the decoder stage, since the inputs for the second stage of training are base and detailed features from different modalities, in order to maintain design consistency with the SFE (feature extractor), the decoder employs a restorer block as its base building block.

3.5. Two-stage training

The lack of field validation due to the need to extract base and detail features separately from different source images makes state-of-the-art semi-supervised and, fully supervised learning methods inapplicable. Therefore, we use a two-stage learning scheme for end-to-end training of these images, for which the MCADFusion method is proposed, and the two-stage framework is shown in Figure 5.

In the training stage I, the infrared and visible images $\{I \in R^{H*W*1}, V \in R^{H*W*3}\}$ are used as the inputs of the encoder, the multi-scale features $\{\Phi_I^M, \Phi_V^M\}$ of the infrared and visible images are obtained by the shared feature encoder based on multi-scale feature extraction (MSF), $\{\Phi_I^M, \Phi_V^M\}$ are used as the inputs of the encoder, and then the shallow features $\{\Phi_I^{SF}, \Phi_V^{SF}\}$ are obtained by the shared feature encoder based on the restorer [SFE] block. The shared feature encoder based on the restorer [SFE] block obtains the shallow features $\{\Phi_I^{SF}, \Phi_V^{SF}\}$, the shallow features are used as the inputs of the dual channel (based on the channel-attentive base feature extraction (CAM) encoder and spatial-attentive detailed feature extraction (SAM) encoder), to obtain the base and detailed features of the infrared (or visible light) image $\{\Phi_I^C, \Phi_I^{SA}\}$ (or $\{\Phi_V^C, \Phi_V^{SA}\}$), the base and detail features are concatenated as the input of the decoder, and finally the IR (or visible light) original image $\{\hat{I}, \hat{V}\}$ is obtained by the decoder.

In the training phase II, infrared and visible images ($\{I \in R^{H*W*1}, V \in R^{H*W*3}\}$) are fed into the encoder pre-trained in training phase I. We extract the image on the basis of characteristics of $\{\Phi_I^C, \Phi_V^C\}$ and the detail characteristics of $\{\Phi_I^{SA}, \Phi_V^{SA}\}$. These features are then fed into the corresponding fusion module F_C or F_{SA}. Finally, the fused features $\{\Phi^C, \Phi^{SA}\}$ are fed into the decoder to generate the fused image F.

The total loss function used in the training phase I is:

$$L_{total}^I = \alpha_1 L_{ir} + \alpha_2 L_{vis} + \alpha_3 L_{decomp} + \alpha_4 L_{grad} \tag{3.7}$$

where L_{ir} and L_{vis} are the similarity and error between the IR and visible images and the reconstructed image, respectively, and both L_{ir} and L_{vis} are known as reconstruction losses. L_{decomp} is the similarity between the IR and visible images at the base and detail level for finer and more balanced feature fusion called base and detail decomposition loss, and L_{grad} is the gradient of the fused image between the IR and visible images and the fused image loss. In Eq (3.7), the values of each hyperparameter are $\alpha_1 = 1$, $\alpha_2 = 1$, $\alpha_3 = 2.5$, and $\alpha_4 = 6$. These values have been experimentally verified to ensure that the model can balance the infrared image loss and visible light image loss to achieve optimal image restoration. The weights of the infrared and visible light image losses are both 1, indicating their equal importance. The weight of the decomposition loss is 2.5, emphasizing the model's image decomposition ability, while the weight of the gradient loss is 6, focusing on edge information and detail preservation. Adjusting these hyperparameters will degrade the model's performance in certain aspects, affecting image quality. By maintaining these optimized values, the model can achieve the best balance between the losses and ensure high-quality image restoration. The reconstruction loss, i.e.,

$$L_{ir} = \beta_1 L_{\text{ssim}}(I, \hat{I}) + L_{\text{MSE}}(I, \hat{I}) \tag{3.8}$$

In Eq (3.8), $L_{ssim}(I, \hat{I}) = 1 - SSIM(I, \hat{I})$ is the structural similarity index, which is used to measure the similarity of the visual effect of the two images, and L_{ssim} is mainly used to evaluate the quality loss between the recovered image and the original image. $L_{MSE} = \|I - \hat{I}\|_2^2$ is mainly used to calculate the average of the sum of squares of the pixel-level differences between the original image and the

recovered image, which is used to accurately control the consistency of each pixel of the reconstructed image with the original image, and to improve the pixel-level accuracy. β_1 is set to 5 for regulating the effect of the L_{ssim} loss in the total loss function. The weighted combination of loss functions using β_1 and β_2, a combination that primarily utilizes the perceived quality benefits of SSIM losses while still retaining the numerical accuracy of MSE losses. L_{vis} is similar to L_{ir}, which is used for assessing the the reconstruction loss of the visible image with the reconstructed image after reconstruction, i.e.,

