IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 January 2024, accepted 24 January 2024, date of publication 19 February 2024, date of current version 23 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3359715

RESEARCH ARTICLE

MDDCMA: A Distributed Image Fusion Framework Based on Multiscale Dense Dilated Convolution and Coordinate Mean Attention

TONG TONG[1], AIPING YE[1], YONGQI LU[2], AND ZHENLU WU[1]
[1]Faculty of Mathematics and Computer Science, Guangdong Ocean University, Zhanjiang 524088, China
[2]School of Electronics and Information Engineering, Guangdong Ocean University, Zhanjiang 524088, China

Corresponding author: Zhenlu Wu (zlwu@gdou.edu.cn)

This work was supported in part by the Guangdong University Student Science and Technology Innovation Cultivation Special Fund Support Project under Grant pdjh2023a0243, and in part by the Undergraduate Innovation Team Project of Guangdong Ocean University under Grant CXTD2024011.

ABSTRACT In the task of fusing infrared and visible images, the extraction of features and fusion strategy significantly impacts the outcome of the fusion. However, prevailing fusion methods are often manually designed, unlearnable, and neglect to consider context adequately. To address these issues, this paper introduces a distributed architecture network based on attention mechanism and dense dilated convolution, realizing three-channel data fusion. This network employs a distributed fusion framework to fully utilize the fusion output of the previous step, capitalizing effectively on the target and texture information in infrared and visible images. Initially, two channels gather ample context from the source images through a dense dilated convolution module with multiscale channel attention. Subsequently, a fusion strategy based on coordinate mean attention is adopted to facilitate the fusion of results between the two channels. Then, the fused features and the preceding fusion results are fed into the fusion channel, minimizing loss of target and texture information in infrared and visible images. Furthermore, we incorporate an edge correction block, capable of refining the edge details of the fusion results and effectively suppressing noise. The proposed method demonstrates good fusion performance and extensive ablation experiments validate the effectiveness of the proposed methodology. Simultaneously, both subjective qualitative and objective quantitative comparison results, conducted on public datasets such as RoadScene, TNO, and MSRS, indicate that the visual quality and evaluation metrics of our fusion images are comparable to those achieved by the state-of-the-art image fusion methods.

INDEX TERMS Distributed architecture, dilated convolution, infrared and visible image fusion, edge correction, coordinate attention.

LIST OF ABBREVIATIONS

CNN	Convolutional neural network
GAN	Generative adversarial network
AE	Autoencoder
PID	Proportional-Integral-Derivative control systems
MDDC	Multiscale dense dilated convolution module
CA	Coordinate attention
CMA	Coordinate mean attention block
CMB	Coordinate mean attention fusion block
ECB	Edge correction block
MAT	Multiscale channel attention block
SAM	Spatial attention module
CAM	Channel attention module
SCM	Spatial and channel attention fusion module

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

I. INTRODUCTION

Image fusion is a critical image processing technique designed to address the challenge of providing a comprehensive depiction of a scene, a task that single-mode sensors struggle to accomplish due to limitations in hardware, theory, and technology. Among the current image fusion scenarios, the most prevalent involves the fusion of visible and infrared images. Infrared imagery can reveal the thermal distribution of objects, enabling effective target emphasis even under adverse weather conditions. However, infrared images lack clarity concerning background details. On the other hand, visible images offer rich color information and high-resolution texture details, but their visual quality and resolution are highly dependent on lighting and environmental conditions. By fusing infrared and visible images, one can combine the rich detail of visible image with the target characteristics of infrared image to create a composite image. This image contains more useful information than any single source image and significantly aids subsequent high-level visual tasks such as object detection [1], [2], object tracking [3], [4], and semantic segmentation [5], [6].

Over the past several decades, a number of image fusion methods have been proposed, broadly classified into two categories: traditional methods and deep learning-based methods. Classic traditional fusion methods include multiscale transform methods [7], [8], [9], [10], sparse representation methods [11], [12], [13], subspace methods [14], [15], total variation methods [16], and various hybrid methods [8], [17]. These primarily employ relevant mathematical transformations to manually analyze the activity level of source image information and design fusion rules in the spatial or transform domain. Existing traditional methods exhibit excellent performance in image fusion applications. However, their disadvantages are becoming increasingly pronounced [18], such as the lack of abbility of direct evaluation and learning from data and the application of the same transformations to extract features from different source images, disregarding the feature differences between the source images. Moreover, most fusion rules of traditional fusion methods are manually designed, often overly simple or superficial, potentially leading to subjective bias and subpar fusion results.

In recent years, researchers have introduced deep learning methods, currently the most popular, to overcome the limitations of traditional image fusion methods. Typical methods based on deep learning include those based on Autoencoder (AE), Generative Adversarial Network (GAN), Convolutional Neural Network (CNN) and transformer-based method. All these methods aim to solve three sub-problems of image fusion, namely feature extraction, feature fusion, and image reconstruction. Specifically, AE-based methods such as Densefuse [19] and Nestfuse [20], use pre-trained autoencoders to accomplish the fusion task. Here the encoder is responsible for feature extraction, the decoder for image reconstruction, and the intermediate feature fusion is achieved according to some simple fusion rules, such as element-wise addition and concatenate operation. CNN was initially combined with traditional methods for image fusion applications [18]. Specifically, CNN-based methods integrate CNN with traditional methods for image fusion, utilizing traditional techniques for feature extraction and image reconstruction, while employing well-trained CNN to establish the fusion rules. For instance, Liu et al. [21] proposed a CNN-based method for infrared and visible image fusion, where a pre-trained CNN is used to generate fusion weights, and the laplacian pyramid method is used for image fusion. There is another type of method involves the use of CNN [18], [22], [23]. Guided by meticulously designed loss functions and network structures, these approaches achieve end-to-end feature extraction, feature fusion, and image reconstruction. Zhang et al. [24] developed a general image fusion framework based on CNN, which includes multiple CNN blocks for feature extraction and image reconstruction. Subsequently, feature fusion is achieved through a simple element-wise mean method. Xu et al. [25] proposed a universal end-to-end CNN for image fusion, named U2Fusion. U2Fusion introduced an information measurement method to evaluate the essential information of images, which is utilized in conjunction with the VGG model to measure the amount of information in each modality image, aiding in the training of CNN networks. These methods achieved fully convolutional image fusion and yielded impressive results. However, these methods still have some drawbacks: (1) These methods perform feature extraction at a single scale, neglecting multiscale local/global information, which can reduce the quality of fusion to some extent. (2) CNN-based approaches typically use only the final output of the feature extraction layers as input for the fusion layers, potentially leading to significant loss of information extracted by the convolutional layers. (3) The fusion strategies adopted in these methods are mostly manually designed, which greatly limits the enhancement of fusion results.

Consequently, fusion methods based on GAN and methods based on Transformer are proposed. Methods based on GAN construct a generator and a discriminator, estimating the probability distribution of pixels in the fused image through adversarial learning. The advantages of this method include the ability to achieve end-to-end fusion and to eliminate the constraints of manually designing fusion rules. FusionGAN [26] is the first GAN-based model for fusing infrared and visible images. It generates fused images by simultaneously inputting infrared and visible images into the generator, while the discriminator, through evaluating and learning from the visible image, arrives at the classification results between the two types of image information. However, due to the limited ability of a single discriminator to analyze multi-modal data, the authors of FusionGAN proposed DDcGAN [27], which evaluates feature information through two discriminators that assess the infrared and visible images separately. However, due to the absence of an ideal fused

image, balancing the generator and discriminator during the training process poses a significant challenge. Therefore, the performance of these fusion methods still exhibits certain deficiencies. The new generation neural network framework, Transformer architecture, has been applied to image fusion, achieving remarkable results. These methods utilize the self-attention mechanism to capture global dependencies, facilitating effective feature representation learning. Ma et al. [28] proposed a multi-task fusion model based on cross-domain distant learning, modeling both intra-domain and cross-domain distant dependencies to better integrate complementary features. Tang et al. [29] developed a model that combines CNN and Transformer, efficiently integrating global complementary information with local details. However, Transformer-based methods require significant computational resources and may perform inefficiently on devices with limited resources [29], [30].

Recently, Proportional-Integral-Derivative (PID) control systems have been introduced as an image fusion strategy within a multiscale framework [31]. PID control systems measure the difference between the fusion result and the source images in real-time and adjust the weights of the function accordingly, enabling adaptive fusion capabilities. This allows source mappings to guide the fusion process, avoiding the need for manually designed complex fusion rules. However, the effectiveness of PID control systems largely depends on precise tuning of their parameters. Finding the optimal settings for the proportional, integral, and derivative components is a complex and time-consuming task [32].

This paper introduces a novel end-to-end method for infrared and visible image fusion, named MDDCMA, designed to overcome the aforementioned limitations. MDDCMA adopts a three-channel distributed fusion framework, where two channels receiving source images are used for feature extraction, and the other channel is used to fuse these extracted features. This structure is able to effectively utilize information from the previous step, preserving more comprehensive source image information for the fusion results. We construct a Multiscale Dense Dilated Convolution Module (MDDC) and Edge Correction Block (ECB) in the two feature extraction channels, which can help the network to extract multiscale global features and improve the edges of coarse features to achieve better fusion results. Simultaneously, a meticulously designed module, Coordinate Mean Attention Fusion Block (CMB), is utilized to precisely fuse the information from feature extraction channels. In the section IV, we conducted ablative analyses and supplementary experiments on ECB and CMB to validate the performance of the proposed module.

The contributions of this paper can be summarized as follows:

- We propose an end-to-end framework for the fusion of infrared and visible images. This framework adopts a distributed structure, implemented through three branches for image feature extraction and fusion. Two branches are applied to feature extraction while the other one handles feature fusion. This structure enables maximal utilization of the output from the previous fusion step, thereby enhancing the fusion quality.
- In a bid to capture comprehensive feature information, we have designed a MDDC for feature extraction. The dense dilated convolution structure extracts features under multiple receptive fields. Furthermore, a multi-scale channel attention block (MAT) and a spatial and channel attention fusion module (SCM) are incorporated to aid in the acquisition and merging of features from different regions, thereby achieving superior feature extraction results.
- To overcome drawbacks associated with manually designed fusion strategies, CMB is designed as the network's fusion strategy. This module is capable of simultaneously considering information from both spatial levels and coordinate positions, effectively enhancing the feature's expressive capacity, and thereby facilitating more precise and efficient feature merging. Additionally, to improve the quality of fusion results, we design an ECB to refine the edge information in the fusion results. It can effectively suppress noise and enhance the texture detail information in the fused images.
- We conducted extensive experiments on publicly available infrared and visible image fusion datasets. The experimental results indicate that, compared to existing fusion methods, our fusion framework exhibits superior performance in both subjective and objective evaluations. Moreover, we conducted ablation experiments, which verified the functionality and effectiveness of the proposed method.

