DSA-Net: Infrared and Visible Image Fusion via Dual-stream Asymmetric Network

Ruyi Yin , Bin Yang * , Zuyan Huang , Xiaozhi Zhang

Posted Date: 30 June 2023

doi: 10.20944/preprints202306.2220.v1

Keywords: infrared and visible image fusion; transformer; deep learning; residual dense block

Article

DSA-Net: Infrared and Visible Image Fusion via Dual-stream Asymmetric Network

Ruyi Yin, Bin Yang *, Zuyan Huang and Xiaozhi Zhang

College of Electrical Engineering, University of South China, Hengyang 421001 China
* Correspondence: yangbin01420@163.com

Abstract: Infrared and visible image fusion technologies are used to characterize the same scene by diverse modalities. However, most existing deep learning-based fusion methods are designed as symmetric networks, which ignore the differences between modal images and lead to the source image information loss during feature extraction. In this paper, we propose a new fusion framework for the different characteristics of infrared and visible images. Specifically, we design a dual-stream asymmetric network with two different feature extraction networks to extract infrared and visible feature maps respectively. The transformer architecture is introduced in the infrared feature extraction branch, which can force the network to focus on the local features of infrared images while still obtaining their contextual information. And the visible feature extraction branch uses residual dense blocks to fully extract the rich background and texture detail information of visible images. In this way, it can provide better infrared targets and visible details for the fused image. Experimental results on multiple datasets indicate that DSA-Net outperforms state-of-the-art methods in both qualitative and quantitative evaluations. In addition, we also apply the fusion results to the target detection task, which indirectly demonstrates the fusion performances of our method.

Keywords: infrared and visible image fusion; transformer; deep learning; residual dense block

1. Introduction

Image fusion can combine images of the same scene captured by different sensors to obtain an image with rich information to make up for the shortage of information in single-sensor imaging, which is beneficial to the subsequent application of images. Infrared (IR) and visible (VIS) image fusion is a widely used branch of image fusion applications. The infrared images are obtained by the sensor capturing the infrared wavelength of the scene with significant thermal radiation information, which can effectively distinguish the target even under poor lighting or extreme weather conditions. However, the target contour edges as well as the background in the infrared images are always blurred. On the contrary, the visible image records the reflected light captured by the sensor and has rich texture details and structure information. So it is in accordance with human visual cognition. The infrared and visible fusion algorithm combines the advantages of both to generate a fused image with prominent targets and abundant texture information, which is widely used in military reconnaissance [1], industrial production [2], civilian surveillance [3], and other fields [4].

The purpose of infrared and visible image fusion is to extract and integrate the essential feature information from source images acquired by distinct imaging devices into a single fused image. Therefore, extracting the significant features of the fusion image is one of the central problems. Over the past few decades, numerous fusion methods have been proposed by researchers, which can be roughly divided into two categories: traditional fusion methods [5–7] and deep learning-based fusion methods [8–10]. Traditional fusion methods measure pixel's salience in the spatial domain or transform domain, and later design specific fusion rules to fuse them to obtain the fused image. Typical traditional methods include sparse representation-based methods [11,12], multi-scale transform-based methods [13,14], subspace-based methods [15,16], and hybrid methods [17]. Such methods have good performances by integrating reasonable fusion strategies into the fusion results. Unfortunately, traditional methods generally exhibit restricted fusion performance due to two reasons. First, the dependence of the traditional theory on manual design leads to the complexity of

the algorithmic framework, and this approach that uses fixed mathematical variations to extract features ignores the modal differences between source images. Secondly, the limited choice of reasonable fusion rules also somewhat limits the performances.

Over the past few years, deep learning techniques have gained considerable action in the domain of computer vision [18,19]. Relying on powerful nonlinear representational capability of convolutional neural networks (CNN), fusion performance of deep learning-based image fusion methods is typically superior to that of traditional methods. However, they cannot conquer the problem of long-range contextual information dependencies, which leads to the loss of some significant global features. Moreover, these methods extract differ modal features by symmetric models, ignoring the modal differences between infrared and visible images. During the process of feature extraction, it is challenging to prevent the loss of intricate details, resulting in a decay of features in the fused results. To solve these problems, we propose an asymmetric network for infrared and visible image fusion. For the facts that the modalities of infrared and visible images are different, we design two unique feature extraction modules separately. The residual dense block is used in the visible image feature extraction module to minimize the degradation of intermediate feature and retain the intricate features of visible images to the fullest extent. The transformer is embedded in the infrared image feature extraction module. This module utilizes the attention mechanism of the transformer to focus on important thermal radiation information in infrared images, ignoring redundant background information. The infrared image feature extraction module also uses dense concatenation to repeatedly utilize previous features, which can reduce feature loss. We also design a depth feature fusion block for global depth features extraction and fusion, and use residual connections to reduce network degradation. Finally, the fused images with high brightness and clear edge contours as well as significant infrared targets are obtained. In brief, the principal achievements and contributions of this paper can be briefly summarized as follows.

Asymmetric network: we design an asymmetric network to build two different feature extraction modules. It is good for the different modal characteristics of infrared and visible images, preserving the infrared target and visible texture information respectively. Subsequently, the two modality features were fused via the main pathway.