$$L_{vis} = \beta_2 L_{ssim}(V, \hat{V}) + L_{MSE}(V, \hat{V}) \tag{3.9}$$

In Eq (3.9), we introduce the correlation CC(.), which is used to calculate the degree of similarity between the base or detail feature vectors $\{\Phi_I^{SA}, \Phi_V^{SA}\}$ or $\{\Phi_I^C, \Phi_V^C\}$ of the infrared and visible images, and is mainly used as a measure of the loss of correlation between different features. In order to make it combinable, we also introduce $\epsilon = 1.01$, mainly to prevent the denominator from being 0 or close to 0, which leads to divergence or non-divergence of the calculation results, and to enhance the robustness and stability of the formula. The base and detail total decomposition loss equations, i.e.,

$$L_{decomp} = \frac{(L_{CC}^{SA})^2}{L_{CC}^C + \epsilon} = \frac{(CC(\Phi_I^{SA}, \Phi_V^{SA}))^2}{CC(\Phi_I^C, \Phi_V^C) + \epsilon} \tag{3.10}$$

In addition, the gradient loss in Eq (3.7) is calculated as shown in Eq (3.11) below, i.e.,

$$L_{grad} = \frac{1}{HW} || \, |\nabla I_f| - max(|\nabla I_{ir}|, |\nabla I_{vis}|_1) ||_1 \tag{3.11}$$

Subject to [42, 43], during training phase II, the total loss function is:

$$L_{total}^{II} = L_{MSE} + \alpha_5 L_{decomp} + \alpha_6 L_{grad} \tag{3.12}$$

4. Experimental validation

In this section, we first describe the experimental setup, detailing the datasets, evaluation metrics, and training process. We then conduct quantitative and qualitative analyses on two publicly available datasets, comparing their performance with seven state-of-the-art models in both subjective and objective aspects. Finally, we validate the impact of different model components through ablation experiments.

4.1. Experimental configuration

Datasets: we validate our MCADFusion model using two publicly available datasets, MSRS [44], and TNO [45]. We compared the fusion results with seven state-of-the-art fusion models: DIDFuse [7], SDNet [46], U2Fusion [6], TarDAL [20], DeFusion [47], ReCoNet [48], CDDFuse [43]. We used entropy (EN), standard deviation (SD), spatial frequency (SF), $Q^{AB/F}$, difference correlation sum (SCD), and structural similarity index measure (SSIM) as six metrics to quantitatively measure the fusion results. Detailed information on these metrics can be found in [49].

Implementing rules: The experiments in this study were completed on a high-performance computer configured with Intel(R) Xeon(R) Platinum 8352V CPUs, 90GB RAM, NVIDIA RTX 4090 GPU

(quantity: 1), and running Ubuntu 18.04, PyTorch 1.8.1, and CUDA 11.1, which ensured efficient complex model processing and accurate reproduction of complex models. In the preprocessing stage, the training samples were randomly cropped into small blocks of 128 × 128 pixels. The whole training process consists of 120 rounds divided into two phases: 40 rounds in the first phase and 80 rounds in the second phase, with 8 samples processed in each batch. The Adam optimizer was used, and the initial learning rate was set to 0.0001, and the learning rate was reduced by 0.5 after each round of training. The hyper parameters settings of the network included: the number of restorer blocks in the SFE was 4, with 8 attention heads and 64 dimensions. The weight parameters of the loss function were set to 1 for both α_1 and α_2, 2.5 for α_3 and 6 for α_4, 5 for both β_1 and β_2, 2.5 for α_5, and 10 for α_6.

4.2. Experimental results

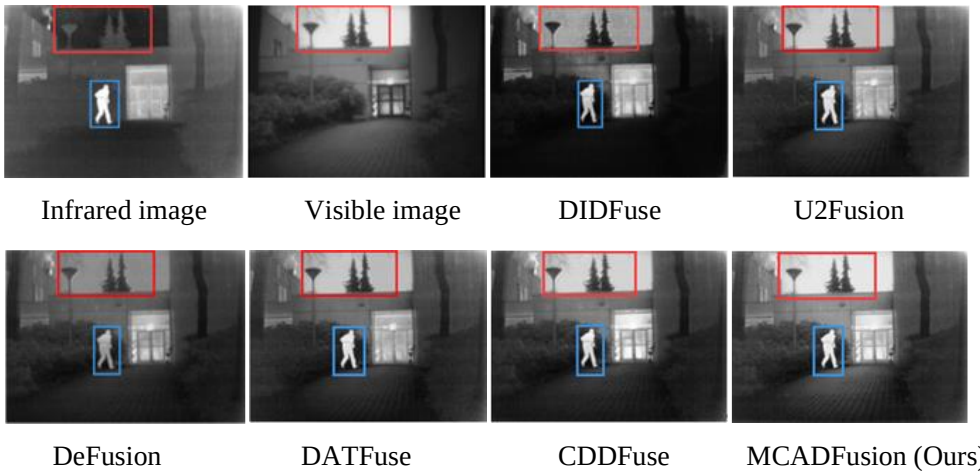

Figure 6. Qualitative comparison of our MCADFusion against 5 state-of-the-art methods on 1 typical VIS and IR image pair in the TNO dataset.