The arrangement for the remaining sections of the paper is as follows. Section II presents a review and introduction of the relevant works related to this study. In Section III, we provide a detailed description of the proposed method. In Section IV, we demonstrate the fusion effects of the proposed method, compare it against advanced image fusion methods quantitatively and qualitatively, and present the results of the ablation experiments and some additional experiments to attest to the method's effectiveness. The conclusion of this paper is given in Section V.

II. RELATED WORKS

In this chapter, we review several techniques that are closely related to our proposed method, which include distributed fusion structures and attention mechanisms, among others.

A. DISTRIBUTED FUSION ARCHITECTURE

Distributed architecture is initially introduced into multi-sensor data fusion to address the drawbacks of traditional centralized fusion frameworks [33]. In distributed fusion, each sensor can have its own processor to fuse local data and cooperate with other sensor nodes. This architecture can fully utilize known prior conditions to maximize the accuracy

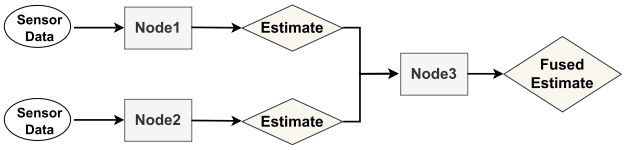

FIGURE 1. Distributed fusion structure of multiple sensors.

of the fusion trajectory [34]. The distributed fusion structure features low latency, high speed, and high survivability, effectively integrating data from different sensors and extracting relevant information about the target [33]. Figure 1 illustrates the principle of the distributed fusion algorithm. Each node generates an optimal estimate given its data and sends it to the fusion node, which subsequently combines the local estimates to achieve the best centralized estimate. Wu et al. [23] proposed a multi-branch image fusion network for multi-focus image fusion. Similarly, Wu et al. [35] proposed a distributed fusion framework specifically for low-resolution multispectral images and panchromatic images. Brännström et al. [36] applied distributed fusion structures to data from ground sensor networks, measuring data using five acoustical sensor nodes, each with three sensors, on five different types of vehicles.

B. ATTENTION MECHANISM IN DEEP LEARNING

The attention mechanism was initially derived from machine translation tasks, but over time, its applications have far exceeded its original realm. Due to its ability to enhance the interpretability of neural networks, the attention mechanism has stood out in the field of artificial intelligence. Specifically, it endows the model with a unique ability to globally scrutinize the entire input sequence and then, based on computed weights, precisely focus on key portions of the input sequence, thus allowing selective attention to important information. In recent years, various variants of attention mechanisms have emerged and rapidly gained prominence in the field of computer vision, demonstrating their capabilities in areas such as object detection [37], [38], semantic segmentation [39], image fusion [40], and image restoration [41], [42]. Liu and Liu [38] proposed a new attention-based feature aggregation module that processes and fuses features extracted at multiple levels, resulting in a feature map that aggregates both high-level and low-level features. Fu et al. [39] proposed a dual attention network that captures spatial dependencies between any two positions in the feature map through two parallel self-attention modules, thereby obtaining rich contextual dependencies. Li et al. [40] achieved extensive mapping attention feature maps by applying attention mechanisms at multiple scales. Suganuma et al. [41] introduced a straightforward yet efficient neural network layer structure wherein multiple operations, governed by an attention mechanism, are stacked in parallel. This configuration enables the selection of appropriate operations based on the input. Qin et al. [42] proposed an end-to-end feature fusion attention network for defogging image restoration, which treats features and pixels unequally through a novel feature attention module, helping to assign more weight to important features.

III. METHODS

This section provides a detailed introduction to the framework for MDDCMA. Firstly, we introduce the overall structure of the fusion network. Then, we describe in sequence the network structures of the MDDC, CMB, ECB, and other modules. Lastly, we present the detailed design of the loss function and its mathematical representation.

A. OVERALL FRAMEWORK

The overall framework of the proposed MDDCMA is shown in the bottom-left corner of Figure 2. This network is an end-to-end network composed of three channels. It includes two feature extraction channels and one fusion channel. In MDDCMA, two branches incorporating MDDC and ECB are used for feature extraction from visible and infrared images. The middle branch fuses the features extracted from the two branches with the results of the previous step in a layer-by-layer manner, with the final layer generating a fused image. MDDCMA employs a quad-layer network structure, continuously extracting deeper features. After the image is input into the feature extraction channel, feature extraction is completed through the MDDC. Within the MDDC, the image is first processed by a MAT to obtain an enhanced feature, which is then fed into four different convolution branches to acquire multiscale features through convolution kernels with different receptive fields. Subsequently, the features extracted from each branch are aggregated through SCM to preserve as much important information as possible from each convolutional branch. At the end of each layer, the features of the feature extraction channel and the fusion channel are jointly input into the CMB. This module obtains useful information from different features through the Coordinate mean attention block (CMA), amplifies areas of interest, and suppresses unnecessary noise. Then, based on the obtained weighted information, the features extracted from the three channels are fused and used as the input for the next layer of the fusion channel. FM1-FM4, consisting of 3×3 convolution blocks, batch norm, and activation function layers, represent different convolution modules at different layers in the fusion channel, used for processing the features after fusion at each layer. After completing the feature processing of the four layers, a fused image is constructed through a 1×1 convolution layer.

B. MULTISCALE DENSE DILATED CONVOLUTION MODULE

During the fusion process of infrared and visible images, image feature extraction is of paramount importance. This is because the distribution of important information in images is not regular. For example, in infrared images, high-heat targets can appear anywhere in the image. Therefore, it is crucial to extract comprehensive features. Traditional deep learning methods use CNN to automatically learn and

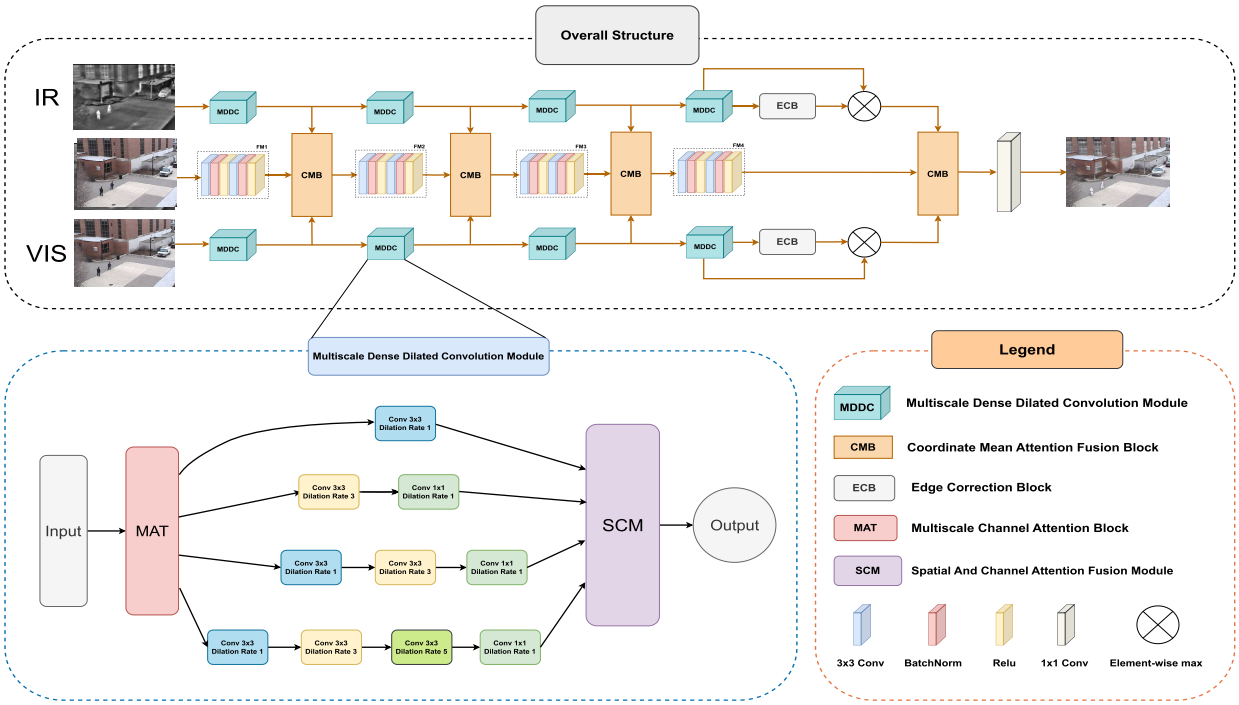

FIGURE 2. The overall network structure of MDDCMA.

extract features from multiple input images. However, due to the fixed receptive field, a single convolution layer can only extract a limited amount of local image features. For instance, convolution layers with smaller receptive fields can extract low-frequency pixels but fail to capture high-frequency content. Similarly, convolution layers with larger receptive fields extract more prominent image features, but not the low-frequency details of the image. To address this, we introduced dilated convolution [43]. Dilated convolution can expand the model's receptive field without changing the size of the image or increasing the amount of parameters. The dilated convolution filter fills zero elements as holes along the spatial dimension of the standard convolution kernel. When the hole rate of dilated convolution is 1, it becomes a standard convolution. The calculation formula for the equivalent convolution kernel in dilated convolution is shown in Equation (1).

$$C = (d-1)(f-1)+f \tag{1}$$

where C represents the effective kernel size and d represents the dilation rate and f represents the standard convolution filter size.

Inspired by dilated convolution, we propose a MDDC which can flexibly increase the receptive field of the network, enabling the extraction of richer features. The structure of the MDDC is shown in the bottom-left of Figure 2, which consists of four branches, each with different receptive fields, used to capture feature map information at different scales. For clarity, we define the convolution branch with a single 3×3 convolution block as the first branch, and the branch with three 3×3 convolution blocks and one 1×1 convolution block as the fourth branch. Therefore, as shown in the bottom-left of the Figure 2, the structural diagram of MDDC is presented with the branches arranged sequentially from top to bottom as the first, second, third, and fourth branches. To enhance the performance of the module, MAT is designed to weight the input features, enabling the network to pay more attention to important areas across multiple scales. Within the MDDC, the input features first pass through MAT, resulting in enhanced features. These features are then input into the four branches, which capture feature map information through branches with different receptive fields. At the end of each convolution branch, a 1×1 convolution is used for linear activation of the feature map, followed by a ReLU activation function to obtain the results of the convolution along each convolution path. Subsequently, SCM is introduced to handle the output of each branch, achieving the fusion of multiscale deep features. The detailed information of all convolutional layers in MDDC of each layer, such as the input channels, the output channels, the list of channel, and the activation functions, is presented in Table 1.

