Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.051556

ARTICLE

BDPartNet: Feature Decoupling and Reconstruction Fusion Network for Infrared and Visible Image

Xuejie Wang[1], Jianxun Zhang[1,*], Ye Tao[2], Xiaoli Yuan[1] and Yifan Guo[1]

[1]Department of Computer Science and Engineering, Chongqing University of Technology, Chongqing, 400054, China

[2]Liangjiang Institute of Artificial Intelligence, Chongqing University of Technology, Chongqing, 400054, China

*Corresponding Author: Jianxun Zhang. Email: zjx@cqut.edu.cn

Received: 08 March 2024   Accepted: 17 April 2024   Published: 20 June 2024

ABSTRACT

While single-modal visible light images or infrared images provide limited information, infrared light captures significant thermal radiation data, whereas visible light excels in presenting detailed texture information. Combining images obtained from both modalities allows for leveraging their respective strengths and mitigating individual limitations, resulting in high-quality images with enhanced contrast and rich texture details. Such capabilities hold promising applications in advanced visual tasks including target detection, instance segmentation, military surveillance, pedestrian detection, among others. This paper introduces a novel approach, a dual-branch decomposition fusion network based on AutoEncoder (AE), which decomposes multi-modal features into intensity and texture information for enhanced fusion. Local contrast enhancement module (CEM) and texture detail enhancement module (DEM) are devised to process the decomposed images, followed by image fusion through the decoder. The proposed loss function ensures effective retention of key information from the source images of both modalities. Extensive comparisons and generalization experiments demonstrate the superior performance of our network in preserving pixel intensity distribution and retaining texture details. From the qualitative results, we can see the advantages of fusion details and local contrast. In the quantitative experiments, entropy (EN), mutual information (MI), structural similarity (SSIM) and other results have improved and exceeded the SOTA (State of the Art) model as a whole.

KEYWORDS

Deep learning; feature enhancement; computer vision

1 Introduction

Image fusion is an important image processing task, aiming to merge images from different sensors or bands into a single composite image, which has important applications and demands in multiple fields [1–3]. For instance, it plays an important role in spacecraft relative navigation, among other diverse application areas [4]. In addition, in the field of remote sensing, image fusion is instrumental for tasks such as detection with small infrared targets [5], ship detection in remote sensing image [6–8].

Firstly, the demand for image fusion stems from the purpose of improving image quality and information content [9–11]. In many cases, a single image cannot provide enough information to meet the needs of specific applications. Visible light images can retain delicate gradient information and high spatial resolution, but they are easily affected by lighting and obstacles. However, infrared images complement the advantages and disadvantages of visible light images, obtaining a large amount of thermal radiation information, with lower texture and spatial resolution. The fusion of the two can improve the contrast, clarity, and details of the image, thereby improving the performance of target detection [12], tracking [13], pedestrian re-identification [14] and semantic segmentation [15], Earth observations and Spacecraft relative navigation [16].

Secondly, image fusion has important applications in night vision, thermal imaging, military, security monitoring, and other fields. IVIF (infrared and visible image fusion) tasks can help military and security departments achieve target detection [17] and navigation [18,19] under invisible light conditions, enhancing safety. In addition, fields such as agriculture [20], environmental monitoring, and resource management also need to fuse multi-modal images to obtain comprehensive information. With the continuous development of sensor technology, multi-modal data [21] from multiple sensors have become easier to obtain. Therefore, image fusion [22–24] has become a way to better integrate and use these data to achieve broader applications. It involves complex data processing and algorithm development, aiming to improve image quality and information richness.

The fusion of visible and infrared images can generally be divided into two categories: The first is traditional methods, while the second is depth based fusion. The traditional fusion method can be roughly divided into the following three steps. Initially, an exclusive transformation is applied to extract features from the original image, forming the feature extraction stage. Following this, diverse fusion strategies are utilized in the feature fusion stage to amalgamate these features. Then, the inverse transform is applied to the feature reconstruction stage, and the fused image is merged from the reconstructed features. These traditional methods use different mathematical methods, and we can classify them into five types based on different methods: Multi-scale transformation-based [25], sparse representation-based [26], saliency-based [27], subspace-based fusion methods [28], and hybrid methods.

With the development of technology, the field of deep learning has also made rapid progress, and there are better solutions and effects for infrared and visible light image fusion based on deep learning. In processing image fusion tasks, deep learning methods can be roughly divided into three categories: Methods based on traditional Convolutional Neural Network (CNN) convolution, methods based on Generative Adversarial Networks (GANs), and methods based on autoencoders. The traditional convolutional method handles such tasks by constructing a specific loss function and limiting a very complex network, ultimately achieving feature extraction of the image, followed by feature fusion, and finally feature reconstruction, achieving good fusion results. The use of GAN method involves applying the generator and discriminator to infrared and visible light image fusion tasks. Using generator and discriminator adversarial learning, the fusion image generated in unsupervised conditions is made to contain two types of modal image information. For IVIF tasks, current popular deep learning methods are primarily focused on improving the fusion process, but these methods overlook an important decomposition process and are only improving the follow-up tasks, which will reduce the efficiency. Paying attention to two features in the decomposition stage can better enhance the features. We believe that to better obtain the fused image, we should focus on the decomposition of features. As the two modalities in IVIF tasks each have important information that needs to be retained, if we can effectively decompose the prominent information from each modality, we can easily obtain a fused image containing a large amount of information. SDNet integrates the concepts of extrusion and decomposition into image fusion. It not only compresses the information from the source image to the fusion result but also considers the decomposition process from the fusion result to the source image. This method enhances the inclusion of scene details in the fused image, and produces more informative and visually attractive results. Inspired by SDNet, we also apply the decomposition idea to the fusion task, and combine the AutoEncoder to propose a groundbreaking network called BDPartNET to complete the multimodal image fusion task, allowing the generated images to retain as much infrared and visible highlight information as possible. For the BDPartNET network based on deep image decomposition, our main contributions are as follows:

- For the decomposition stage, we designed a decomposition loss function to be able to forcibly decompose the intensity and texture features of the image. The general model uses the pixel loss and structural loss of the source image and the fused image as the main loss. However, we consider that the key information of the source images of infrared and visible light images are intensity information and texture information, respectively. We just decomposed these two features, so our loss function mainly consists of two parts, which calculate the pixel loss of the infrared image with intensity information and texture feature image. When we reconstruct an image, our loss function restricts the reconstructed image to the intensity and gradient features of the source image. In the testing phase, we perform channel fusion and dimensionality reduction on the two feature maps decomposed from the network and finally generate a fused image through a decoder.
- For the decomposed intensity feature map, we designed a contrast enhancement module (CEM). We designed a multi-scale convolution that enhances contrast using standard deviation to process the intensity information branch. We calculate the standard deviation of each convolution area and perform standard deviation normalization on the input before the convolution operation. This enhances the contrast of the feature map and helps subsequent fusion.
- Inspired by the residual network, we designed a gradient enhancement module (DEM). We directly perform convolution operations on the texture feature map, calculate the Laplacian gradient map of the texture feature map, and then use residual connections in parallel, and finally use depth separable convolution to output the gradient-enhanced texture feature map, which greatly improves the effect in the fusion stage.

2 Related Work

This section will introduce the typical methods based on deep learning currently used for multimodal image fusion, and briefly describe the U-Net skip link used in BDPartNet. Furthermore, we will discuss the datasets involved in the training of the proposed algorithms to offer additional insights for the readers.

2.1 Image Fusion Based on Deep Learning

2.1.1 Fusion Methods Based on CNN

The difficulty of using convolutional neural networks to achieve infrared and visible light image fusion tasks is to design a sufficiently excellent loss function and a very complex network structure to achieve feature extraction, feature fusion, and reconstruction of the source image. In convolutional neural network-based methods, PMGI can serve as a representative work. The PMGI network designs a complex network and a loss function with fixed gradient and intensity ratios, enabling the network to extract features from multimodal source images and generate the final fused image. However, manually designing the proportion of information can lead to some serious issues, such as poor performance on other scene datasets.