Qualitative comparison: In a qualitative comparison, the MCADFusion methodology was scrutinized against several cutting-edge, state-of-the-art (SOTA) techniques. The assessment of these fusion approaches was conducted using the TON dataset, employing two distinct scenarios, as illustrated in Figures 6–8. Furthermore, the outcomes achieved following the integration of multiple scenes are presented on the MSRS dataset, showcasing the effectiveness of the proposed method. In Figure 6, the fused images produced by the DIDFuse, DeFusion, and DATFuse methods exhibit more noise in the red box, and the salient features are not clear. In contrast, the fused images generated by the MCADFusion method contain less noise in the red box and appear more natural. Additionally, the MCADFusion method is more natural in appearance compared to the U2Fusion, DATFuse, and CDDFuse methods. In the blue box, the human thermal imaging information generated by the DIDFuse, U2Fusion, and DeFusion methods is darker, while there is almost no visual difference in terms of human sensitivity compared to the CDDFuse method. In Figure 7, within the group 1 images, the MCADFusion method provides clearer signage information in the red box relative to the DIDFuse, DATFuse, and CDDFuse

methods. In the yellow box, the MCADFusion method retains more detailed information. Meanwhile, in the blue box, there is almost no visual difference between the information retained by the MCADFusion method and the DATFuse. CDDFuse methods in terms of human perception. In the second set of images, the fusion results of the DIDFuse method in the blue basket and the red box are darker in terms of feature information, while the CDDFuse and MCADFusion methods retain more detailed feature information. In Figure 8, the results of the MCADFusion method demonstrate superior preservation of content within the red-framed area and the basket. This approach effectively retains both the thermal radiation information from the infrared image and the detailed feature information from the visible image, resulting in a more natural visual effect.

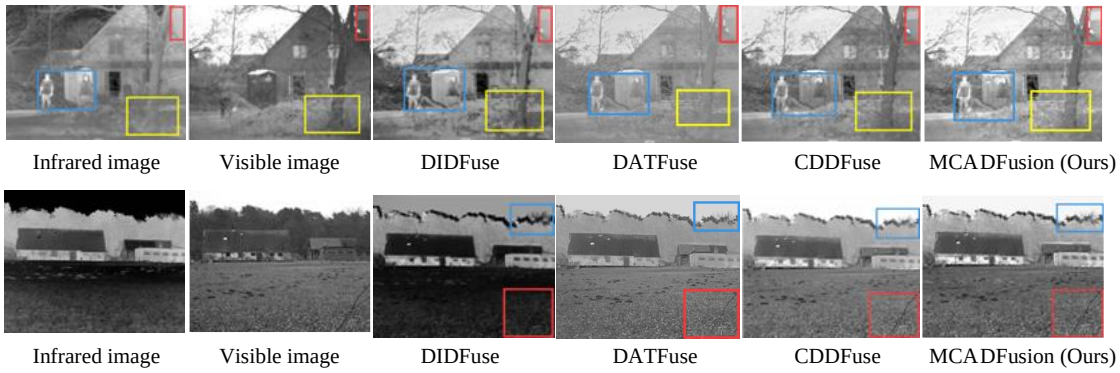

Figure 7. Qualitative comparison of our MCADFusion against 2 state-of-the-art methods on 1 typical VIS and IR image pair in the TNO dataset (continued).

Quantitative comparison: In a quantitative comparison, the MCADFusion method was compared with eight existing SOTA methods. Six metrics, namely entropy (EN), standard deviation (SD), spatial frequency (SF), $Q^{AB/F}$, difference correlation (SCD), and structural similarity index (SSIM), were used to quantitatively assess the fusion results. The quantitative comparison results on the publicly available TNO dataset are shown in Table 1 below. It is obvious from Table 1 that the MCADFusion method has better performance in all four metrics, EN, SD, SF, and SCD, compared to the eight existing SOTA methods. The higher these four metrics are, the better the fused image is in terms of information, contrast, and similarity of detail information to the original image. Meanwhile, the two metrics, $Q^{AB/F}$ and SSIM, have little difference between the metrics in compared with the other eight methods, while the higher of these two metrics indicates that the fused image is better in terms of quality and detail retention, and the structural similarity with the original image. It indicates that the MCADFusion method does not differ much from the other eight methods in terms of image quality and degree of detail retention. The quantitative comparison results on the publicly available MSRS dataset are shown in the following Table 2. It is obvious from Table 2 that the MCADFusion method has comparable or better performance compared to the seven existing SOTA methods in the five metrics of EN, SD, SF, SCD, and SSIM. It can be shown that the fused image obtained by the MCADFusion method performs better in terms of information, contrast, detail information, similarity of the original image, and structural similarity compared to the other 8 SOTA methods. Meanwhile, the $Q^{AB/F}$ metrics are not much different from the other methods, indicating that they are also not much different in terms