TABLE 1. Input channel, output channel and channel list of each layer's mddc and activation function of all convolutional layers in the mddcma.

Layer	Input Channel	Output Channel	Channel list of MDDC	Activation Function
Layer1	1	64	[1,16,32,48,64]	Relu
Layer2	64	128	[64,80,96,112,128]	Relu
Layer3	128	192	[128,144,160,176,192]	Relu
Layer4	192	256	[192,208,224,240,256]	Relu

Given source images from different angles within the same scene, objects in the image may appear at different locations, resulting in changes in size and shape. Single-scale features are unable to extract all the necessary spatial information, hence the need for a multiscale mechanism to capture multiscale features through different kernel sizes. Inspired by the SENet [44], MAT is proposed to enable the network to focus more on important areas and effectively capture information across multiple scales. The structure of MAT is shown in Figure 3. Features are inputted into MAT, followed by average pooling with 1×1, 2×2, and 4×4 respectively, to generate multiscale features containing more necessary spatial information. Subsequently, a fully connected layer and an activation function are used to calculate the weights of the global features. The weight $W_{t_i}^k$ for the k-th feature $f_{t_i}^k$ of the t-th pooling scale in the MAT at the i-th layer of MDDCMA which can be formulated as Equation (2):

$$W_{t_i}^k = \sigma\left(L_2\left(\sigma\left(L_1 G(f_{t_i}^k)\right)\right)\right) \tag{2}$$

where G(·) represents the global average pooling operation. $L_1 \in \mathbb{R}^{k\times k}$ and $L_2 \in \mathbb{R}^{k\times k}$, σ denotes sigmoid function.

Then, by upsampling the extracted multiscale features and performing element-wise multiplication with the weights, important features within these multiscale features are emphasized. Subsequently, an attention map is obtained through normalization operations. In the attention mapping operation, a maximum selection strategy is adopted to concentrate more attention on the most prominent spatial locations [43]. The attention map can be achieved in Equaltion (3):

$$M_i = \left\{\triangle\left(W_{1_i}^k * \left[f_{1_i}^k\right]\right), \triangle\left(W_{2_i}^k * \left[f_{2_i}^k\right]\right), \triangle\left(W_{3_i}^k * \left[f_{3_i}^k\right]\right)\right\} \tag{3}$$

where $\triangle$ denotes the normalization operation and $\{\cdot,\cdot,\cdot\}$ refers to the operation for obtaining the maximum value in the corresponding channel and position in the feature map. $*$ represents the element-wise multiplication. $[\cdot]$ denotes the upsampling operation and M_i represents the final attention map. This result is the output of MAT.

Given that the extracted features are three-dimensional tensors, we also incorporate a SCM based on Spatial Attention Module (SAM) and Channel Attention Module (CAM) to fully consider information in the spatial and channel dimensions. This enables the fusion of multiscale deep features that have been acquired. Specifically, the results of the first and second convolution branches are fused using the SCM, and the results of the third and fourth convolution branches are also combined through SCM. Subsequently, the two fused results are further fed into SCM for aggregation, thereby yielding the final features. The information merged is the output of the MDDC. The structure of SCM is illustrated in Figure 4.

In Figure 4, $\hat{\Phi}_1^n$ and $\hat{\Phi}_2^n$ represent the feature extraction results from two different branches in MDDC, where n∈{1,2,3,4} denotes the layer in MDDCMA. After feeding the inputs into SAM and CAM, the corresponding spatial and channel attention weights are obtained. The computation processes for channel attention weight and spatial attention weight on input features are as follows:

$$\hat{\beta}_k^n = \partial\left(\text{Conv}\left(P\left(\hat{\Phi}_k^n\right)\right)\right) \tag{4}$$

$$\hat{\alpha}_k^n = \partial\left(\text{Conv}\left(L\left(\hat{\Phi}_k^n\right)\right)\right) \tag{5}$$

where k∈{1,2}, Conv denotes the convolution operation, $\partial(\cdot)$ represents the ReLU activation function, $P(\cdot)$ signifies the global pooling operation, and $L(\cdot)$ indicates the l1-norm operation. $\hat{\beta}_k^n$ and $\hat{\alpha}_k^n$ are respectively the channel attention weight and the spatial attention weight. Then, the weighting maps are calculated by soft-max operation from attention weights, this process is as follows:

$$\hat{\theta}_k^n = \frac{\hat{\beta}_k^n}{\sum_{i=1}^M \hat{\beta}_i^n} \tag{6}$$

$$\hat{\delta}_k^n = \frac{\hat{\alpha}_k^n}{\sum_{i=1}^M \hat{\alpha}_i^n} \tag{7}$$

where $M = 2$. In the end, the enhanced feature maps $\widehat{\Phi}_{fc}^n$ and $\widehat{\Phi}_{fs}^n$ can be represented by Equations (8) and (9), respectively.

$$\hat{\Phi}_{fc}^n = \sum_{i=1}^{M} (\hat{\theta}_i^n \times \hat{\Phi}_i^n) \tag{8}$$

$$\hat{\Phi}_{fs}^n = \sum_{i=1}^{M} (\hat{\delta}_i^n \times \hat{\Phi}_i^n) \tag{9}$$

Once $\widehat{\Phi}_{fc}^n$ and $\widehat{\Phi}_{fs}^n$ are obtained, the final features are generated according to Equation (10).

$$\hat{\Phi}_f^n = \left(\hat{\Phi}_{fc}^n + \hat{\Phi}_{fs}^n\right) \times 0.5 \tag{10}$$

Under the collective influence of the aforementioned modules, MDDC is capable of extracting abundant feature information, laying the groundwork for the subsequent fusion results to retain rich feature information from the source images.

C. COORDINATE MEAN ATTENTION FUSION BLOCK

After feature extraction, a fusion strategy is employed to merge the extracted multiscale features. Effective fusion of the features are crucial for the restoration of high-quality fused images. A good fusion strategy can preserve more information-rich image features during the fusion process. Current mainstream fusion strategies mainly involve direct fusion of extracted convolution features through elemental fusion strategies, such as element-wise addition [19], element-wise mean, and element-wise maximum [22]. However, elemental fusion strategies belong to a rather rudimentary fusion method. This method equally merges the feature maps of multiple input images without fully

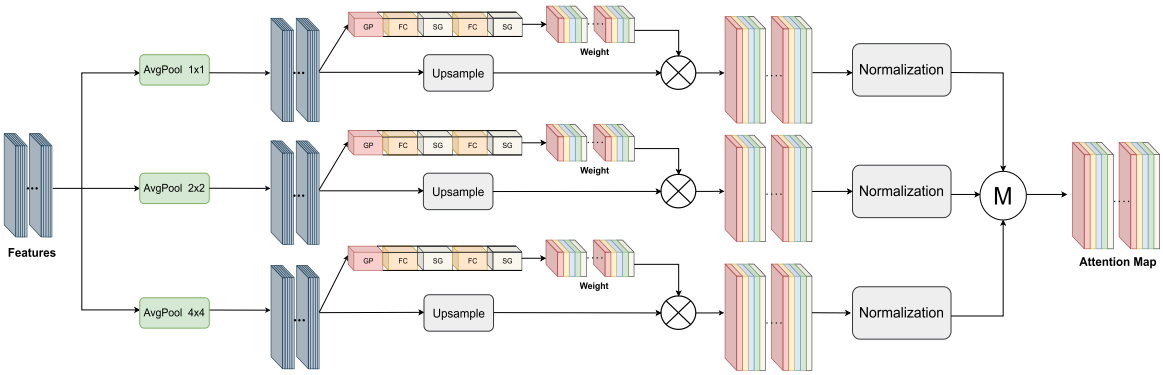

FIGURE 3. The network structure of multiscale channel attention block. GP, FC, SG represent the global average pooling operation, fully connected layer, and sigmoid function respectively. ⊗ denotes the element-wise multiplication. Ⓜ denotes the element-wise max operation.

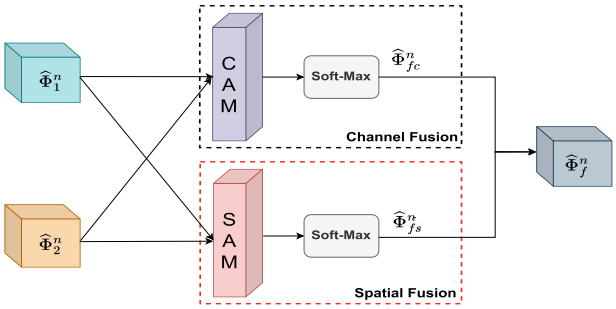

FIGURE 4. The procedure of spatial and channel attention fusion module.

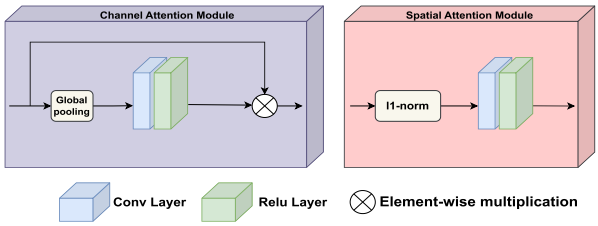

FIGURE 5. The network structure of spatial attention module and channel attention module.

taking into account the unique differences between these feature maps. Moreover, elemental fusion strategies are unlearnable, thus limiting the performance of the fusion results. The attention mechanism can effectively improve the performance of tasks based on CNN, amplifying areas of interest, ignoring other irrelevant noise, and allowing the network model to focus only on parts beneficial to improving the results. Therefore, to enhance the performance of the fusion network, inspired by the coordinate attention block (CA) [45], we propose a novel coordinate mean attention block as the fusion strategy, enabling the network to pay more attention to important information features. In CMB, we use CMA to guide feature learning so that the network can pay more detailed and extensive attention to spatial features, facilitating the aggregation of input features more precisely. The coordinate attention block can effectively aggregate the relevant information of the two coordinates of the feature map using a one-dimensional average pooling method. It has been proven to be very suitable in the field of image processing. However, its extraction of spatial features is not comprehensive, leaving room for improvement. Therefore, we improved the original network structure and designed CMA. The network structure of CMA is shown in Figure 6. On the basis of the original CA using a one-dimensional average pooling operation, we parallel a one-dimensional maximum pooling operation to enhance feature extraction. While average pooling can pay attention to the central area of the significant target, maximum pooling can focus on the edge area with prominent changes.