This problem has been improved in subsequent research, such as SDNet [29]. This network can simultaneously focus on both fusion and decomposition stages so that the fusion result contains more scene details. STDFusionNet [30] uses salient target masks to assist in the early stages of the network, making it easier to integrate intensity information into the generated images. Taking advantage of the multi-scale feature integration in the fusion process, RXDNSort [31] adopts the structural advantages of ResNet [32] and DenseNet [33]. By comprehensively extracting features from each layer, effective fusion has been achieved. Li et al. [34] utilized the advantages of meta-learning to complete the fusion task of infrared and visible light images at different resolutions. They only used one CNN model, significantly expanding the applicability of the fusion model. The purpose of image fusion is to serve advanced visual tasks. Inspired by this, Tang and others added semantic loss to the IVIF task, which is more in line with advanced visual tasks such as object segmentation and detection. The recently introduced Transformer structure [35,36] similar to CNN performs well in natural language processing and also performs very well in various visual tasks. So, Ma et al. [37] designed a universal fusion method based on the Transformer architecture, where attention guided cross domain modules can integrate global complementary information. However, lighting conditions also play a decisive role in the quality of fusion images in IVIF tasks. PIAFusion [38] improved the model based on the influence of lighting conditions during fusion, but the compatibility of the model is insufficient, and the effect is still not ideal in relatively complex environments.

2.1.2 Image Fusion Based on GAN

Generative Adversarial network (GAN) [39] is a very excellent generative network, which was proposed by Ian goodflow and others in 2014. The core idea of Gan is to learn the generative model of data distribution through confrontation. The purpose of the generator is to generate as realistic data as possible, while the discriminator can distinguish false data as much as possible, forcing the generator to achieve better results. Through the training process of this game, the generator will generate more and more realistic data, to improve the performance of the generated model. Ma et al. proposed Fusiongan [40] network for IVIF tasks based on GAN network for the first time. This method defined the fusion algorithm as a confrontation game between maintaining infrared thermal radiation information and maintaining visible appearance texture information. The generator attempts to generate an image that combines the main infrared intensity and additional visible gradients, and the goal of the discriminator is to force the fused image to have more texture details. This enables the fused image to maintain the thermal radiation of the infrared image and the texture details of the visible image at the same time.

2.1.3 Image Fusion Based on Auto-Encoder

As research progresses, scholars have proposed numerous image fusion methods based on autoencoders. These methods mainly use autoencoders to extract features from source images and complete image reconstruction. During the feature fusion stage, manually designed fusion rules are usually applied. One typical fusion method based on AutoEncoder (AE) is DenseFuse [41], which consists of convolutional layers, fusion layers, and dense blocks. However, considering the limited capability of the autoencoder structure in feature extraction, Li and others proposed NestFuse [42] and RFN-Nest [43]. NestFuse introduces nested connections in the network to extract multi-scale features from source images, while RFN-Nest designs a detail preservation loss function and feature enhancement loss function to encourage the network to achieve a higher degree of detail feature integration.

As multi-scale features have not completely eliminated redundant information in source images, Jian et al. [44] adopted an attention mechanism, focusing on the salient targets and texture details of source images. To further explore the interpretability of the fusion field, DRF [45] decomposes the source image into scene components and attribute components and fuses them separately. Although DRF considers the interpretability of feature extraction, it ignores the interpretability of fusion rules, that is, manually formulated fusion rules may not be suitable for merging deep features. Therefore, Xu et al. [46] proposed a learnable fusion rule, which improves the interpretability of the network by evaluating the importance of each pixel to the classification result.

2.2 U-Net and Skip Connection

U-Net [47,48] is a structure of Convolutional Neural Networks (CNN), initially proposed by Ronneberger and others in 2015, specifically designed for image segmentation tasks. What makes it unique is that the network presents a U-shaped structure. Although it was initially used for medical image segmentation, it was later widely applied in the field of computer vision, covering various image segmentation problems.

U-Net introduced the concept of skip connections, connecting the feature maps of the encoder with those of the corresponding decoder layer. This connection mechanism helps to fuse global and local information at different scales, thereby enhancing the model's perception of details such as target boundaries. The design of this structure aims to optimize image segmentation tasks, enabling the network to more effectively capture and understand complex structures in images.

It is worth noting that in addition to image segmentation tasks, U-Net is also used as a feature extractor in various applications. Samkari et al. [49] also used U-Net in the field of human pose estimation.

2.3 Datasets Introduction

The experiments in this paper utilize three main datasets: FLIR, TNO [50], MSRS [38], and Roadscene [51]. The FLIR dataset (available at www.flir.com/oem/adas/adas-dataset-form/) was released in July 2018, 7,498 total video frames recorded at 24 Hz and are 1:1 match between thermal and visible frames. This dataset assists developers in training convolutional neural networks, enabling the automotive industry to utilize FLIR's cost-effective thermal imagers and develop safer and more efficient autonomous driving systems. The TNO dataset comprises near-infrared, long-wave infrared, or thermal infrared nighttime images along with visible light images captured in military and various other scenes. These images are suitable for research on image fusion algorithms in complex scenarios. The MSRS dataset consists of high-quality aligned infrared and visible light image pairs. These image pairs include diverse scenes such as roads, vehicles, and pedestrians. The Roadscene dataset contains 221 aligned pairs of visible light and infrared images, showcasing a variety of scenes including roads, vehicles, and pedestrians.

These datasets were selected to provide diverse scenes and ample data support for our image fusion algorithm research.

3 Method

In this section, we will introduce the BDPartNET algorithm and the proposed network structure. In Section 3.2, we will provide a detailed description of our intensity enhancement module and texture enhancement module. In addition, we will also explain the details of the training and testing phases.

3.1 Problem Formulation

When visible light images suffer from illumination degradation during nighttime, the problem of fusing nighttime infrared and visible light images can be divided into two sub-problems: Feature enhancement for visible light and infrared images, and fusion strategy. However, existing approaches that simply combine enhancement and fusion algorithms suffer from significant texture loss and insufficient contrast. Therefore, bridging the gap between enhancement and fusion tasks and jointly modeling them becomes crucial for infrared and visible light image fusion. Addressing these challenges, we propose BDPartNet, which extracts prominent features from both modalities before performing the fusion task.

Thus, through feature enhancement and our designed encoder, the intensity information of the infrared image and the texture information of the visible image can be better preserved. During training, only individual images are used as input and output, without involving multi-modal fusion. Visible light and infrared images are inputted into the shared encoder and decoder during training to better train the model's ability to decompose images. Through our carefully designed loss function, it is straightforward to decompose the intensity and texture features of an image, which conveniently aligns with the emphasis on these two features in infrared and visible light, respectively.

In the fusion stage, as depicted in Fig. 1, registered infrared and visible light images are separately fed into the encoder for decomposition. At this point, utilizing the trained network, we obtain the intensity feature map of the infrared image and the texture detail feature map of the visible light image. Knowing that infrared images excel in intensity information while visible light images excel in texture information, we perform a channel-wise fusion of these advantageously modal-specific feature maps, followed by dimensionality reduction and input to the decoder for image reconstruction, thus yielding a fused image with rich information.

During the training phase, our objective is to train a model with a well-designed loss function capable of effectively decomposing the features of a source image into two parts. Subsequently, these decomposed features are fused, aiming to maintain as much similarity as possible to the original image before decomposition. By training on datasets of both modalities, we aim to develop a robust shared encoder and decoder for subsequent fusion tasks. The training process for decomposition can be represented as follows.