of the quality of the fused images. The evaluation metrics in Figures 3 and 4 clearly demonstrate that removing various modules from our MCADFusion model results in decreased performance across nearly all metrics. This substantiates the importance and effectiveness of each module within our model.

Figure 8. The images are sourced from the MSRS dataset. The left column displays infrared images, the middle column shows visible light images, and the right column presents the fusion results of MCADFusion applied to the MSRS dataset.

Table 1. Quantitative results from the IVF task using the TNO dataset: Optimal values are highlighted in bold and italics indicate suboptimal values.

	EN	SD	SF	$Q^{AB/F}$	SCD	SSIM
DIDFuse [7]	6.97	45.12	12.59	0.40	1.71	0.81
SDNet [46]	6.64	32.66	12.05	0.44	1.49	*1.00*
U2Fusion [6]	6.83	34.55	11.52	0.44	1.71	0.99
TarDAL [20]	6.84	45.63	8.68	0.32	1.52	0.88
DeFusion [47]	6.95	38.41	8.21	0.41	1.64	0.96
ReCoNet [48]	7.10	44.85	8.73	0.39	1.70	0.88
DATFuse [50]	6.58	29.65	10.09	0.51	1.45	0.94
CDDFuse [43]	*7.12*	*46.00*	*13.15*	**0.54**	1.76	**1.03**
MCADFusion	**7.20**	**48.39**	**13.58**	*0.48*	**1.87**	0.99

Table 2. Quantitative IVF task results on the MSRS dataset: Optimal values are highlighted in bold and italics indicate suboptimal values.

	EN	SD	SF	$Q^{AB/F}$	SCD	SSIM
DIDFuse [7]	4.27	31.49	10.15	0.2	1.11	0.24
SDNet [46]	5.25	17.35	8.67	0.38	0.99	0.72
U2Fusion [6]	5.37	25.52	9.07	0.42	1.24	0.77
TarDAL [20]	5.28	25.22	5.98	0.18	0.71	0.47
DeFusion [47]	6.46	37.63	6.60	0.54	1.35	0.94
ReCoNet [48]	6.61	43.24	9.77	0.50	1.44	0.85
DATFuse [50]	6.58	40.45	*11.63*	0.64	1.44	0.90
CDDFuse [43]	*6.70*	*43.38*	11.56	**0.69**	*1.62*	**1.00**
MCADFusion	**6.83**	**50.77**	**12.84**	*0.66*	**1.76**	*1.00*

4.3. Ablation experiments

We conducted a large number of ablation experiments to validate the plausibility of the different modules in the MCAD model. In these experiments, we used six evaluation metrics such as entropy (EN), standard deviation (SD), spatial frequency (SF), $Q^{AB/F}$, difference correlation (SCD), and structural similarity index (SSIM) to assess the validity and reasonableness of our method. The results are shown in Tables 3 and 4.

First, removing the multi-scale feature extraction module and keeping the other hyperparameters and model framework unchanged showed that not adding the multi-scale feature extraction module was less effective. Second, simplifying the training phase to use only the second phase and performing 120 rounds of training, the results show that two-stage training is better than single-stage training. Third, comparing the effect of using four and three restorer blocks, the experiment shows that the configuration of four restorer blocks is superior to three. Finally, replacing the channel attention module with a module using a 3×3 convolutional kernel, the results demonstrate that the channel attention module achieves better fusion results.

Table 3. Ablation experiment results on the TNO test set. Bold values indicate the best performance.

Configurations	EN	SD	SF	$Q^{AB/F}$	SCD	SSIM
Remove the multi-scale feature extraction module	6.49	45.05	13.10	0.40	1.72	0.96
Phase II training only	*7.19*	*48.18*	*13.50*	0.46	1.84	0.98
The number of restorer blocks is 3	7.13	47.03	13.32	*0.47*	1.73	0.98
Removing channel attention	7.14	47.64	13.27	0.41	*1.76*	**1.00**
MCADFusion (Ours)	**7.20**	**48.39**	**13.58**	**0.48**	**1.87**	*0.99*

Table 4. Ablation experiment results on the MSRS test set. Bold values indicate the best performance.