CMA first performs two parallel one-dimensional pooling operations on the input X ($X \in \mathbf{R}^{CxHxW}$ represents the feature map, and C, H, W represent the number of channels, height, and width of the feature map, respectively) and merges them, followed by decomposing the fused features. Then, the two extracted feature vectors containing coordinate attention are multiplied by the original input X to locate the attention. The pooling methods are one-dimensional average pooling and one-dimensional maximum pooling. Each parallel pooling method is performed in two directions: pooling along the width direction and pooling along the height direction. The result of pooling is transposed and concatenated for ease of subsequent processing. Then, a 1×1 convolution operation, batch normalization, and a non-linear activation operation are performed on the feature map obtained after concatenation. The two activated feature maps are fused through an element-wise maximum operation that extracts the maximum values from the corresponding channels and positions in the feature maps, resulting in feature maps that contain important information in the central and edge areas. The feature vectors in the height and width directions are obtained by decomposing the fused feature in the spatial dimension. Subsequently, two 1×1 convolution transformations and non-linear activation operations are performed on these two vectors to obtain the attention weights in the height and width directions. Finally, the original feature map is weighted with

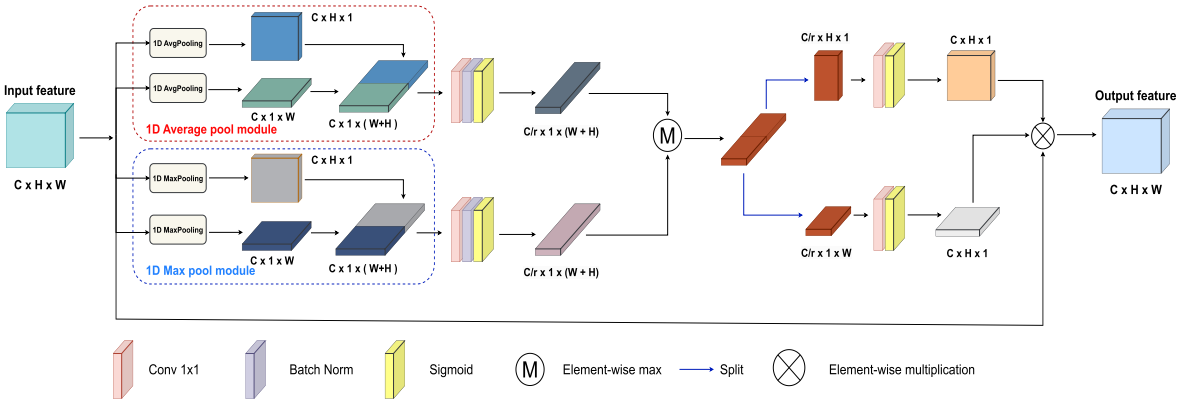

FIGURE 6. The network structure of coordinate mean attention block.

the two types of attention weights to generate the attention map.

After obtaining the feature map through one-dimensional average pooling and max pooling, convolution and feature activation are applied to the feature map. This process can be expressed by Equation (11) and Equation (12):

$$F_{\text{avg}} = \partial\left(\text{Conv}_1\left(\left[avgz^H; avgz^W\right]\right)\right) \tag{11}$$

$$F_{\max} = \partial\left(\text{Conv}_1\left(\left[maxz^H; maxz^W\right]\right)\right) \tag{12}$$

where ∂ represents the ReLU activation function, Conv_1 denotes a 1×1 convolution, and $avgz^H$, $avgz^W$, $maxz^H$ and $maxz^W$ represent the feature maps obtained after pooling operations along the height and width directions of the feature map. $[\cdot;\cdot]$ signifies transposition followed by concatenate operation along the spatial dimension. Subsequently, the element-wise maximum operation is used to fuse the features obtained from average pooling and maximum pooling, as shown in Equation (13)

$$F = \max\left(F_{\text{avg}}, F_{\max}\right) \tag{13}$$

Then, the feature map is decomposed into two feature vectors along the spatial dimensions. The generated pair of feature tensors are then transformed back to the original channel size using a 1×1 convolution. Ultimately, the positional information weights of the input features in the two directions are obtained. They are illustrated as follows:

$$S^H = \sigma\left(\text{Conv}_1\left(F^H\right)\right) \tag{14}$$

$$S^W = \sigma\left(\text{Conv}_1\left(F^W\right)\right) \tag{15}$$

where $F^H \in \mathbf{R}^{(C/r)\times H}$ and $F^W \in \mathbf{R}^{(C/r)\times W}$ represent two decomposed vectors associated with the height and width of F respectively. σ denotes the sigmoid activation function, and Conv_1 indicates a 1×1 convolution.

Finally, the original feature map is multiplied by the two types of attention weights to produce the final attention map. The formula for the output feature map of CMA is shown in Equation (16):

$$O(i,j) = X(i,j) \times S^H(i) \times S^W(j) \tag{16}$$

where i∈{1, 2, …, H}, j∈{1, 2, …, W}. The network structure of CMB is shown in Figure 7. Input the two features that need to be fused into CMB, obtain the weight attention map through CMA. Subsequently, perform a soft-max operation on these features to adaptively generate respective probability weights, and multiply them with the inputs to weight the important features. This process can be expressed by Equation (17):

$$X_{out} = X_A \times \omega_A + X_B \times \omega_B \tag{17}$$

where ω_A and ω_B are attention weights generated by soft-max after the inputs undergo processing through CMA. X_{out} is the final output result. The feature map processed by the CMA possesses spatial hierarchical information along the height direction and coordinate position information along the width direction. simultaneously. Compared to the original coordinate attention block, the addition of a parallel one-dimensional maximum pooling module in the coordinate mean attention block does not significantly increase the model's parameter quantity. However, it can extract richer edge feature information. Therefore, by enhancing features through the coordinate mean attention block, it is possible to comprehensively capture the information of the input features, improving fusion quality.

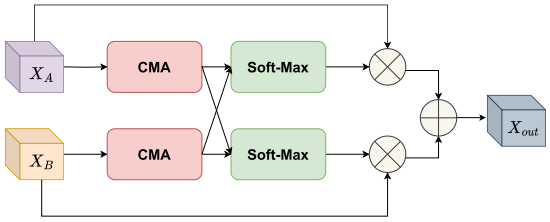

FIGURE 7. The network structure of coordinate mean attention fusion block. ⊗ represents element-wise multiplication operation, ⊕ indicates element-wise addition operation.

D. EDGE CORRECTION BLOCK

When we talk about edge information in an image, we are actually referring to those features in the image that exhibit sudden changes at certain locations. These changes could involve abrupt variations in grayscale, brightness, color, or texture. Within an image, edge information essentially serves as key markers, helping us in distinguishing between different objects or regions in the image. Therefore, before outputting the fused image, the coarse features extracted by the neural network are fed into an edge correction module to adjust the edge textures, reducing the impact of noise on the image and improving the quality of fusion. As shown in Figure 2, after the fourth layer's MDDC completes feature extraction, the output features are individually placed into ECB for edge revision. First, edge information for image correction is obtained. The edge gradient map ∇E, of size m × n, can be represented by Equaltion (18)

$$\nabla E = \sum_{mn}^{i=1} \sqrt{\left(\nabla_i^h \mathbf{u}\right)^2 + \left(\nabla_i^v \mathbf{u}\right)^2} \tag{18}$$

where $\nabla_i^h \mathbf{u} = \mathbf{u}_i - \mathbf{u}_{l(i)}$ and $\nabla_i^v \mathbf{u} = \mathbf{u}_i - \mathbf{u}_{b(i)}$ represent linear operators used for calculating the first layer of horizontal and vertical differences respectively, while $\mathbf{u}_{l(i)}$ and $\mathbf{u}_{b(i)}$ correspond to the nearest neighbor pixels to the left and below the source pixel i, respectively.

Subsequently, a spatial gradient filter ($\mathcal{S}^{\mathcal{G}}$) is utilized to optimize the texture details of the edges, yielding more pronounced gradient information. The output of $\mathcal{S}^{\mathcal{G}}$ can be represented by Equation (19):

$$\mathcal{S}_{\text{out}}^{\mathcal{G}} = \max_{i \in M} \left(\max_{j \in N} \left(\nabla \boldsymbol{E}(i+1, j+1), \nabla \boldsymbol{E}(i, j) \right) \right) \tag{19}$$

where M = {1, …, m-1}, N = {1, …, n-1}. I is the horizontal pixel of the map, and j is the vertical pixel. The structure of the edge correction block is shown in the Figure 8. The coarse features extracted are input into the edge correction block. After being processed by $\mathcal{S}^{\mathcal{G}}$, the edge features used for repair are obtained. These features are then placed into the connected convolutional layer to enhance the features. Ultimately, the enhanced feature map is combined with the output of the fourth layer from the feature extraction channel.

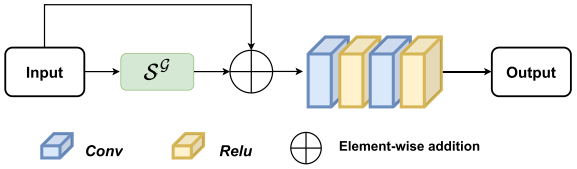

FIGURE 8. The network structure of edge correction block.