3.2 BDPartNet Architecture

Our neural network comprises an encoder, a decoder, and feature enhancement modules. As illustrated in Fig. 1, the encoder takes input from either infrared or visible light images. To facilitate subsequent feature enhancement, we employ a dual-branch convolutional approach to generate two distinct feature maps, which we term as intensity feature map and the detail feature map. To better process these two types of features, we design two specific feature enhancement modules, namely the contrast enhancement module and texture enhancement module, which are applied to each feature map, respectively. Subsequently, the enhanced detail feature map and intensity feature map are simply concatenated for fusion. Given that our model follows an end-to-end architecture, a decoder is designed to decode the fused feature map, ultimately restoring the original image. To prevent loss of detail and expedite network convergence, the results of the first and second convolutional layers in the encoder are respectively added to the inputs of the penultimate and antepenultimate layers in the decoder.

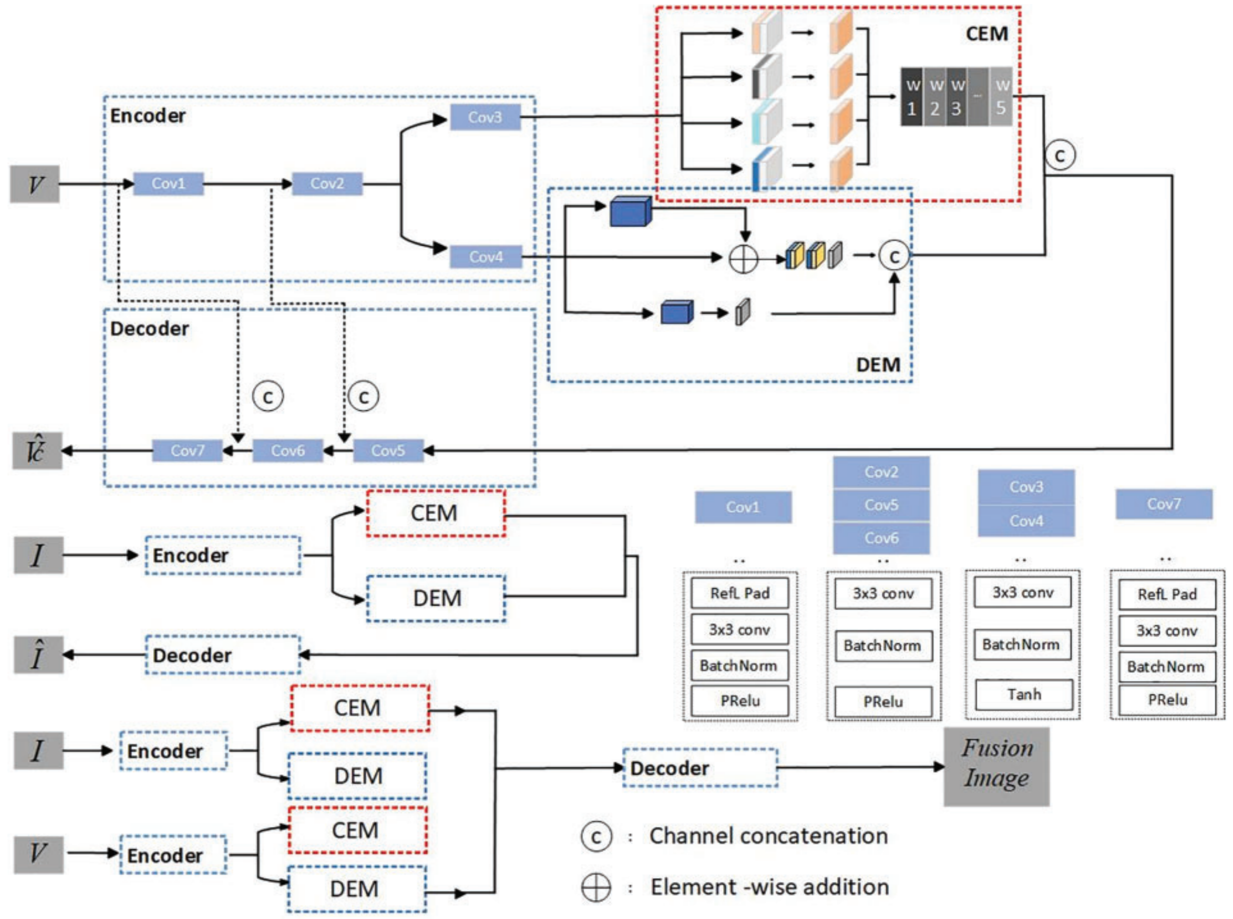

Figure 1: The overall framework of the proposed BDPartNet

Thus, through feature enhancement and our designed encoder, the intensity information of the infrared image and the texture information of the visible image can be better preserved. During training, only individual images are used as input and output, without involving multi-modal fusion. Visible light and infrared images are inputted into the shared encoder and decoder during training to better train the model's ability to decompose images. Through our carefully designed loss function, it is straightforward to decompose the intensity and texture features of an image, which conveniently aligns with the emphasis on these two features in infrared and visible light, respectively.

In the fusion stage, as depicted in Fig. 1, registered infrared and visible light images are separately fed into the encoder for decomposition. At this point, utilizing the trained network, we obtain the intensity feature map of the infrared image and the texture detail feature map of the visible light image. Knowing that infrared images excel in intensity information while visible light images excel in texture information, we perform a channel-wise fusion of these advantageously modal-specific feature maps, followed by dimensionality reduction and input to the decoder for image reconstruction, thus yielding a fused image with rich information.

During the training phase, our objective is to train a model with a well-designed loss function capable of effectively decomposing the features of a source image into two parts. Subsequently, these decomposed features are fused, aiming to maintain as much similarity as possible to the original image before decomposition. By training on datasets of both modalities, we aim to develop a robust shared encoder and decoder for subsequent fusion tasks. The training process for decomposition can be represented as follows:

$$\phi = E(\phi_D, \phi_C) \tag{1}$$

In the equations below, ϕ represents the feature map of the input image. If the image is a color image, we only utilize the Y channel from the Ycrcb color space. E denotes the dual-branch feature enhancement block, while ϕ_D and ϕ_C represent the separated and enhanced features, respectively, used for subsequent fusion. The fusion process can be represented as follows:

$$\phi_F = F(\phi_C^{ir}, \phi_D^{vis}) \tag{2}$$

where ϕ_F represents the fused image, F denotes the process of channel concatenation followed by input to the decoder for fusion. ϕ_C^{ir} represents the intensity feature map of the infrared image after decomposition and enhancement, while ϕ_D^{vis} represents the texture feature map of the visible light image after decomposition and enhancement.

3.3 Detail Enhancement Module (DEM)

After encoding in the encoder, two features were forcibly decomposed, with one part defined as the texture module, as we designed a detail enhancement module (DEM) as shown in Fig. 2. Inspired by ResBlock [52–54], DEM consists of a main convolutional stream and two parallel residual streams. To preserve the most fundamental features, the mainstream passes through two 3 × 3 convolutions and one 1 × 1 convolution, where the 3 × 3 convolutions are with the LReLU activation function. The two residual streams are positioned before and after, respectively. The first residual stream directly calculates feature maps using the Laplacian operator, then concatenates them directly before the mainstream convolutions, enhancing the input gradient of the mainstream. The second residual stream preserves texture features through the Sobel operator, connecting a 1 × 1 convolution to match channel differences, and finally performs Element-wise addition after the main stream's convolution operation.