Configurations	EN	SD	SF	$Q^{AB/F}$	SCD	SSIM
Remove the multi-scale feature extraction module	6.24	47.38	11.91	0.64	1.68	1.00
Phase II training only	6.80	*50.00*	12.70	0.64	*1.74*	1.00
The number of restorer blocks is 3	6.72	49.41	*12.70*	*0.65*	1.70	*1.00*
Removing channel attention	*6.81*	49.60	12.24	0.62	1.67	0.99
MCADFusion (Ours)	**6.83**	**50.77**	**12.84**	**0.66**	**1.76**	**1.00**

Table 5. Results of ablation experiments with different parameter weights in the test set of TNO. Bold indicates the best value.

Configurations: (α_1, α_2, α_3, α_4, β_1, β_2)	EN	SD	SF	$Q^{AB/F}$	SCD	SSIM
(1, 1, 2.5, 6, 2, 2)	7.14	44.12	11.74	0.49	1.83	1.01
(1, 1, 2.5, 6, 4, 4)	7.12	46.00	13.15	**0.54**	1.76	*1.03*
(1, 1, 3, 6, 5, 5)	7.12	13.75	11.84	*0.50*	1.79	**1.03**
(1, 1, 2.5, 7, 5, 5)	*7.20*	*48.10*	*13.36*	0.46	*1.85*	0.98
Ours (1, 1, 2.5, 6, 5, 5)	**7.20**	**48.39**	**13.58**	0.48	**1.87**	0.99

Table 6. Results of ablation experiments with different parameter weights in the test set of MSRS. Bold indicates the best value.

Configurations: (α_1, α_2, α_3, α_4, β_1, β_2)	EN	SD	SF	$Q^{AB/F}$	SCD	SSIM
(1, 1, 2.5, 6, 2, 2)	6.72	48.58	11.26	0.66	*1.74*	**1.03**
(1, 1, 2.5, 6, 4, 4)	6.65	49.02	12.05	0.64	1.68	*1.02*
(1, 1, 3, 6, 5, 5)	6.70	*50.56*	**13.01**	*0.65*	1.69	1.00
(1, 1, 2.5, 7, 5, 5)	*6.80*	50.48	11.56	0.65	1.62	0.96
Ours (1, 1, 2.5, 6, 5, 5)	**6.83**	**50.77**	*12.84*	**0.66**	**1.76**	1.00

We conducted ablation experiments on the hyperparameters α_1, α_2, α_3, α_4, β_1, and β_2. Specifically, β_1 and β_2 are the weight parameters between the SSIM and MSE losses for the infrared (or visible) image and the reconstructed image, trained during the first stage of the total loss functions. For comparison purposes, we set β_1 and β_2 to 2, 4, or 5. The results indicated that all evaluation indexes were

relatively better when β_1 and β_2 were set to 5. This proves that when the weight of the structural loss function is 5, the quality and detail retention of the IR and visible images in the fused images are better.

Additionally, we compared the different weights of the modified parameters α_3 and α_4. We concluded that the evaluation indexes were relatively better when α_3 and α_4 were set to 5 and 6, respectively. The results of these ablation experiments on the TNO and MSRS datasets are presented in Tables 5 and 6.

4.4. Extensibility of the MCADFusion method

The MCADFusion method proposed in this paper not only excels in infrared and visible image fusion tasks but also shows significant potential for applications in other areas, such as RGB-D image segmentation and the fusion of point clouds and images. Specifically, MCADFusion can be extended to RGB-D image segmentation by leveraging multi-scale convolution and attention mechanisms. This approach draws on the strategies of depth-aware convolutional neural networks [51] and cascaded feature networks [52] to enhance segmentation accuracy. More likely, since the MCADFusion method ultimately produces a fused image that incorporates the thermal radiation information and detailed texture features of the infrared and combinable light images, there is a considerable advantage for the method to be used in the future for preprocessing of small target detection [53]. Meanwhile, the multi-scale feature extraction module and restormer module proposed in this paper can be easily used to migrate to other methods, especially multimodal image fusion techniques [54, 55], which can be used to extract features and recover fused medical images of different morphologies from different source images. Additionally, literature [56] provides a comprehensive review of RGB-D semantic segmentation, presenting various deep learning methods and their effectiveness in different application scenarios. This offers valuable insights for further optimizing the MCADFusion method.

Moreover, MCADFusion performs well in tasks involving the fusion of point clouds and images, which are relevant to autonomous driving and robot navigation. For instance, in the RGB-IR pedestrian re-identification task [57], the method improves re-identification accuracy by fusing RGB and infrared image features. Similarly, in the 3D pedestrian re-identification task [58], combining global semantic guidance with local feature aggregation strategies further enhances the fusion of different modal data. Future research can explore the application of the MCADFusion method in more practical scenarios, aiming for breakthroughs across various fields.