E. LOSS FUNCTION

Determining the valid information to retain from the source image is key to the task of image fusion. The content of the loss function determines the optimization direction of the fusion network and the type and proportion of information included in the fused image. Therefore, designing a good loss function is critically important for improving the quality of the fused image. The fusion objective is to enhance the preservation of key information from the source image as much as possible. In this research, the fused image should retain both the texture details in the visible image and highlight the infrared thermal information of the target of interest. Therefore, we adopted intensity loss and detail loss to jointly constrain the fusion target. The total loss function is shown in Equation (20):

$$\mathcal{L}_{\text{fusion}} = \alpha \mathcal{L}_{\text{intensity}} + \beta \mathcal{L}_{\text{detail}} \tag{20}$$

The pixel intensity loss $\mathcal{L}_{\text{intensity}}$ is used to constrain the image to contain more heat information, while the detail loss $\mathcal{L}_{\text{detail}}$ is used to constrain the image to contain more texture details. α and β are set as hyperparameters to balance pixel intensity loss and detail loss, in order to achieve better visual quality and higher evaluation indicators. The calculation methods for $\mathcal{L}_{\text{intensity}}$ and $\mathcal{L}_{\text{detail}}$ are as follows:

$$\mathcal{L}_{\text{intensity}} = \frac{1}{HW} \sum_{i}^{H} \sum_{j}^{W} \left| I_f^{(i,j)} - \max\left(I_{ir}^{(i,j)}, I_{\text{vis}}^{(i,j)} \right) \right| \tag{21}$$

$$\mathcal{L}_{\text{detail}} = \frac{1}{HW} \sum_{i}^{H} \sum_{j}^{W} \left| \nabla I_f^{(i,j)} - \max\left(\nabla I_{ir}^{(i,j)}, \nabla I_{\text{vis}}^{(i,j)} \right) \right| \tag{22}$$

where H denotes the image height, W represents the image width, max(·) signifies the operation of selecting the maximum element, and | · | represents absolute value. In the $\mathcal{L}_{\text{intensity}}$ loss function, the maximum selection strategy is employed to highlight significant target information in the infrared image, making it more prominent in the fused image. ∇ denotes the sobel gradient operator. In the $\mathcal{L}_{\text{detail}}$ loss function, the maximum gradient operation is performed to ensure that the most prominent texture detail information from the source image is retained in the fused image. In summary, under the joint constraint of pixel intensity loss and detail loss, the fusion result of the proposed method achieves optimal pixel distribution and the richest detail information. The fused image presents good visual effects and favorable objective evaluation indicators.

IV. EXPERIMENTAL VALIDATION

In this section, we will provide a detailed overview of the experimental configurations and implementation details of this study. On this basis, we present a qualitative and quantitative comparison of our method with state-of-the-art techniques to validate its superiority. Moreover, we conducted several ablation studies to help demonstrate the effectiveness and advancements of the specific designs within our method. Lastly, an efficiency analysis experiment and a research on object detection are conducted to further demonstrate the effectiveness of the proposed method.

A. EXPERIMENTAL CONFIGURATIONS

To comprehensively and accurately evaluate the performance of the fusion network, we conducted numerous qualitative and quantitative experiments on the OSU [46], TNO [47], RoadScene [25], and MSRS [48] datasets. In the testing phase, we compared the fusion results of our proposed model with those of current advanced fusion methods, including STDFusionNet [1], SeAfusion [6], DenseFuse [19], Nestfuse [20], IFCNN [24], U2Fusion [25], DDcGAN [27], SwinFusion [28], RFN-Nest [49], and SDNet [50]. To ensure fairness, these methods all used the parameter configurations recommended in the corresponding literature, and were fine-tuned for optimal performance before experimentation. Of note, in the replication of methods such as DenseFuse, NestFuse, IFCNN, RFN-Nest, which utilize various fusion strategies, we respectively adopted element-wise addition, average attention fusion strategy, element-wise maximum fusion strategy, and Residual fusion network for deep feature fusion.

Subjective visual perception systems can easily be influenced by individual emotions and visual environments, among other human factors. Therefore, it is unreliable to solely evaluate image fusion performance based on subjective qualitative visual effects. To more objectively and fairly evaluate the performance of the fusion network, we selected six commonly-used objective evaluation indicators from various perspectives, including: Mutual Information (MI) [51], Entropy (EN) [52], Visual Information Fidelity (VIF) [53], Standard Deviation (SD) [54], Spatial Frequency (SF) [55], and Average Gradient (AG) [56]. MI quantifies the amount of information the fused image obtains from the source image, while EN evaluates the amount of information contained in the fused image based on information theory. VIF primarily calculates the information fidelity in the fused image, indicating whether it aligns with human visual perception. SD reflects the image contrast based on static concepts. The larger the SD value, the better the contrast distribution in the image, and the more information the image carries. SF reflects the rate of change in image grayscale, while AG can measure the clarity of the fused image. The higher these indicators, the better the performance of the fusion network.

B. IMPLEMENTATION DETAILS

During the model training process, we used images from the OSU dataset to construct the training dataset. Due to differences in imaging sensors, the images in the OSU dataset are not strictly registered, leading to the presence of black edges in the infrared images. We therefore cropped the infrared and visible images to the same size of 280×200. This provided us with 8544 pairs of images to train the model. Since the visible images in the OSU dataset are colored and the infrared images are grayscale, fusing them as such would be meaningless. To make the channel count of the input image pairs identical, we preprocessed the visible images into grayscale. Additionally, all images were normalized to [0,1] before being fed into the network to accelerate model convergence. The batch size was set to 4. The fusion model was optimized using the Adam optimizer with parameters $\beta 1$ set to 0.9, $\beta 2$ set to 0.999, epsilon set to 10^{-8}, weight decay of 0.0002, and an initial learning rate of 0.001. The hyperparameters for loss were set at $\alpha = 1$, $\beta = 1$, and the fusion model was optimized under the guidance of the loss function $\mathcal{L}_{\text{fusion}}$. The entire training process was performed using the Pytorch framework. The RoadScene and TNO datasets used for testing contain colored visible images, but we trained the proposed network using input grayscale images. To achieve better visual results at the testing stage, we adopted the strategy [57] to process the colored images, instead of converting the input colored images into grayscale. Specifically, we first converted the colored images into YCbCr color space, then input the Y channels of the infrared and visible images into the model. Finally, the fusion results were channel-wise concatenated with the Cb and Cr channels from the visible image, then converted into an RGB color image, which is the final output result of the network. The process of the study handling the fusion of RGB images is shown in Figure 9.

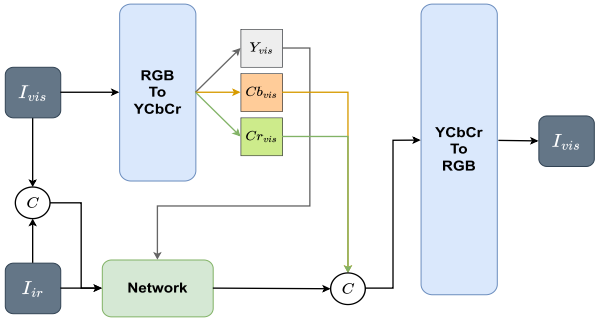

FIGURE 9. Process of processing RGB-images. © denotes the concatenate operation.

C. COMPARATIVE EXPERIMENT

To thoroughly assess the performance of the fusion network, we conducted comparative experiments with ten other methods on the TNO and RoadScene datasets. In this section, we will analyze and compare the performance of different fusion methods in both qualitative and quantitative experiments.

1) QUALITATIVE RESULTS

The qualitative comparison of different methods on the TNO dataset is shown in Figure 10. We selected six different scenes from TNO as representative examples, including: (1) bunker; (2) kaptein_1123; (3) kaptein_1654; (4) lake; (5) man in door; (6) marne etc., to demonstrate the superiority of the proposed method compared to others. Key targets in the fusion results are identified and enlarged by red boxes. From the comparison images in the bunker scene, it can be seen that DenseFuse, NestFuse, RFN-Nest, and U2Fusion did not effectively retain the heat information from the infrared images. Their fusion results have weaker

target information at the location indicated by the red box in the image. The DDcGAN method preserved the heat information well, but the texture details at the location indicated by the red box are blurry, indicating a lack of preservation of features from the visible images. In contrast, SeAFusion, IFCNN, and MDDCMA retained almost all the texture details from the visible images, and important heat targets were well preserved. Furthermore, it is easy to find that the fusion results of MDDCMA also have the clearest texture details. In the Kaptein_1123 image pair, RFN-Nest, DenseFuse, STDFusion, and NestFuse retain most of the texture information, but the fusion results do not exhibit the brightness factor present in the visible images, leading to a nearly uniform brightness in the grass. In the fusion results of IFCNN, U2Fusion, and SeAFusion, the grass's edge textures are blurred, resembling enhanced features of an infrared image. DDcGAN display prominent thermal information but suffer from significant noise and loss of detail. SwinFusion and MDDCMA almost completely preserve the grass background from the visible images, appearing more natural and in line with human visual perception. In Kaptein_1654, SeAFusion, SwinFusion, NestFuse, and IFCNN exhibit less clarity in the grassy background's edge textures. Moreover, U2Fusion, STDFusion, and RFN-Nest struggle to preserve salient targets effectively. MDDCMA nearly retain all texture details and salient information, with clear visibility of grass and street lamps, resulting in a more natural appearance. In the lake scene, the bush textures in IFCNN, U2Fusion, SeAFusion, NestFuse, and SwinFusion were blurry. DDcGAN's bushes had superior contrast and clear edge textures but miss information on the wooden bench and the letters in the top right corner. By contrast, MDDCMA's fusion results feature noticeable contrast and clearer edge textures, as indicated in the red boxes in the images.

Qualitative comparisons of different methods on the RoadScene dataset are shown in Figure 11, where key targets in the fusion results are highlighted with red boxes and magnified. The RoadScene dataset primarily includes daytime and nighttime road scenes with pedestrians and vehicles. Subjective evaluations are conducted on both daytime and nighttime images for intuitive comparative assessments. In the first column of Figure 11, it is evident that, except for SeAFusion, SwinFusion, and MDDCMA, other methods dimmed the street lamps to some extent, leading to information loss. Additionally, MDDCMA own the most distinct color contrast and the clearest edge textures, making objects in its fusion results more easily recognizable. In nighttime scenes, the information-providing capacity of infrared and visible images is limited, resulting in some redundant information, such as the glare of lights and blurred objects. As seen in the second column, all methods inevitably retain the glare of lights, thus reducing visual quality. Among these methods, DenseFuse and NestFuse's wires and cables nearly vanish against the black sky background, and the road signs in IFCNN, U2Fusion, SDNet, SeAFusion, and SwinFusion have unclear edge textures. MDDCMA's fusion results preserve rich texture details and salient information, with road signs being most prominent compared to other methods. Overall, this method effectively utilize the information from infrared and visible images to generate high-quality fusion images. MDDCMA show commendable fusion performance, though it still has its shortcomings. In the Kaptein_1654 scene in Figure 10, the fusion results of MDDCMA lose the information of the smoke, and in the marne scene, the results of MDDCMA results miss cloud features. This indicates that while MDDCMA own a robust feature extraction capability, it still falls short in distinguishing important information.