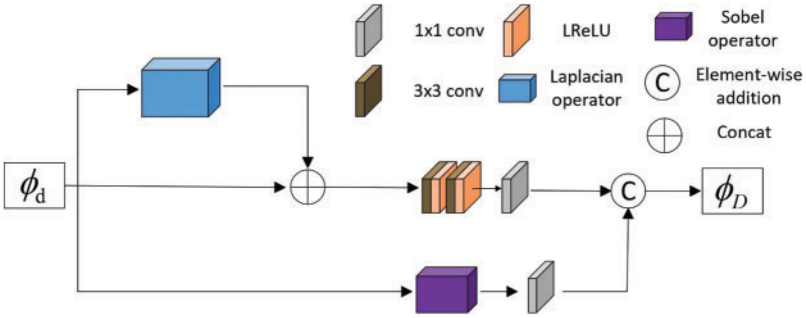

Figure 2: The detailed structure of detail enhancement module

3.4 Contrast Enhancement Module (CEM)

As shown in Fig. 3, the decomposed intensity features are input into the Contrast Enhancement Module (CEM), allowing feature maps of different scales to be input into the network. This enhances the details of objects of different sizes. Typically, convolution kernels of the same size struggle to capture information at different scales. Therefore, four convolution kernels of different sizes are used to capture multi-scale depth features. To minimize the loss of detail information during the process, no pooling layers are added in the multi-scale convolution operation. The kernel sizes of the multi-scale convolution layer are 1 × 1, 3 × 3, 5 × 5, and 7 × 7, respectively. The convolution layer chooses the LReLU activation function. Then, the obtained features are concatenated along the channel dimension and sent to the standard deviation calculation layer. We designed a convolution layer to calculate the standard deviation of each small area of the feature map. Unlike ordinary convolution, the parameters in the standard deviation convolution kernel are determined by the standard deviation of the feature maps of different sizes and do not undergo backpropagation during the network training process. The calculation formula is computed as

$$\sigma_{ij} = \frac{1}{2r+1}\sqrt{\sum\nolimits_{-r\le p,q\le r}\left(\phi_b(i+p,j+q)-\mu_{ij}\right)^2} \tag{3}$$

where r is the radius of the window. μ_{ij} is the mean value of the window with radius r centered on (i, j), and σ_{ij} is the standard deviation. ϕ_b represents the eigenvalue on the corresponding coordinate. Then we normalize the obtained standard deviation parameters and convolve the feature map again. We incorporate the obtained feature maps into the attention mechanism so that feature maps of different channels have corresponding weights. Finally, the contrast-enhanced feature map and the texture-enhanced feature map are concatenated and sent into the decoder. We designed a contrast block in the CEM to enhance the local contrast information at the feature level.

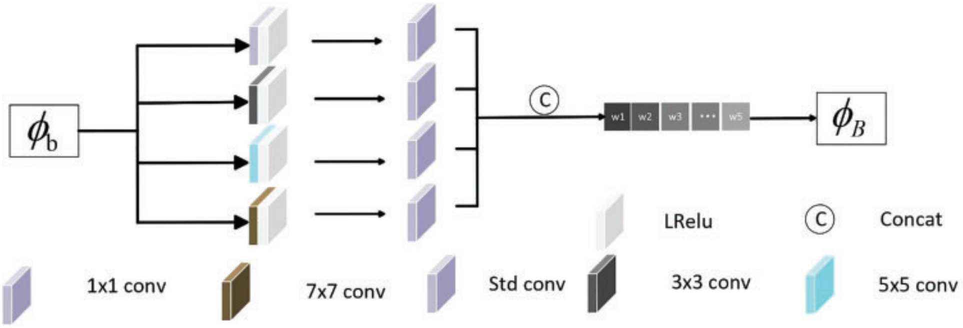

Figure 3: The detailed structure of contrast enhancement module

3.5 Encoder and Decoder

For feature extraction and image restoration, we have designed an encoder and decoder.

The encoder is responsible for feature extraction, where the Cov1 convolutional block extracts the raw features from the source image. We first perform reflection padding on the input image, followed by a 3 × 3 convolution for feature extraction. A BatchNorm layer is added for accelerated convergence and prevention of overfitting. The Cov1 convolutional block utilizes the PReLU activation function. The features are then passed through Cov2 for convolution and subsequently undergo dual-branch convolution to forcibly decompose them into intensity and texture features for enhancement.

The decoder concatenates the enhanced intensity and texture features and performs three deconvolutions for image restoration. Cov2, Cov5, and Cov6 share the same structure. The first two convolutional blocks in the decoder are intended to increase the depth of the network, thereby enhancing its robustness. Finally, the features are passed through Cov7 for deconvolution and dimensionality reduction to restore the source image.

Inspired by U-Net skip connections, we have integrated short-circuit connections into our network design. As depicted in Fig. 1, it can be observed that there are two skip connections between the encoder and decoder. The skip connection between the sixth and seventh convolutional blocks is aimed at minimizing the loss of texture detail between the source image and the restored image during training. Similarly, the skip connection between the fifth and sixth convolutional blocks aims to minimize the loss of intensity information.

3.6 Loss Function

Image decomposition The Encoder forcibly decomposes the feature maps extracted from the source images into two types of features: One is the intensity feature used to extract pixel intensity information of the two images, and the other is detail feature map used to extract gradient information of the two images. The $L_1 - Loss$ which is the loss function for image decomposition is as follows:

$$L_1 = \Phi\left(\|G_V - G_I\|_2^2\right) + \beta_1\left(\|T_V - T_I)\|_2^2\right) \tag{4}$$

G_V and T_V represent the intensity and detail feature map of visible light image V, while G_I and T_I represent the background and detail feature map of infrared image I. Φ is the *tanh* function, which is used to restrict the gap to the interval $(-1, 1)$. β_1 is the tuning parameter.

Image reconstruction During the image fusion stage, also known as the reconstruction stage, to better preserve the intensity and texture information of the source images, we designed a well-crafted reconstruction loss function $L_2 - Loss$ given by

$$L_2 = \beta_2(f(I,\hat{I}) + \beta_3(f(V,\hat{V}) + \beta_4||\nabla V - \nabla\hat{V}||_1 + \beta_5||\nabla I - \nabla\hat{I}||_1 \tag{5}$$

The input infrared image and visible light image are denoted by I and V, respectively, and the reconstructed images are denoted by $\hat{I}$ and $\hat{V}$. ∇ denotes the gradient operator, and

$$f\left(P,\hat{P}\right) = \|P - \hat{P}\|_2^2 + \lambda L_{SSIM}\left(P,\hat{P}\right) \tag{6}$$

where P represent the input image and $\hat{P}$ represent the reconstructed image, and λ is the hyperparameter. SSIM [55] is used in the formula to measure the structural similarity of two images. L_{SSIM} can be described as

$$L_{SSIM}\left(P,\hat{P}\right) = \frac{1 - SSIM\left(P,\hat{P}\right)}{2} \tag{7}$$

It is important to note that the $L_2 - norm$ measures the consistency of pixel intensity between the original and reconstructed images, while L_{SSIM} calculates differences in brightness, contrast, and structure of the images. Particularly, visible images are rich in texture. In order to ensure texture consistency, we use gradient sparse penalty to regularize and reconstruct the visible image.

Combining L_1 and L_2 , the total loss L_{Total} can be expressed as

$$\begin{aligned} L_{Total} &= L_1 + L_2 \\ &= \Phi(\|G_V - G_I\|_2^2) + \beta_1(\|T_V - T_I)\|_2^2) + \beta_2(f(I,\hat{I}) \\ &\quad + \beta_3(f(V,\hat{V}) + \beta_4||\nabla V - \nabla V\hat{V}||_1 + \beta_5||\nabla I - \nabla I\hat{I}||_1 \end{aligned}$$

where β_1, β_2, β_3, β_4, β_5 are the tuning parameters.

4 Experimental Results

In this section, we demonstrate the generalization and feasibility of our method through experimental processes and results. Firstly, we describe the experimental setup. Then, we compare the qualitative and quantitative results of our method with the experimental results of other networks to demonstrate the superiority of our approach. Finally, we validate the effectiveness of certain modules in our network through a series of ablation experiments.