5. Conclusions

This paper presents a novel multi-scale convolutional attention decomposition method for enhanced infrared and visible light image fusion in computer graphics (MCADFusion). The method fuses the thermal radiation information from infrared images and the detail information from visible images by fusing multi-scale convolution, attention mechanism, and decomposition techniques to achieve high-quality fusion of infrared and visible images. Among the infrared and visible image fusion tasks, significant advantages are demonstrated, and experiments on publicly available TNO and MSRS datasets show that the method outperforms existing methods in both subjective visual assessment and objective assessment.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This study was financially supported by the following projects: Science, and Technology Research Project of Henan Province (No. 242102211110, No. 242102210217); the Key Research Project of Higher Education Institutions of Henan Province (No. 24A510013); and the Open Project of Key Laboratory of Microsystems Technology (No. 6142804231002).

W.Z.W and M.H.D proposed and conceptualized the innovations of this paper. M.H.D and B.Z prepared the experimental data and verified the experimental results in the paper. M.H.D and C.H.W were responsible for the overall framework, implementation, etc., and evaluated the performance of the proposed methodology in this paper. All authors reviewed the manuscript.

The data in this study are available upon request to the second author.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

1. B. Meher, S. Agrawal, R. Panda, A. Abraham, A survey on region based image fusion methods, *Inf. Fusion*, **48** (2019), 119–132. https://doi.org/10.1016/j.inffus.2018.07.010
2. S. Karim, G. Tong, J. Li, A. Qadir, U. Farooq, Y. Yu, Current advances and future perspectives of image fusion: A comprehensive review, *Inf. Fusion*, **90** (2023), 185–217. https://doi.org/10.1016/j.inffus.2022.09.019
3. K. Ma, Z. Duanmu, H. Yeganeh, Z. Wang, Multi-exposure image fusion by optimizing a structural similarity index, *IEEE Trans. Comput. Imaging*, **4** (2018), 60–72. https://doi.org/10.1109/TCI.2017.2786138
4. X. Zhang, Deep learning-based multi-focus image fusion: A survey and a comparative study, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 4819–4838. https://doi.org/10.1109/TPAMI.2021.3078906
5. J. Liu, X. Fan, J. Jiang, R. Liu, Z. Luo, Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion, *IEEE Trans. Circuits Syst. Video Technol.*, **32** (2022), 105–119. https://doi.org/10.1109/TCSVT.2021.3056725
6. H. Xu, J. Ma, J. Jiang, X. Guo, H. Ling, U2Fusion: A unified unsupervised image fusion network, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 502–518. https://doi.org/10.1109/TPAMI.2020.3012548
7. Z. Zhao, S. Xu, C. Zhang, J. Liu, J. Zhang, P. Li, DIDFuse: Deep image decomposition for infrared and visible image fusion, in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, (2021), 970–976. http://doi.org/10.24963/ijcai.2020/135

8. W. G. C. Bandara, V. M. Patel, Hypertransformer: A textural and spectral feature fusion transformer for pansharpening, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 1767–1777. https://doi.org/10.1109/CVPR52688.2022.00181
9. S. Xu, J. Zhang, Z. Zhao, K. Sun, J. Liu, C. Zhang, Deep gradient projection networks for pansharpening, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 1366–1375. https://doi.org/10.1109/CVPR46437.2021.00142
10. Z. Zhao, J. Zhan, S. Xu, K. Sun, L. Huang, J. Liu, et al., FGF-GAN: A lightweight generative adversarial network for pansharpening via fast guided filter, in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, (2021), 1–6. https://doi.org/10.1109/ICME51207.2021.9428272
11. D. Cheng, R. Liao, S. Fidler, R. Urtasun, DARNet: Deep active ray network for building segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 7431–7439. https://doi.org/10.1109/CVPR.2019.00761
12. H. Qin, M. Zhang, Y. Ding, A. Li, Z. Cai, Z. Liu, et al., BiBench: Benchmarking and analyzing network binarization, in *Proceedings of the 40th International Conference on Machine Learning*, (2023), 28351–28388.
13. C. He, K. Li, Y. Zhang, Y. Zhang, Z. Guo, X. Li, et al., Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects, preprint, arXiv:2308.03166.
14. C. He, K. Li, Y. Zhang, L. Tang, Y. Zhang, Z. Guo, et al., Camouflaged object detection with feature decomposition and edge reconstruction, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2023), 22046–22055. https://doi.org/10.1109/CVPR52729.2023.02111
15. C. He, K. Li, Y. Zhang, G. Xu, L. Tang, Y. Zhang, et al., Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping, in *Advances in Neural Information Processing Systems*, **36** (2023).
16. J. Wang, Z. Yin, P. Hu, A. Liu, R. Tao, H. Qin, et al., Defensive patches for robust recognition in the physical world, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 2456–2465. https://doi.org/10.1109/CVPR52688.2022.00249
17. H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, J. C. Ye, Diffusion posterior sampling for general noisy inverse problems, preprint, arXiv:2209.14687.
18. X. Deng, P. L. Dragotti, Deep convolutional neural network for multi-modal image restoration and fusion, *IEEE Trans. Pattern Anal. Mach. Intell.*, **43** (2021), 3333–3348. https://doi.org/10.1109/TPAMI.2020.2984244
19. A. Ben Hamza, Y. He, H. Krim, A. Willsky, A multiscale approach to pixel-level image fusion, *Integr. Comput.-Aided Eng.*, **12** (2005), 135–146. https://doi.org/10.3233/ICA-2005-12201
20. J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, et al., Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 5802–5811. https://doi.org/10.1109/CVPR52688.2022.00571