TABLE 2. Quantitative comparisons of the six metrics, I.E., AG, EN, MI, SD, SF, VIF, on 20 image pairs from the tno dataset (unit: red indicates the best result and blue represents the second best result).

Method	AG	EN	MI	SD	SF	VIF
DDcGAN	4.529733	7.361207	1.731497	9.685053	0.042514	0.689603
DenseFuse	3.338208	6.827795	2.725177	8.571426	0.033986	0.880978
IFCNN	4.144991	6.539053	2.300365	8.231863	0.042023	0.711111
NestFuse	3.533797	6.875468	2.952278	8.807706	0.035725	0.921433
RFN-Nest	2.792688	6.81824	1.99314	8.727294	0.0246	0.784724
SeAFusion	4.784406	7.016459	2.697099	8.897805	0.046737	0.928254
SwinFusion	3.811741	6.71552	3.053536	8.580392	0.038084	0.83382
U2Fusion	3.381549	6.275898	1.846464	8.139052	0.03142	0.641917
MDDCMA	7.113011	7.461745	2.886558	10.03327	0.072776	1.051869

2) QUANTITATIVE RESULTS

We conducted a quantitative assessment on 20 images from the TNO dataset and 221 images from the RoadScene dataset. We calculated the average scores of nine fusion methods across various objective evaluation metrics, as shown in Table 2 and Table 3. The best results are highlighted in red, and the second-best results in blue, providing an intuitive analysis of the evaluation. It is clear to see from Table 2 that MDDCMA achieves the best results in terms of AG and EN, signifying the strong ability of MDDCMA to retain information from source images. Additionally, the superior result on VIF indicates that the proposed method can generate high-quality images that align well with human visual systems. Although the proposed method is not the best in terms of MI and SD, multiple objective metrics indicate that the proposed method offers superior fusion performance, maintaining the feature information of the source images and presenting excellent visual quality, consistent with subjective evaluation results. Similarly, for the RoadScene dataset, MDDCMA also achieve better performance in AG, EN, SD, SF, and VIF metrics. As shown in the Table 3, the fusion results of MDDCMA are the best in terms of SD, EN, and AG, indicating that our fusion results have higher contrast and retain more edge details from the source images. Additionally, MDDCMA has the best SF scores, suggesting high image quality and clarity in our fusion results. The highest VIF further validates that our fusion results present excellent visual effects with low distortion. Considering the multiple objective evaluation metrics, MDDCMA has the best fusion performance, which is in line with the subjective evaluation and substantiates the effectiveness of our method.

FIGURE 10. The comparison results of different methods on the TNO dataset.

D. ABLATION STUDIES

We conducted ablation experiments to determine the optimal parameters for the loss function and to validate the effectiveness of the CMB and ECB in enhancing the fusion network. All ablation studies were carried out on the RoadScene dataset. To ensure a fair comparison, the network was trained using the same parameters and iterations, with the exception of the ablation parameters.

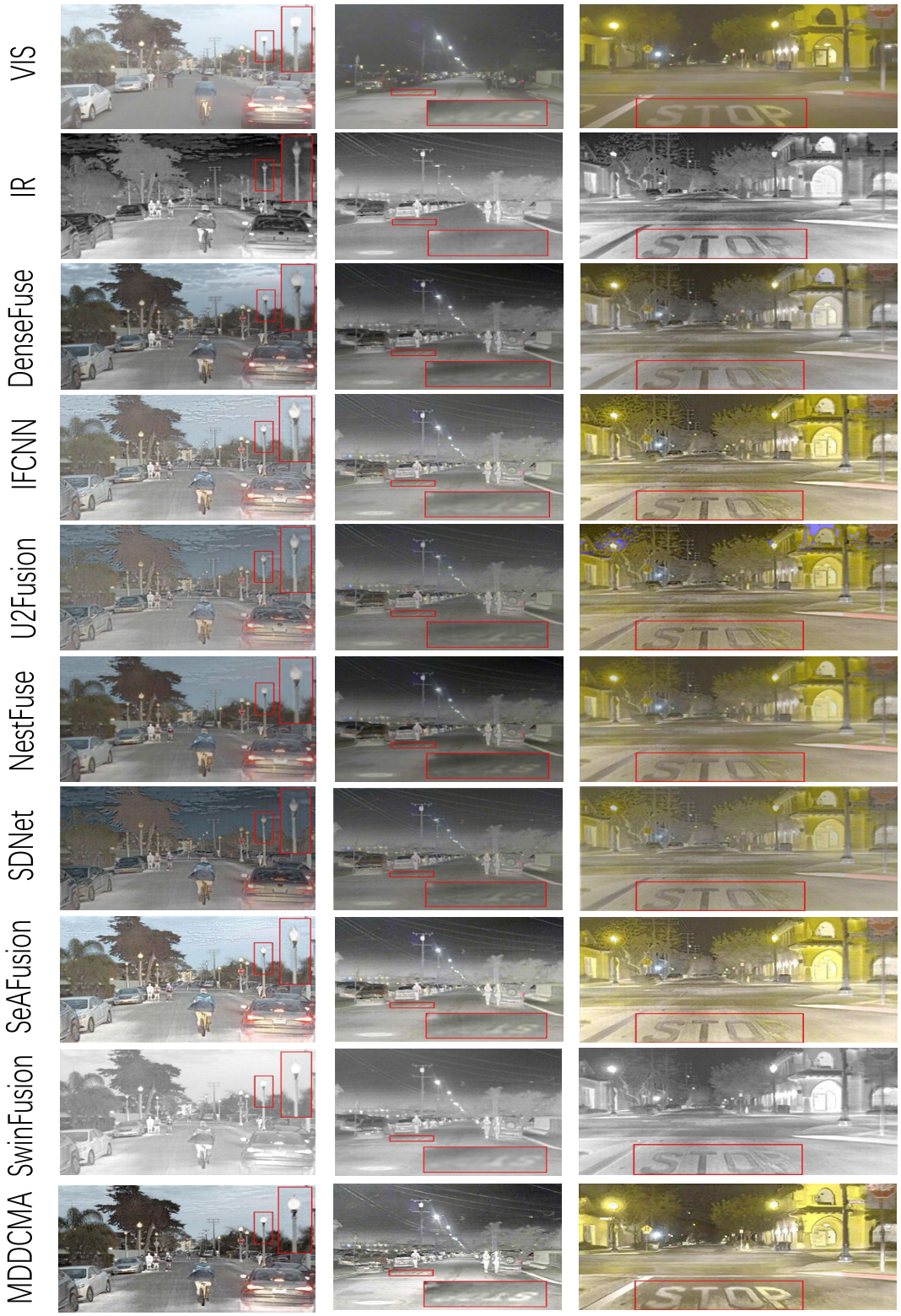

FIGURE 11. The comparison results of different methods on the RoadScene dataset.

1) ANALYSIS ON LOSS FUNCTION PARAMETERS

In the loss function proposed in this paper, which consists of $\mathcal{L}_{\text{intenity}}$ and $\mathcal{L}_{\text{detail}}$, we employed parameters α and β to control the weight of the two parts of the loss function. To this end, this research analyzed the impact of varying weights of pixel intensity loss and detail loss on the performance of the fusion network. The experimental results are presented in Table 4, which displays the calculated objective metrics

FIGURE 12. Representative examples of ablation results under different parameters. The first and second columns are infrared and visible images respectively, while the last three columns are subjective ablation results with hyperparameters $\alpha = 10$, $\beta = 1$, $\alpha = 1$, $\beta = 10$ and $\alpha = 1$, $\beta = 1$ respectively.

TABLE 3. Quantitative comparisons of the six metrics, I.E., AG, EN, MI, SD, SF, VIF, on 221 image pairs from the roadscene dataset (unit: red indicates the best result and blue represents the second best result).

Method	AG	EN	MI	SD	SF	VIF
DenseFuse	4.706425	7.279267	3.353331	10.31224	0.048973	0.800583
U2Fusion	5.036924	6.901157	2.662739	9.761545	0.048933	0.660084
IFCNN	6.016415	7.205511	3.005133	10.26107	0.061104	0.757717
NestFuse	4.25602	7.105744	3.492139	10.24845	0.045051	0.801904
SeAFusion	6.491336	7.329656	3.028792	10.51581	0.065195	0.813902
SDNet	4.85651	6.868144	3.016707	9.65337	0.046083	0.695712
STDFusionNet	5.613659	7.344679	4.110191	10.02578	0.061181	0.8196
SwinFusion	4.465027	6.921515	3.42225	10.02213	0.047407	0.780881
MDDCMA	8.823721	7.653395	3.260415	10.70291	0.085189	0.85391

TABLE 4. Quantitative average fusion performance of different group of loss parameter on the roadscene dataset (unit: red indicates the best result and blue represents the second best result).

Parameter	AG	EN	MI	SD	SF	VIF
α =10 β =1	8.881387	7.616307	2.614843	10.52508	0.094663	0.726951
α =1 β =10	7.615336	7.556276	3.131668	10.24579	0.084441	0.768192
α =1 β =1	8.823721	7.663513	3.260415	10.70291	0.085189	0.85391

of the fusion results under different parameters. The optimal results are indicated in bold, while the suboptimal results are underlined. It can be intuitively observed that among the six objective evaluation metrics, when the combination of α =1 and β =1 is chosen, several metrics of the fusion results, such as EN, MI, SD, and VIF, are optimal and exhibit excellent performance in visual quality. Therefore, compared with other loss parameters, this set of parameters possesses better fusion performance, and we adopted this set of parameters in subsequent experiments. Moreover, to facilitate the observation of the impact of loss parameters on fusion performance, Figure 12 provides a comparison of the images of the fusion results from different parameter groups. When the parameter combination is α=10, β=1, the weight of the intensity loss is larger, hence more intensity information features are retained. It can be observed that the edges of the utility poles and tree leaves are covered by strong light. When the parameter combination is α =1, β=10, the edge details are quite blurry, and the overall image becomes darker, with less preservation of pixel intensity information. When the parameter combination is α =1, β=1, the edge details of utility poles and tree leaves are clearer in the fused image, the image contrast is higher, and it possesses better visual quality. In summary, compared with the other two sets of parameters, the parameter combination α=1, β=1 has better fusion performance, which is the parameter used in subsequent experiments.