4.1 Experimental Configurations

We trained our network using the FLIR dataset and tested it on three different datasets: TNO, MSRS, and RoadScene. The TNO dataset comprises multispectral nighttime images of various military-related scenarios. Since our network performs fusion on single-channel grayscale images, for RGB images in the dataset, we experimented by decomposing the Y channel from the Ycrcb color space to obtain grayscale images.

For quality assessment of image fusion, six commonly adopted metrics in academia are typically used, including entropy (EN), standard deviation (SD), mutual information (MI), visual information fidelity (VIF), spatial frequency (SF), Qabf, and structural similarity index measure (SSIM). EN measures the amount of information contained in the fused image, and MI evaluates the amount of information transmitted from the source image to the fused image. Both EN and MI evaluate fusion performance from the perspective of information theory. VIF evaluates the information fidelity of the fused image from the perspective of the human visual system. SF measures the spatial frequency information contained in the fused image. SD reflects the distribution and contrast of the fused image from a statistical perspective. Qabf evaluates the amount of edge information transmitted from the source image to the fused image. EN, SF, and SD are no-reference indicators. In addition, the larger the EN, MI, VIF, SF, SD, Qabf, and SSIM of the fusion algorithm, the better the fusion performance.

4.2 Implementation Details

We randomly select 180 images from the FLIR dataset as training samples. Before training, all images are converted to grayscale. Meanwhile, all input images are cropped to a size of 128 × 128 pixels before being fed into the encoder for training.

4.3 Hyperparameters Setting

Through experience summary, the adjustment parameters used in BDPartNet are set as follows: $\beta_1 = 0.5$, $\beta_2 = 2$, $\beta_3 = 3$, $\beta_4 = 7$, $\beta_5 = 3$ and $\lambda = 5$. During the training phase, Adam optimized the network for 180 batches with a size of 24. We set the initial learning rate to 10^{-2}, and reduce it by a factor of 10 every 40 epochs. Fig. 4 shows the relationship between the loss curve and the epoch index. The results show that after 180 epochs, no matter the total loss or other characteristic loss, the decline curve is very flat in the learning process, so the network matches our parameter settings very well.

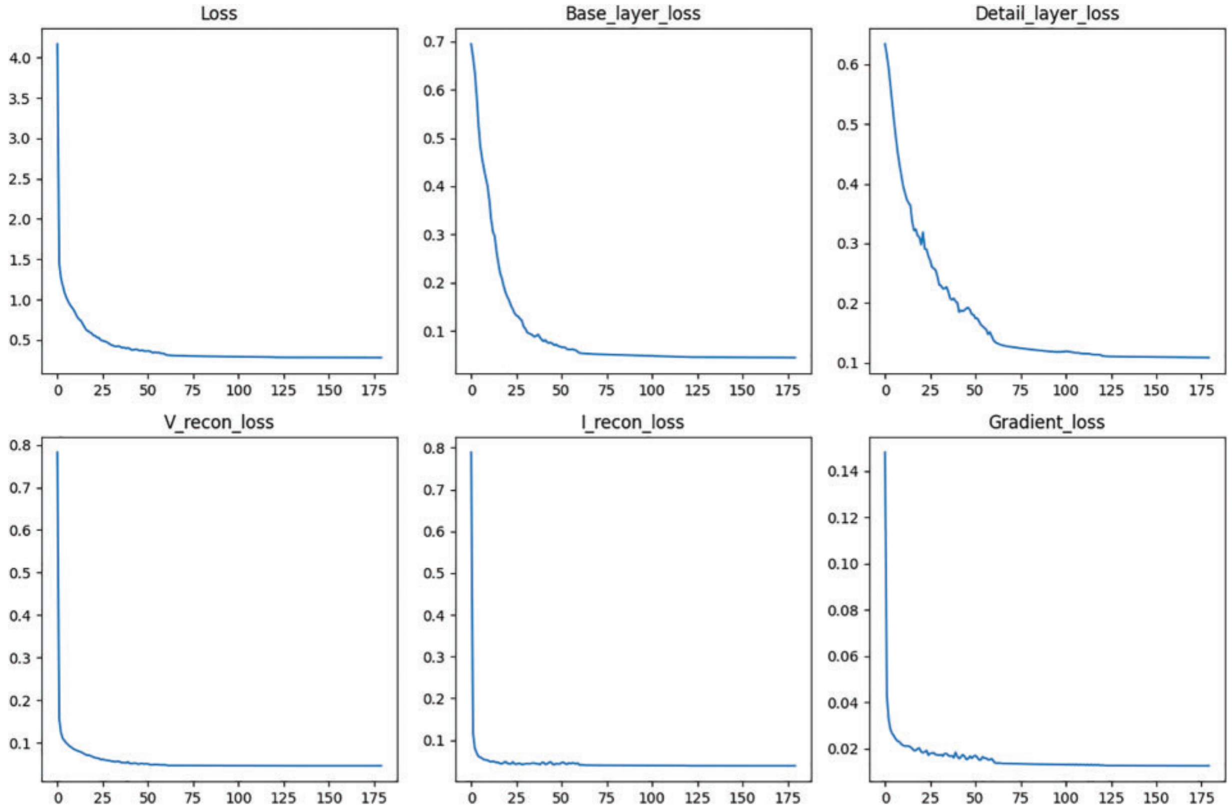

Figure 4: Loss curves over 180 epochs

4.4 *Experimental Results*

In this section, we test BDPartNET on three test sets and compare our method with nine state-of-the-art methods, including DIDFuse [56], U2Fusion [1], SDNet, RFNet [57], TarDAL [58], DeFusion [59], and ReCoNet [60]. All these seven methods are publicly implemented, and we set parameters according to the reports in the original papers.

Qualitative results. We show the qualitative comparison in Fig. 5 Apparently, our method better integrates the thermal radiation information in the infrared image and the detailed texture in the visible image. Objects in dark areas are clearly highlighted, making it easy to distinguish foreground targets from the background. In addition, background details that are difficult to recognize due to low illumination have clear edges and rich contour information, which helps us better understand the scene.

Quantitative results. Then, we use eight metrics to quantitatively compare the aforementioned results, as shown in Tables 1–3. Our method exhibits excellent performance on nearly all metrics, demonstrating that our approach is applicable to a variety of lighting conditions and target categories.