21. L. Tang, Y. Deng, Y. Ma, J. Huang, J. Ma, Superfusion: A versatile image registration and fusion network with semantic awareness, *IEEE/CAA J. Autom. Sin.*, **9** (2022), 2121–2137. https://doi.org/10.1109/JAS.2022.106082

22. H. Li, X. J. Wu, J. Kittler, RFN-Nest: An end-to-end residual fusion network for infrared and visible images, *Inf. Fusion*, **73** (2021), 72–86. https://doi.org/10.1016/j.inffus.2021.02.023

23. X. Li, X. Guo, P. Han, X. Wang, H. Li, T. Luo, Laplacian redecomposition for multimodal medical image fusion, *IEEE Trans. Instrum. Meas.*, **69** (2020), 6880–6890. https://doi.org/10.1109/TIM.2020.2975405

24. H. Li, X. J. Wu, J. Kittler, Infrared and visible image fusion using a deep learning framework, in *2018 24th International Conference on Pattern Recognition (ICPR)*, (2018), 2705–2710. https://doi.org/10.1109/ICPR.2018.8546006

25. H. Li, X. J. Wu, DenseFuse: A fusion approach to infrared and visible images, *IEEE Trans. Image Process.*, **28** (2019), 2614–2623. https://doi.org/10.1109/TIP.2018.2887342

26. S. Li, X. Kang, J. Hu, Image fusion with guided filtering, *IEEE Trans. Image Process.*, **22** (2013), 2864–2875. https://doi.org/10.1109/TIP.2013.2244222

27. M. Li, Y. Dong, X. Wang, Pixel level image fusion based the wavelet transform, in *2013 6th International Congress on Image and Signal Processing (CISP)*, **2** (2013), 995–999. https://doi.org/10.1109/CISP.2013.6745310

28. Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, *Inf. Fusion*, **36** (2017), 191–207. https://doi.org/10.1016/j.inffus.2016.12.001

29. J. Ma, H. Zhang, Z. Shao, P. Liang, H. Xu, GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion, *IEEE Trans. Instrum. Meas.*, **70** (2021), 1–14. https://doi.org/10.1109/TIM.2020.3038013

30. H. Li, X. J. Wu, T. Durrani, NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models, *IEEE Trans. Instrum. Meas.*, **69** (2020), 9645–9656. https://doi.org/10.1109/TIM.2020.3005230

31. S. Huang, Y. Yang, X. Jin, Y. Zhang, Q. Jiang, S. Yao, Multi-sensor image fusion using optimized support vector machine and multiscale weighted principal component analysis, *Electronics*, **9** (2020), 1531. https://doi.org/10.3390/electronics9091531

32. R. Liu, J. Liu, Z. Jiang, X. Fan, Z. Luo, A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion, *IEEE Trans. Image Process.*, **30** (2021), 1261–1274. https://doi.org/10.1109/TIP.2020.3043125

33. J. Ma, L. Tang, M. Xu, H. Zhang, G. Xiao, STDFusionNet: An infrared and visible image fusion network based on salient target detection, *IEEE Trans. Instrum. Meas.*, **70** (2021), 1–13. https://doi.org/10.1109/TIM.2021.3075747

34. D. Wang, J. Liu, X. Fan, R. Liu, Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration, preprint, arXiv:2205.11876.

35. J. Wang, Z. Wei, T. Zhang, W. Zeng, Deeply-fused nets, preprint, arXiv:1605.07716.

36. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., Going deeper with convolutions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 1–9. https://doi.org/10.1109/CVPR.2015.7298594

37. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, preprint, arXiv:1706.03762.

38. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16x16 words: Transformers for image recognition at scale, preprint, arXiv:2010.11929.