2) ANALYSIS ON COORDINATE MEAN ATTENTION FUSION BLOCK

In image fusion, an appropriate fusion strategy is crucial for enhancing fusion performance. A good fusion strategy can retain more information-rich image features during the fusion process. We designed a fusion module based on coordinate attention, termed as the CMB, to enhance the utilization of more meaningful features. To validate the effectiveness of CMB, we conducted ablation studies. Specifically, we set up two groups of comparative experiments. In the first group of experiments, we used concatenate operations to replace CMB to complete the fusion of features extracted from different branches. In the second group of experiments, we used element-wise operations including element-wise addition, element-wise maximum, and element-wise mean as the fusion strategy. In each group of experiments, we used the same images as input, followed by quantitative calculation of objective metrics for the fusion results. Table 5 presents the fusion results and the objective metric calculation results of the methods used in the first and second group of experiments, along with this study's method. The optimal results are indicated in red, and the suboptimal results are in blue. It can be seen that the fusion results of the network using CMB to execute fusion operations have the best effects on metrics such as EN, MI, SD, and VIF. This demonstrates the key influence of the fusion strategy on network performance and validates the effectiveness of the coordinate mean attention fusion block, which can widely and accurately guide feature fusion, yielding significant effects in improving fusion quality. Figure 13 provides a comparison of images of fusion results under different fusion strategies, with key areas marked and enlarged in red and green. Fusion results using element-wise addition, element-wise maximum, and concatenate operations as fusion strategies all contain a large amount of noise, and there is certain information missing in the areas marked in red, severely affecting visual quality. Fusion images using CMB and element-wise mean as fusion strategies retain most of the key features and allows clear identification of the street lamp while observing its texture details. However, in the fused image using the

FIGURE 13. Ablation results using different fusion strategies on the RoadScene dataset. The fusion strategy for different images is as follows: (a) Visible image (b) Infrared image (c) Element-wise addition (d) Element-wise max (e) Element-wise mean (f) concatenate (g) CMB.

TABLE 5. Quantitative average fusion performance of different fusion strategy on the roadscene dataset (unit: red indicates the best result and blue represents the second best result).

Method	AG	EN	MI	SD	SF	VIF
MDDCMA+ Elementwise add	12.74108	7.653395	2.623712	10.69085	0.13967	0.699545
MDDCMA+ Element-wise max	8.484373	7.592528	2.860704	10.18381	0.092838	0.692323
MDDCMA+ Element-wise mean	8.439486	7.62449	3.155478	10.60115	0.092787	0.822452
MDDCMA+ concatenate	9.247316	7.477409	2.596183	10.68175	0.100052	0.666444
Ours	8.823721	7.663513	3.260415	10.70291	0.085189	0.85391

element-wise mean as the fusion strategy, the features of power lines in the sky are lost, and the texture details of manhole covers on the road are unclear while the fused image using CMB as the fusion strategy allows for clear observation of the texture features of street lamps, power lines, and manhole covers, with no noticeable distortion or noise. Therefore, CMB can effectively enhance the quality of fusion results.

3) ANALYSIS ON EDGE CORRECTION BLOCK

Given that the features extracted by the feature extraction branches are coarse, we designed an edge correction block to supplement the extracted features in order to ensure the visual performance of the fusion results. The edge correction block can reduce the impact of noise on fusion results and enhance the visual quality of feature maps. To validate the role of ECB in improving the fusion results of the network, we tested the differences between models with and without edge correction block under the same input. Specifically, we used a visible image and an infrared image of a streetscape from the RoadScene dataset as input to obtain fusion results. Figure 14 presents the fusion results of the two models and their regional comparisons. From comparison images on the first row, it can be seen that due to the features in the red box area of the source image being relatively blurry, the letters in this area are difficult to observe and recognize. In such cases, the letters in the fusion image after ECB processing are clearer and contain less noise compared to those without ECB processing. This is because ECB effectively suppresses noise. In comparison images on the second row, as indicated by the red box, the edges of buildings without ECB processing are mostly covered by

FIGURE 14. Ablation results corresponding to different network structures. The first and second columns represent visible and infrared images respectively, while the third and fourth columns represent fusion results without ECB processing and fusion results using our method respectively.

TABLE 6. Quantitive average fusion performance of different network structures on the roadscene dataset (unit: red indicates the best result).

Method	AG	EN	MI	SD	SF	VIF
NO-ECB	9.199743	7.576102	2.739462	10.51581	0.094998	0.753928
Ours	8.823721	7.663513	3.260415	10.70291	0.085189	0.85391

halos, making them difficult to observe. In contrast, after ECB processing, the buildings exhibit clearer edge textures. In addition, we also calculated objective metrics to eliminate the influence of human subjective factors, as shown in Table 6. The results processed by ECB have higher EN, SCD, AG, SD, and SF values on the RoadScene dataset, which is consistent with the subjective evaluation results. Therefore, experimental results indicate that the model with ECB has better fusion effects compared to the network model that removed ECB, proving that edge correction block can improve the quality of fusion images.

E. EFFICIENCY COMPARISON

In order to evaluate the overall efficiency of different methods, we present the average runtime of various methods in Table 7. The runtime refers to the duration from the input of the image into the model to the generation of the final fused image. Notably, the DenseFuse method takes a longer time to produce fusion results. We attribute this to its use of a dense structure, which involves a significant amount of model parameters, and its implementation based on TensorFlow, which requires more time. IFCNN demonstrates the shortest runtime across all datasets, which we believe is due to its use of a pretrained residual network for feature extraction and a network structure with better generalization capabilities. DDcGAN, representing the GAN-based methods, also takes longer compared to CNN-based approaches. Our fusion model employs a four-layer feature extraction structure to comprehensively capture features from source images, thereby involving substantial computational load. Fortunately, in comparison to other methods, our approach still maintains competitive runtime efficiency.

TABLE 7. Mean of the running times of all methods on 500 image pairs with a size of 256 × 256 and 500 image pairs with a size of 512 × 512 (unit: red indicates the best result and blue represents the second best result).

Method	256 x 256	512 x 512
DenseFuse	2.5665 ± 0.1857	6.2335 ± 0.1488
DDcGAN	1.2166 ± 0.1138	2.4548 ± 0.1172
IFCNN	0.0109 ± 0.1241	0.0171 ± 0.1241
NestFuse	0.0659 ± 0.02033	0.2180 ± 0.2472
RFN-Nest	0.0493 ± 0.2395	0.0698 ± 0.1256
SDNet	0.0500 ± 0.0163	0.1472 ± 0.0194
STDFusionNet	0.0934 ± 0.0179	0.3730 ± 0.0232
SeAFusion	0.02944 ± 0.2576	0.0507 ± 0.2260
SwinFusion	0.3286 ± 0.1906	1.2569 ± 0.1391
U2Fusion	0.2656 ± 0.0132	1.1579 ± 0.0235
MDDCMA	0.1911 ± 0.1321	1.1251 ± 0.1288

F. OBJECT DETECTION APPLICATIONS

As previously mentioned, fused images encompass more useful information than any individual source image, substantially benefiting subsequent advanced visual tasks, such as object detection, tracking, and semantic segmentation. To validate this, we employed both infrared and visible source images, along with our fused images, in object detection tasks. Specifically, we utilized YOLOv8, one of the most sophisticated object detection networks currently available, pretrained on the COCO dataset, to demonstrate the enhanced detection performance of our fused images. The object detection experiments were conducted on the MSRS dataset, which depicts high quality urban scenes. We randomly chose 200 images from this dataset and manually annotated them, identifying people and vehicles as the primary subjects for detection. Visible images, infrared images, and the fused images of our method are put into the YOLOv8 detector, respectively.

1) QUALITATIVE ANALYSIS

We selected three representative scenes for a qualitative comparison of our experiment results, as depicted in the Figure 15. In the 00054N scene, the lighting is dim, rendering the model unable to detect pedestrian information in the visible image. Simultaneously, the thermal information emitted by the inactive car is too scant for detection in the infrared image. In the 01042N scene, the detector yields low confidence in identifying pedestrians in the dark areas of the visible image. Similarly, the minimal thermal emission from the car in the infrared image leads to poor detection results. In contrast, our fused images demonstrate superior detection performance. As can be seen in the last column of images, both pedestrians and cars are correctly detected with high confidence in the fusion results.

2) QUANTITATIVE ANALYSIS

To objectively assess the performance of our monitoring tasks, we conducted a detailed analysis using quantitative metrics, the results of which are presented in the Table 8. Precision, defined as the percentage of correctly predicted data, and recall, indicating the number of accurately predicted positive class instances, represent all positive classes in the sample. For both precision and recall, we set a threshold of 0.5. The mean average precision (mAP), which ranges from 0 to 1, evaluates model performance by combining accuracy and recall. mAP@0.5 denotes the mAP value at a confidence threshold of 0.5. The F1 score, a crucial metric in classification problems, balances recall and precision, considering them equally important. It is the harmonic mean of precision and recall, with possible values ranging from 0 to 1. As shown in the Table 8, the quantitative results reveal that our method, MDDCMA, ranks first in precision and F1 score for detecting people, and exhibits the best recall, F1, and mAP@0.5 for vehicle detection. When it comes to the average performance across all categories, MDDCMA consistently shows the best results, in line with subjective evaluation outcomes. These objective analyses lead us to conclude that our fusion results can enhance the performance of object detection tasks.

FIGURE 15. Results of object detection. (Top to bottom) 00689N, 01042N, and 00054N. (Left to right) visible images, infrared images and fused images of our method, respectively.

TABLE 8. Object evaluation metrics of 80 images in the msrs dataset for object detection (unit: red indicates the best result, blue indicates the second best result).