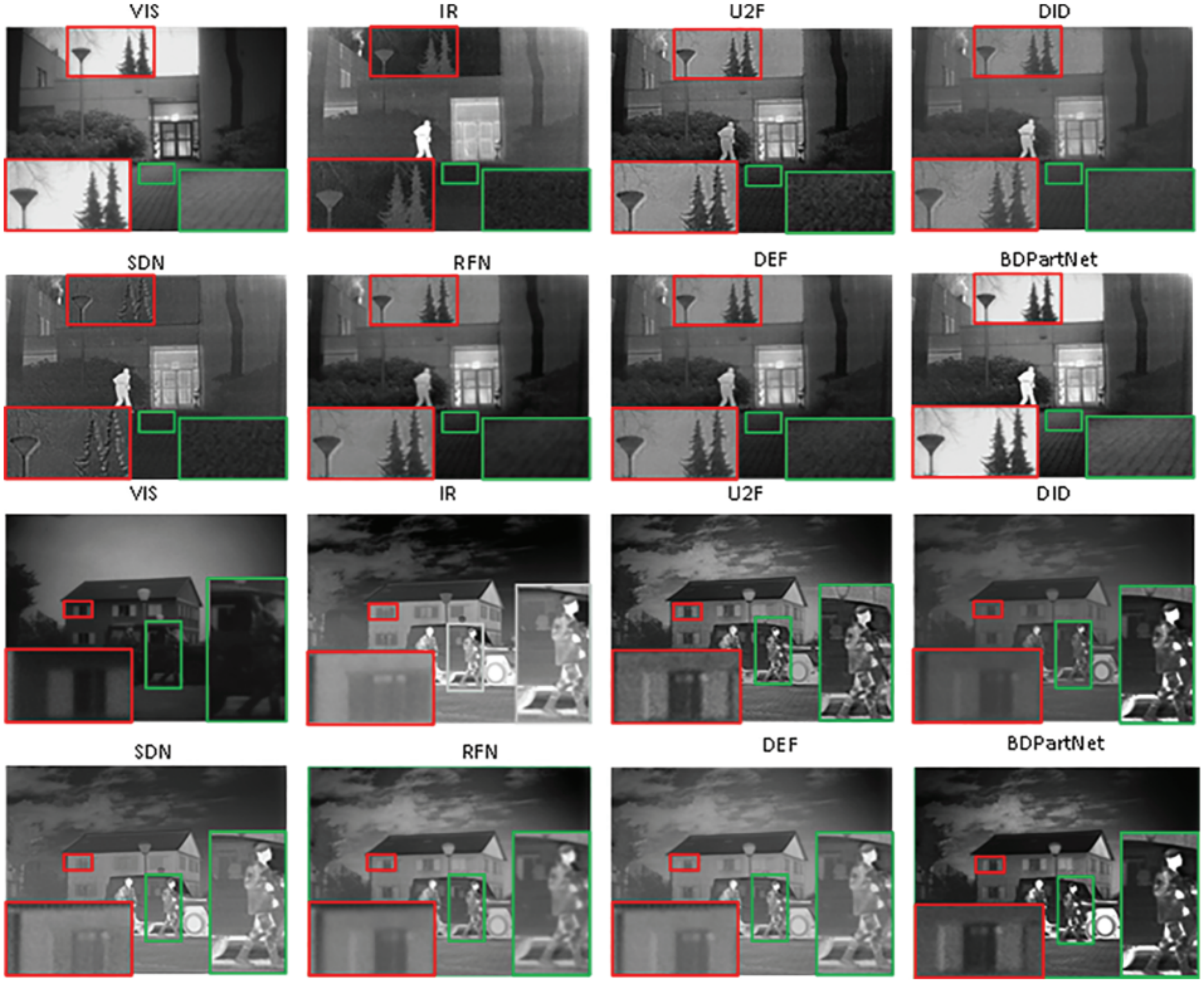

Figure 5: Comparison of qualitative results with SOTA fusion methods. Areas marked by red and green boxes are amplified for ease of inspection

Table 1: Quantitative experimental results based on the TNO dataset. Bold indicates the best value [61]

	Dataset: TNO infrared-visible fusion dataset							
Method	EN	SD	SF	MI	SCD	VIF	Qbaf	SSIM
DID	6.97	45.12	12.59	1.70	1.71	0.60	0.40	0.81
U2F	6.83	34.55	11.52	1.37	1.71	0.58	0.44	0.99
SDN	6.64	32.66	12.05	1.52	1.49	0.56	0.44	1.00
RFN	6.83	34.50	**15.71**	1.20	1.67	0.51	0.39	0.92
TarD	6.84	45.63	8.68	1.86	1.52	0.53	0.32	0.88
DeF	6.95	38.41	8.21	1.78	1.64	0.60	0.41	0.96
ReC	7.10	44.85	8.73	1.78	1.70	0.57	0.39	0.88
BDPartNet	**7.24**	**46.69**	13.19	**2.17**	**1.84**	**0.63**	**0.55**	**1.02**

Table 2: Quantitative experimental results based on the MSRS dataset. Bold indicates the best value

	Dataset: MSRS infrared-visible fusion dataset							
Method	EN	SD	SF	MI	SCD	VIF	Qbaf	SSIM
DID	4.27	31.49	10.15	1.61	1.11	0.31	0.20	0.24
U2F	5.37	25.52	9.07	1.40	1.24	0.54	0.42	0.77
SDN	5.25	17.35	8.67	1.19	0.99	0.50	0.38	0.72
RFN	5.56	24.09	**11.98**	1.30	1.13	0.51	0.43	0.83
TarD	5.28	25.22	5.98	1.49	0.71	0.42	0.18	0.47
DeF	6.46	37.63	8.60	2.16	1.35	0.77	**0.54**	0.94
ReC	6.61	43.24	9.77	2.16	1.44	0.71	0.50	0.85
BDPartNet	**6.80**	**43.87**	10.67	**3.10**	**1.65**	**0.94**	0.52	**0.98**

Table 3: Quantitative experimental results based on the RoadScene dataset. Bold indicates the best value

	Dataset: RoadScene infrared-visible fusion dataset							
Method	EN	SD	SF	MI	SCD	VIF	Qbaf	SSIM
DID	7.43	51.58	14.66	2.11	1.70	0.58	0.48	0.86
U2F	7.09	38.12	13.25	1.87	1.70	0.60	0.51	0.97
SDN	7.14	40.20	13.70	2.21	1.49	0.60	0.51	**0.99**
RFN	7.21	41.25	16.19	1.68	1.73	0.54	0.45	0.90
TarD	7.17	47.44	10.83	2.14	1.55	0.54	0.40	0.88
DeF	7.23	44.44	10.22	2.25	1.69	0.63	0.48	0.89
ReC	7.36	52.54	10.78	2.18	1.74	0.59	0.43	0.88
BDPartNet	**7.54**	**54.47**	**16.27**	**2.41**	**1.90**	**0.69**	**0.54**	0.97

4.5 Ablation Study

Ablation studies involve assessing the effectiveness of specific modules within a network by systematically removing these modules and evaluating the overall performance. Through such experiments, the contribution of individual modules to the network's functionality can be discerned. Separately removing the L2 Loss, CEM module, and DEM module, the impact on performance was observed to conduct ablation experiments on our architecture. The qualitative findings of the experimental groups are depicted in Fig. 6, while the quantitative results are summarized in Table 4.

4.5.1 Reconstruction Loss study

During the process of image reconstruction, we imposed pixel-level constraints on both the decomposed feature maps of the two modalities and the source image. This ensures that the resulting fused image effectively retains the salient features of each modality. To further validate our approach, we conducted an ablation study aimed at eliminating the reconstruction loss. From Fig. 6, it can be observed that without the addition of L2-Loss, the fused image essentially only preserves the general outline of objects. For instance, as shown in Fig. 6, without L2-Loss, it is difficult to discern details such as bushes and bricks on the ground, highlighting a significant disparity in fusion effectiveness compared to our model.

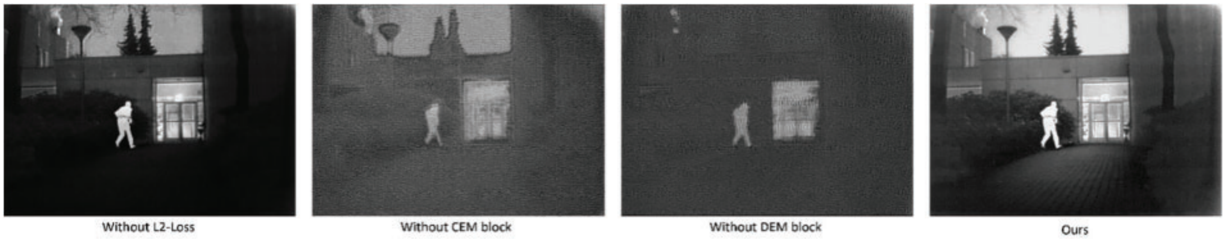

Figure 6: Ablation experiments on the L2-Loss, CEM, and DEM modules separately and compared the results

Table 4: Ablation experiment results in the testset of TNO. **Bold** indicates the best value

	W/O $L_2 - Loss$	W/O CEM	W/O DEM	Ours
EN	6.23	3.87	3.21	**7.24**
SD	37.81	30.26	26.5	**46.69**
VIF	0.48	0.21	0.26	**0.63**
SSIM	0.92	0.83	0.71	**1.02**

4.5.2 CEM and DEM Study

Our Contrast Enhancement Module (CEM) primarily focuses on enhancing edge contrast of objects and shapes of different sizes from a multi-scale perspective, thereby enriching the information in the fused image and maximizing the utilization of thermal radiation information. The Detail Enhancement Module (DEM) enhances textures through Laplacian operation and Sobel operator, making full use of the textures in the visible light images. The integration of these two modules yields an efficient fused image. As shown in Fig. 6 of the ablation experiment results, it is evident that the thermal radiation information of indoor scenes and human subjects becomes blurry, with less defined boundaries of thermal radiation targets and decreased local contrast. Since we concatenate two decomposed feature maps during the fusion stage and then perform dimensionality reduction, the ablation experiments only showcase half of the features. Hence, there is noticeable noise throughout the entire image. However, enhanced textures can still be observed in features such as trees, poles, and floor tiles. The Quantitative results of these experiments are presented in Table 4.