39. Z. Wu, Z. Liu, J. Lin, Y. Lin, S. Han, Lite transformer with long-short range attention, preprint, arXiv:2004.11886.

40. S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M. H. Yang, restormer: Efficient transformer for high-resolution image restoration, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 5728–5739. https://doi.org/10.1109/CVPR52688.2022.00564

41. X. Lin, S. Sun, W. Huang, B. Sheng, P. Li, D. D. F. Feng, EAPT: Efficient attention pyramid transformer for image processing, *IEEE Trans. Multimedia*, **25** (2023), 50–61. https://doi.org/10.1109/TMM.2021.3120873

42. L. Tang, J. Yuan, J. Ma, Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network, *Inf. Fusion*, **82** (2022), 28–42. https://doi.org/10.1016/j.inffus.2021.12.004

43. Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, et al., CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2023), 5906–5916. https://doi.org/10.1109/CVPR52729.2023.00572

44. L. Tang, J. Yuan, H. Zhang, X. Jiang, J. Ma, PIAFusion: A progressive infrared and visible image fusion network based on illumination aware, *Inf. Fusion*, **83** (2022), 79–92. https://doi.org/10.1016/j.inffus.2022.03.007

45. A. Toet, M. A. Hogervorst, Progress in color night vision, *Opt. Eng.*, **51** (2012), 010901. https://doi.org/10.1117/1.OE.51.1.010901

46. H. Zhang, J. Ma, SDNet: A versatile squeeze-and-decomposition network for real-time image fusion, *Int. J. Comput. Vision*, (2021), 1–25. https://doi.org/10.1007/s11263-021-01501-8

47. P. Liang, J. Jiang, X. Liu, J. Ma, Fusion from decomposition: A self-supervised decomposition approach for image fusion, in *European Conference on Computer Vision (ECCV)*, (2022), 719–735. https://doi.org/10.1007/978-3-031-19797-0_41

48. Z. Huang, J. Liu, X. Fan, R. Liu, W. Zhong, Z. Luo, Reconet: Recurrent correction network for fast and efficient multi-modality image fusion, in *European Conference on Computer Vision*, Springer, (2022), 539–555. https://doi.org/10.1007/978-3-031-19797-0_31

49. W. Tan, H. Zhou, J. Song, H. Li, Y. Yu, J. Du, Infrared and visible image perceptive fusion through multi-level gaussian curvature filtering image decomposition, *Appl. Opt.*, **58** (2019), 3064–3073. https://doi.org/10.1364/AO.58.003064

50. W. Tang, F. He, Y. Liu, Y. Duan, T. Si, DATFuse: Infrared and visible image fusion via dual attention transformer, *IEEE Trans. Circuits Syst. Video Technol.*, **33** (2023), 3159–3172. https://doi.org/10.1109/TCSVT.2023.3234340

51. W. Wang, U. Neumann, Depth-aware CNN for RGB-D segmentation, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 135–150. https://doi.org/10.1007/978-3-030-01252-6_9

52. D. Lin, G. Chen, D. Cohen-Or, P. A. Heng, H. Huang, Cascaded feature network for semantic segmentation of RGB-D images, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2017), 1311–1319. https://doi.org/10.1109/ICCV.2017.147

53. G. Qi, Y. Zhang, K. Wang, N. Mazur, Y. Liu, D. Malaviya, Small object detection method based on adaptive spatial parallel convolution and fast multi-scale fusion, *Remote Sens.*, **14** (2022), 420. https://doi.org/10.3390/rs14020420

54. Z. Zhu, H. Yin, Y. Chai, Y. Li, G. Qi, A novel multi-modality image fusion method based on image decomposition and sparse representation, *Inf. Sci.*, **432** (2018), 516–529. https://doi.org/10.1016/j.ins.2017.09.010

55. K. Wang, M. Zheng, H. Wei, G. Qi, Y. Li, Multi-modality medical image fusion using convolutional neural network and contrast pyramid, *Sensors*, **20** (2020), 2169. https://doi.org/10.3390/s20082169

56. C. Wang, C. Wang, W. Li, H. Wang, A brief survey on RGB-D semantic segmentation using deep learning, *Displays*, **70** (2021), 102080. https://doi.org/10.1016/j.displa.2021.102080

57. Y. Chen, L. Wan, Z. Li, Q. Jing, Z. Sun, Neural feature search for RGB-infrared person re-identification, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 587–597. https://doi.org/10.1109/CVPR46437.2021.00065

58. C. Wang, X. Ning, W. Li, X. Bai, X. Gao, 3D person re-identification based on global semantic guidance and local feature aggregation, *IEEE Trans. Circuits Syst. Video Technol.*, **34** (2024), 4698–4712. https://doi.org/10.1109/TCSVT.2023.3328712

AIMS Press