Category	Method	Precision	Recall	F1	mAP@0.5
Person	VI	0.8817	0.5190	0.65	0.6618
	IR	0.7895	0.8544	0.82	0.9111
	MDDCMA	0.8931	0.7405	0.84	0.9002
Cars	VI	0.9054	0.8954	0.89	0.9287
	IR	0.9583	0.6216	0.75	0.7509
	MDDCMA	0.9306	0.9054	0.91	0.9358
Average	VI	0.8935	0.7112	0.78	0.7953
	IR	0.8739	0.7380	0.78	0.8310
	MDDCMA	0.9118	0.8229	0.86	0.9080

V. CONCLUSION

To address the existing challenges in image fusion methods, such as manual design of fusion strategies and insufficient feature extraction, we propose a novel unsupervised deep learning model for the fusion of infrared and visible images. Specifically, we introduce a distributed image fusion network called MDDCMA. This network employs three branches to achieve image fusion, using MDDC in two feature extraction branches to capture rich image features from source images. Subsequently, CMB is utilized to precisely fuse the features from the feature extraction branches with those from the fusion branch. Then, the extracted features and the fusion results from the previous step are fed into the fusion branch to reduce the loss of target information in the infrared image and the loss of texture information in the visible image. Additionally, in order to minimize the impact of noise on the fused image, we design ECB to complement the extracted coarse features. The proposed method can most retain the salient target information in the infrared image and the textural details information in the visible image. Extensive quantitative and qualitative experiments conducted on multiple public datasets demonstrate that the proposed method exhibits performance comparable to the state-of-the-art image fusion methods. Although MDDCMA demonstrates good fusion performance, there is still room for improvement in its visual capabilities. We attribute this to the model's limitations in capturing complementary information within the source images. Once the fusion loss function is determined, the output results become fixed, leading to inherent limitations in the fusion results of fusion model. In the future, we plan to consider cascading additional downstream tasks to guide the training of the fusion model, further enhancing fusion performance.

REFERENCES

[1] J. Ma, L. Tang, M. Xu, H. Zhang, and G. Xiao, "STDFusionNet: An infrared and visible image fusion network based on salient target detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[2] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "DetFusion: A detection-driven infrared and visible image fusion network," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 4003–4011.

[3] Z. Tu, W. Pan, Y. Duan, J. Tang, and C. Li, "RGBT tracking via reliable feature configuration," *Sci. China Inf. Sci.*, vol. 65, no. 4, Mar. 2022, Art. no. 142101.

[4] G. Braso and L. Leal-Taixe, "Learning a neural solver for multiple object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6247–6257.

[5] L. Tang, Y. Deng, Y. Ma, J. Huang, and J. Ma, "SuperFusion: A versatile image registration and fusion network with semantic awareness," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 12, pp. 2121–2137, Dec. 2022.

[6] L. Tang, J. Yuan, and 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*, vol. 82, pp. 28–42, Jun. 2022.

[7] J. Nan, Z. Song, H. Lei, and W. Li, "Fusion of infrared and visible sensor images based on anisotropic diffusion and fast guided filter," *Proc. SPIE*, vol. 12342, pp. 830–837, May 2022.

[8] Z. Chao, X. Duan, S. Jia, X. Guo, H. Liu, and F. Jia, "Medical image fusion via discrete stationary wavelet transform and an enhanced radial basis function neural network," *Appl. Soft Comput.*, vol. 118, Mar. 2022, Art. no. 108542.

[9] J. Ma and Y. Zhou, "Infrared and visible image fusion via gradientlet filter," *Comput. Vis. Image Understand.*, vols. 197–198, Aug. 2020, Art. no. 103016.

[10] J. Chen, X. Li, L. Luo, X. Mei, and J. Ma, "Infrared and visible image fusion based on target-enhanced multiscale transform decomposition," *Inf. Sci.*, vol. 508, pp. 64–78, Jan. 2020.

[11] A. S. Yousif, Z. Omar, and U. U. Sheikh, "An improved approach for medical image fusion using sparse representation and Siamese convolutional neural network," *Biomed. Signal Process. Control*, vol. 72, Feb. 2022, Art. no. 103357.

[12] X. Li, H. Tan, F. Zhou, G. Wang, and X. Li, "Infrared and visible image fusion based on domain transform filtering and sparse representation," *Infr. Phys. Technol.*, vol. 131, Jun. 2023, Art. no. 104701.

[13] D. Tang, Q. Xiong, H. Yin, Z. Zhu, and Y. Li, "A novel sparse representation based fusion approach for multi-focus images," *Expert Syst. Appl.*, vol. 197, Jul. 2022, Art. no. 116737.

[14] J. Liu, D. Shen, Z. Wu, L. Xiao, J. Sun, and H. Yan, "Patch-aware deep hyperspectral and multispectral image fusion by unfolding subspace-based optimization model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1024–1038, 2022.

[15] C. Panigrahy, A. Seal, and N. K. Mahato, "Parameter adaptive unit-linking dual-channel PCNN based infrared and visible image fusion," *Neurocomputing*, vol. 514, pp. 21–38, Dec. 2022.

[16] P.-H. Dinh, "Combining spectral total variation with dynamic threshold neural p systems for medical image fusion," *Biomed. Signal Process. Control*, vol. 80, Feb. 2023, Art. no. 104343.

[17] M. A. Azam, K. B. Khan, S. Salahuddin, E. Rehman, S. A. Khan, M. A. Khan, S. Kadry, and A. H. Gandomi, "A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics," *Comput. Biol. Med.*, vol. 144, May 2022, Art. no. 105253.

[18] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, Dec. 2021.

[19] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.

[20] H. Li, X.-J. Wu, and T. Durrani, "NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9645–9656, Dec. 2020.

[21] Y. Liu, X. Chen, J. Cheng, and H. Peng, "A medical image fusion method based on convolutional neural networks," in *Proc. 20th Int. Conf. Inf. Fusion (Fusion)*, Jul. 2017, pp. 1–7.

[22] C. Lin, C. Qiu, H. Jiang, and L. Zou, "A deep neural network based on prior-driven and structural preserving for SAR image despeckling," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 6372–6392, 2023.

[23] Y. Wu, Y. Li, S. Feng, and M. Huang, "Pansharpening using unsupervised generative adversarial networks with recursive mixed-scale feature fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3742–3759, 2023.

[24] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, Feb. 2020.

[25] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.

[26] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.

[27] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.

[28] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via Swin transformer," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 7, pp. 1200–1217, Jul. 2022.

[29] W. Tang, F. He, and Y. Liu, "TCCFusion: An infrared and visible image fusion method based on transformer and cross correlation," *Pattern Recognit.*, vol. 137, May 2023, Art. no. 109295.

[30] D. Rao, T. Xu, and X.-J. Wu, "TGFuse: An infrared and visible image fusion approach based on transformer and generative adversarial network," *IEEE Trans. Image Process.*, early access, May 10, 2023, doi: 10.1109/TIP.2023.3273451.

[31] L. Dong and J. Wang, "FusionCPP: Cooperative fusion of infrared and visible light images based on PCNN and PID control systems," *Opt. Lasers Eng.*, vol. 172, Jan. 2024, Art. no. 107821.
[32] L. Dong and J. Wang, "FusionPID: A PID control system for the fusion of infrared and visible light images," *Measurement*, vol. 217, Aug. 2023, Art. no. 113015.
[33] M. E. Liggins, C.-Y. Chong, I. Kadar, M. G. Alford, V. Vannicola, and S. Thomopoulos, "Distributed fusion architectures and algorithms for target tracking," *Proc. IEEE*, vol. 85, no. 1, pp. 95–107, Jan. 1997.
[34] Z. Tang, G. Xiao, J. Guo, S. Wang, and J. Ma, "Dual-attention-based feature aggregation network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.
[35] Y. Wu, M. Huang, Y. Li, S. Feng, and D. Wu, "A distributed fusion framework of multispectral and panchromatic images based on residual network," *Remote Sens.*, vol. 13, no. 13, p. 2556, Jun. 2021.
[36] M. Brnnstrm, R. Lennartsson, A. Lauberts, H. Habberstad, E. Jungert, and M. Holmberg, "Distributed data fusion in a ground sensor network," in *Proc. 7th Int. Conf. Inf. Fusion*, Stockholm, Sweden, Jun. 2004, pp. 1–8.
[37] D. Yoo, S. Park, J. Y. Lee, A.S. Paek, and I. S. Kweon, "Attentionnet: Aggregating weak directions for accurate object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, Dec. 2015, pp. 2659–2667.
[38] Z.-Y. Liu and J.-W. Liu, "Hypergraph attentional convolutional neural network for salient object detection," *Vis. Comput.*, vol. 39, no. 7, pp. 2881–2907, May 2022.
[39] J. Fu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
[40] J. Li, H. Huo, C. Li, R. Wang, and Q. Feng, "AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks," *IEEE Trans. Multimedia*, vol. 23, pp. 1383–1396, 2021.
[41] M. Suganuma, X. Liu, and T. Okatani, "Attention-based adaptive selection of operations for image restoration in the presence of unknown combined distortions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 9031–9040.
[42] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "FFA-Net: Feature fusion attention network for single image dehazing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11908–11915.
[43] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[44] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[45] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13713–13722.
[46] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Comput. Vis. Image Understand.*, vol. 106, nos. 2–3, pp. 162–182, May 2007.
[47] A. Toet, "The TNO multiband image data collection," *Data Brief*, vol. 15, pp. 249–251, Dec. 2017.
[48] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware," *Inf. Fusion*, vols. 83–84, pp. 79–92, Jul. 2022.
[49] H. Li, X.-J. Wu, and J. Kittler, "RFN-nest: An end-to-end residual fusion network for infrared and visible images," *Inf. Fusion*, vol. 73, pp. 72–86, Sep. 2021.
[50] H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion," *Int. J. Comput. Vis.*, vol. 129, no. 10, pp. 2761–2785, Jul. 2021.
[51] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electron. Lett.*, vol. 38, no. 7, p. 313, 2002.
[52] J. Van Aardt, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *J. Appl. Remote Sens.*, vol. 2, no. 1, May 2008, Art. no. 023522.
[53] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Inf. Fusion*, vol. 14, no. 2, pp. 127–135, Apr. 2013.
[54] Y. J. Rao, "In-fibre Bragg grating sensors," *Meas. Sci. Technol.*, vol. 8, no. 4, p. 355, 1997.
[55] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.
[56] G. Cui, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition," *Opt. Commun.*, vol. 341, pp. 199–209, Apr. 2015.
[57] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4714–4722.

TONG TONG is currently pursuing the B.S. degree in the Internet of Things engineering with Guangdong Ocean University, Guangdong, China. His research interests include computer vision, image processing, and deep learning.

AIPING YE is currently pursuing the B.S. degree in software engineering with Guangdong Ocean University, Guangdong, China. Her research interests include neural networks and robotics.

YONGQI LU is currently pursuing the B.S. degree in automation with Guangdong Ocean University, Guangdong, China. Her research interests include computer vision and convolutional neural networks.

ZHENLU WU received the B.E. and M.E. degrees from Guangdong Ocean University, Guangdong, China, in 2006 and 2009, respectively. He is currently a Lecturer with the School of Mathematics and Computer Science, Guangdong Ocean University. His current research interests include machine learning and image processing.