5 Conclusion

In response to the IVIF challenge, we introduced a novel Autoencoder (AE) network aimed at addressing the complexity of dual-scale image decomposition and reconstruction. Our approach leverages an encoder-decoder architecture, with the encoder responsible for dual-scale image decomposition and the decoder tasked with image reconstruction. Through extensive experimentation and analysis, we have demonstrated the effectiveness of our proposed method.

During the training phase, our model's encoder is trained to effectively output background and feature maps, which are then faithfully reconstructed by the decoder to reconstruct the original image. This process ensures that our model learns to capture both global contextual information and subtle details present in the input images, thereby enhancing its ability to accurately reconstruct complex scenes.

In the testing phase, we introduced a fusion layer between the encoder and decoder, promoting the fusion of background and detail feature maps through a specific fusion strategy. This fusion mechanism enables our model to effectively combine information from multiple scales and generate fused images that highlight targets and complex details. We conducted extensive experiments on the TNO, MSRS, and RoadScene datasets, yielding qualitative and quantitative evidence supporting the superiority of our proposed method.

Our qualitative analysis indicates that our model adeptly generates fused images, preserving important target features and effectively capturing subtle details present in the input scenes. Furthermore, our quantitative evaluation demonstrates significant improvements in various performance metrics (including SD, SF, etc.), further confirming the effectiveness of our method.

By addressing the key challenges posed by the IVIF problem and achieving superior performance on benchmark datasets, our proposed method represents a significant advancement in the field of image fusion. We believe that our work lays a solid foundation for future research in this area and holds tremendous potential for practical applications in domains such as surveillance, remote sensing, medical imaging, Earth observations, and beyond.

Acknowledgement: We sincerely thank the editors and every anonymous reviewer for their valuable comments, which greatly improved this article. Because your professionalism improves the quality of the journal.

Funding Statement: This work is supported in part by the National Natural Science Foundation of China (Grant No. 61971078) and Chongqing Education Commission Science and Technology Major Project (No. KJZD-M202301901).

Author Contributions: The authors confirm contribution to the paper as follows: Xuejie Wang completed the model design, experiments, and paper writing. Jianxun Zhang supervised the entire project. Ye Tao validated and analyzed the model. Yifan Guo collected data and performed data analysis. Xiaoli Yuan debugged the code and analyzed the model's inference speed.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Jianxun Zhang, upon reasonable request.

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

References

[1] 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, 2022. doi: 10.1109/TPAMI.2020.3012548.

[2] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, no. 4, pp. 153–178, 2019. doi: 10.1016/j.inffus.2018.02.004.

[3] 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, no. 10, pp. 28–42, 2022. doi: 10.1016/j.inffus.2021.12.004.

[4] J. Tao, Y. Cao, M. Ding, Z. Zhang, and G. Palmerini, "Visible and infrared image fusion-based image quality enhancement with applications to space debris on-orbit surveillance," *Int. J. Aerosp. Eng.*, vol. 2022, no. 5, pp. 1–21, 2022. doi: 10.1155/2022/6300437.

[5] S. Liu, P. Chen, and M. Woźniak, "Image enhancement-based detection with small infrared targets," *Remote Sens.*, vol. 14, no. 13, pp. 3232, 2022. doi: 10.3390/rs14133232.

[6] S. Wang, S. Huang, S. Liu, and Y. Bi, "Not just select samples, but exploration: Genetic programming aided remote sensing target detection under deep learning," *Appl. Soft Comput.*, vol. 145, no. 2, pp. 110570, 2023. doi: 10.1016/j.asoc.2023.110570.

[7] S. Liu, P. Chen, and Y. Zhang, "A multi-scale feature pyramid sar ship detection network with robust background interference," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 9904–9915, 2023. doi: 10.1109/JSTARS.2023.3325376.

[8] S. Liu *et al.*, "Human memory update strategy: A multi-layer template update mechanism for remote visual monitoring," *IEEE Trans. Multimed.*, vol. 23, pp. 2188–2198, 2021. doi: 10.1109/TMM.2021.3065580.

[9] F. Zhao, W. Zhao, L. Yao, and Y. Liu, "Self-supervised feature adaption for infrared and visible image fusion," *Inf. Fusion*, vol. 76, pp. 189–203, 2021. doi: 10.1016/j.inffus.2021.06.002.

[10] F. Zhao and W. Zhao, "Learning specific and general realm feature representations for image fusion," *IEEE Trans. Multimed.*, vol. 23, pp. 2745–2756, 2020. doi: 10.1109/TMM.2020.3016123.

[11] W. Zhao, H. Lu, and D. Wang, "Multisensor image fusion and enhancement in spectral total variation domain," *IEEE Trans. Multimed.*, vol. 20, no. 4, pp. 866–879, 2017. doi: 10.1109/TMM.2017.2760100.

[12] Y. Cao, D. Guan, W. Huang, J. Yang, Y. Cao and Y. Qiao, "Pedestrian detection with unsupervised multispectral feature learning using deep neural networks," *Inf. Fusion*, vol. 46, no. 2, pp. 206–217, 2019. doi: 10.1016/j.inffus.2018.06.005.

[13] C. Li, C. Zhu, Y. Huang, J. Tang, and L. Wang, "Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking," in *Proc. Eur. Conf. Comput. Uncoop. Spacecraft Human Pose Estim. Ter Vision*, Munich, Germany, 2018, pp. 808–823.

[14] Y. Lu *et al.*, "Cross-modality person re-identification with shared-specific feature transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, USA, 2020, pp. 13379–13389.

[15] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "MFNet: Towards realtime semantic segmentation for autonomous vehicles with multi-spectral scenes," in *Proc. IEEE Int. Conf. Intell. Rob. Syst.*, Vancouver, Canada, 2017, pp. 5108–5115.

[16] G. L. Civardi, M. Bechini, M. Quirino, A. Colombo, M. Piccinin and M. Lavagna, "Generation of fused visible and thermal-infrared images for uncooperative spacecraft proximity navigation," *Adv. Space Res.*, vol. 73, no. 11, pp. 5501–5520, 2024. doi: 10.1016/j.asr.2023.03.022.

[17] H. Wang and J. Han, "Research on military target detection method based on YOLO method," in *2023 IEEE 3rd Int. Conf. Inform. Technol., Big Data and Arti. Intell. (ICIBA)*, Chongqong, China, 2023, pp. 1089–1093.

[18] Y. Xiao, "Collection, utilization, protection and compliance governance of personal data in the vehicle in the development of auto-drive system," in *Seventh Int. Conf. Mechat. Intell. Rob. (ICMIR 2023)*, Kunming, China, SPIE, 2023, vol. 12779, pp. 636–642.

[19] L. Wang *et al.*, "Multi-modal 3D object detection in autonomous driving: A survey and taxonomy," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 7, pp. 3781–3798, 2023. doi: 10.1109/TIV.2023.3264658.

[20] A. Pandharipande, M. Lankhorst, and E. Frimout, "Luminaire-based multi-modal sensing for environmental building applications," *IEEE Sens. J.*, vol. 22, no. 3, pp. 2564–2571, 2021. doi: 10.1109/JSEN.2021.3137542.

[21] C. Qin *et al.*, "Unsupervised deformable registration for multi-modal images via disentangled representations," in *Int. Conf. Inform. Process. Med. Imag.*, Hong Kong, China, Cham, Springer International Publishing, 2019, pp. 249–261.

[22] S. Karim, G. Tong, J. Li, A. Qadir, U. Farooq and Y. Yu, "Current advances and future perspectives of image fusion: A comprehensive review," *Inf. Fusion*, vol. 90, pp. 185–217, 2023. doi: 10.1016/j.inffus.2022.09.019.

[23] Z. Zhao, S. Xu, J. Zhang, C. Liang, C. Zhang and J. Liu, "Efficient and model-based infrared and visible image fusion via algorithm unrolling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1186–1196, 2021. doi: 10.1109/TCSVT.2021.3075745.

[24] X. Wang, Z. Guan, S. Yu, J. Cao, and Y. Li, "Infrared and visible image fusion via decoupling network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022. doi: 10.1109/TIM.2022.3216413.

[25] S. Li, B. Yang, and J. Hu, "Performance comparison of different multi-resolution transforms for image fusion," *Inf. Fusion*, vol. 12, no. 2, pp. 74–84, 2011. doi: 10.1016/j.inffus.2010.03.002.

[26] J. Zong and T. Qiu, "Medical image fusion based on sparse representation of classified image patches," *Biomed. Signal Process. Control*, vol. 34, no. 2, pp. 195–205, 2017. doi: 10.1016/j.bspc.2017.02.005.

[27] X. Zhang, Y. Ma, F. Fan, Y. Zhang, and J. Huang, "Infrared and visible image fusion via saliency analysis and local edge-preserving multi-scale decomposition," *J. Opti. Soc. of Ame. A*, vol. 34, no. 8, pp. 1400–1410, 2017. doi: 10.1364/JOSAA.34.001400.

[28] U. Patil and U. Mudengudi, "Image fusion using hierarchical PCA," in *2011 Int. Conf. Image Inform. Process.*, Colorado Springs, USA, IEEE, 2011, pp. 1–6.

[29] 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, 2021. doi: 10.1007/s11263-021-01501-8.

[30] 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. 5009513, 2021. doi: 10.1109/TIM.2021.3075747.

[31] Y. Long, H. Jia, Y. Zhong, Y. Jiang, and Y. Jia, "Rxdnfuse: A aggregated residual dense network for infrared and visible image fusion," *Inf. Fusion*, vol. 69, no. 1, pp. 128–141, 2021. doi: 10.1016/j.inffus.2020.11.009.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, USA, 2016, pp. 770–778.

[33] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, USA, 2017, pp. 4700–4708.

[34] H. Li, Y. Cen, Y. Liu, X. Chen, and Z. Yu, "Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion," *IEEE Trans. Image Process.*, vol. 30, pp. 4070–4083, 2021. doi: 10.1109/TIP.2021.3069339.

[35] A. Dosovitskiy *et al.*, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Rep.*, 2021, pp. 1–12.

[36] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, "Endto-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[37] 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. Sin.*, vol. 9, no. 7, pp. 1200–1217, 2022. doi: 10.1109/JAS.2022.105686.

[38] 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*, vol. 83, no. 5, pp. 79–92, 2022. doi: 10.1016/j.inffus.2022.03.007.

[39] I. Goodfellow *et al.*, "Generative adversarial nets," in *27th Int. Conf. on Neu. Inform. Proce. Sys.*, Dec. 2014, vol. 2, pp. 2672–2680.

[40] 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, no. 4, pp. 11–26, 2019. doi: 10.1016/j.inffus.2018.09.004.

[41] 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, 2019. doi: 10.1109/TIP.2018.2887342.

[42] 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, 2020. doi: 10.1109/TIM.2020.3005230.

[43] 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, no. 9, pp. 72–86, 2021. doi: 10.1016/j.inffus.2021.02.023.

[44] L. Jian, X. Yang, Z. Liu, G. Jeon, M. Gao and D. Chisholm, "SEDRFuse: A symmetric encoder-decoder with residual block network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 5002215, 2021. doi: 10.1109/TIM.2020.3022438.

[45] H. Xu, X. Wang, and J. Ma, "DRF: Disentangled representation for visible and infrared image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 5006713, 2021. doi: 10.1109/TIM.2021.3056645.

[46] H. Xu, H. Zhang, and J. Ma, "Classification saliency-based rule for visible and infrared image fusion," *IEEE Trans. Comput. Imaging.*, vol. 7, pp. 824–836, 2021. doi: 10.1109/TCI.2021.3100986.

[47] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Med. Image Comput. Comput.-Assist. Interven.*, Munich, Germany, Springer International Publishing, Oct. 5–9, 2015, pp. 234–241.

[48] X. Mao, C. Shen, and Y. B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *30th Int. Conf. on Neu. Inform. Proce. Syst.*, Dec. 2016, pp. 2810–2818.

[49] E. Samkari, M. Arif, M. Alghamdi, and M. A. Al Ghamdi, "Human pose estimation using deep learning: A systematic literature review," *Mach. Learn. Knowl. Extract.*, vol. 5, no. 4, pp. 1612–1659, 2023. doi: 10.3390/make5040081.

[50] A. Toet and M. A. Hogervorst, "Progress in color night vision," *Opt. Eng.*, vol. 51, no. 1, pp. 1–20, 2012. doi: 10.1117/1.OE.51.1.010901.

[51] G. B. Palmerini, "Combining thermal and visual imaging in spacecraft proximity operations," in *2014 13th Int. Conf. Control Autom. Rob. Vis. (ICARCV)*, Singapore, IEEE, 2014, pp. 383–388. doi: 10.1109/ICARCV.2014.7064336.

[52] J. Zhang, Y. Zhu, W. Li, W. Fu, and L. Cao, "DRNet: A deep neural network with multi-layer residual blocks improves image denoising," *IEEE Access*, vol. 9, pp. 79936–79946, 2021. doi: 10.1109/ACCESS.2021.3084951.

[53] W. Xu *et al.*, "DRB-GAN: A dynamic resblock generative adversarial network for artistic style transfer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6383–6392.

[54] Z. Zhang and J. Yu, "STDGAN: Resblock based generative adversarial nets using spectral normalization and two different discriminators," in *Proc. 27th ACM Int. Conf. Multimed.*, Los Angeles, CA, USA, 2019, pp. 674–682.

[55] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004. doi: 10.1109/TIP.2003.819861.

[56] Z. Zhao *et al.*, "DIDFuse: Deep image decomposition for infrared and visible image fusion," arXiv preprint arXiv:2003.09210, 2020.

[57] Y. Ding, X. Yu, and Y. Yang, "RFNet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3975–3984.

[58] J. Liu *et al.*, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Waikoloa, USA, 2022, pp. 5802–5811.

[59] P. Liang, J. Jiang, X. Liu, and J. Ma, "Fusion from decomposition: A self-supervised decomposition approach for image fusion," in *Proc. ECCV*, Springer, Cham, 2022, pp. 719–735.

[60] Z. Huang, J. Liu, X. Fan, R. Liu, W. Zhong, and Z. Luo, "ReCoNet: Recurrent correction network for fast and efficient multi-modality image fusion," in *Proc. ECCV*, Springer, Cham, 2022, pp. 539–555.

[61] Z. Zhao *et al.*, "CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Waikoloa, USA, 2023, pp. 5906–5916.