Transformer: Although convolutional neural networks are more capable of acquiring local information, they have limited performance in maintaining remote contextual information of the source images. In contrast, Transformer can expand the receptive field of the image and acquire more contextual information by global relationship modeling. Therefore, we embed the transformer into the CNN network so that the network can inherit the advantages of CNN and the transformer and maximize the retention of global and local features.

Generalization capability: The experimental results on the RoadScene dataset and TNO dataset indicate that DSA-Net performs better than the other nine representative advanced methods both in subjective and objective evaluations. Furthermore, the experiment has also been expanded to include object detection, and the results demonstrate that our method has a greater potential in advancing advanced computer vision tasks.

The rest of this paper is organized as follows. Section 2 reviews some works related to our method. Section 3 describes the proposed DSA-Net in detail, including the overall framework, network architecture, and loss function. In Section 4, we conduct comparative experiments to validate the merits of the proposed approach. In addition, we perform ablation experiments, generalization experiments, and applications in target detection. Finally, conclusions are given in Section 5.

2. Related Work

In this section, we first review the existing infrared and visible image fusion algorithms in Section 2.1, followed by a brief introduction to the Transformer in Section 2.2.

2.1. Deep Learning-Based Fusion Methods

In 2017, Liu et al. [20] firstly use CNN to extract image features and design fusion strategies to achieve image fusion tasks. However, this method was limited to multi-focus image fusion tasks.

 doi:10.20944/preprints202306.2220.v1

Later, Liu et al. [21] presented an analogous approach for infrared and visible image fusion. They used CNN to obtain the weight map and obtained the fused image through a series of post-processing. These two methods simply use CNN for feature extraction only, and other parts of the fusion framework still need to be designed manually, without completely getting rid of the traditional algorithms. In 2018, Li et al. [22] proposed an encoding-decoding framework which uses dense concatenation in the encoder to fully extract feature maps and also designs a fusion strategy to combine the extracted features. Then the decoder is utilized to decode the features and reconstruct the fused image. However, this approach still requires the manual design of fusion rules and cannot fully achieve end-to-end fusion. In 2019, Ma et al. [23] introduced generative adversarial networks (GAN) into infrared and visible image fusion, using a pair of simple generators and discriminators to obtain fused images. Subsequently, Ma et al. [24] proposed a model called DDcGAN, which uses a dual discriminator to reduce the loss of source image information. However, GAN is not stable in unsupervised learning tasks and the fused image edge contours may be blurred. In 2020, Xu et al. [25] proposed a unified fusion model that trains the model by learning multiple fusion tasks continuously to avoid catastrophic forgetting, storage, and computation problems. In 2021, Liu et al. [26] propose a deep network for infrared and visible image fusion using a feature learning module with a fusion learning mechanism to optimize the fusion effect. In 2022, Tang et al. [27] proposed a Y-shape fusion framework and used a dynamic transformer module to acquire local features and important contextual information.

2.2. Transformer

In 2017, Vaswani first proposed the concept of a Transformer [28] to capture more long-range information. Which conquers the inherent problem of CNN, i.e., long memory loss, by employing multi-headed self-attention. Since then, Transformer has swept the field of natural language processing (NIP) [29,30]. In 2020, Dosovitskiy proposed a vision Transformer (VIT) for image classification [31], which was the first application of a Transformer in the field of vision. Since then, transformer has been extensively developed in the field of vision. For example, a new Transformer network for medical image segmentation [32], an end-to-end video instance segmentation [33], a pure semantic segmentation [34], and even better models of visual Transformer [35,36] for other vision tasks. Recently, transformer have also been widely used in image fusion tasks. In 2021, Vs et al. [37] put forward a transformer-based multi-scale fusion strategy, which captures local and global features, by using spatial CNN branches and Transformer branches for multi-scale feature fusion. Zhao et al. [38] used density nets for encoding and dual Transformer to focus and integrate information from the infrared and visible images. Subsequently, Fu et al. [39] presented a patch pyramid Transformer (PPT) for image fusion, a patch transformer is designed to transform the image into a series of patches and then leverage the pyramid Transformer for feature extraction. Rao et al. [40] came up with a lightweight fusion framework by combining Transformer and adversarial learning, where a generator was designed for generating the fused image and two discriminators for optimizing the perceptual quality of the fused images.

3. Proposed Method

In this section, we present the structure of our method and the loss function in detail.

3.1. Framework Overview

Figure 1 shows the framework of our proposed network. The main framework consists of three parts, the infrared feature extraction module, the visible feature extraction module, and the merge module. Since infrared images contain strong thermal radiation information and can effectively distinguish targets, we use a combination of CNN and transformer for infrared image feature extraction. The transformer is designed to model global dependencies, the network architecture of which is shown in Figure 2. Visible images are rich in texture details and background information, so we use the densely connected convolution layer to extract local features, the network architecture of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2023 doi:10.20944/preprints202306.2220.v1

which is shown in Figure 6. With these two branches, the useful information from the source images can be fully extracted. Then these features are concatenated together and fed into a merge module. This module uses convolution skip joints to produce a continuous memory mechanism, which can adaptively learn more effective features from previous and current local features and stabilize the training of the broader network. The decoding block is used for the generation of subsampling and fusion results and consists in turn of a convolution layer with a kernel size of 3×3, a batch normalization (BN), and a corrected linear unit (ReLU). Since our network is an end-to-end network, the output of the network is the fused image.

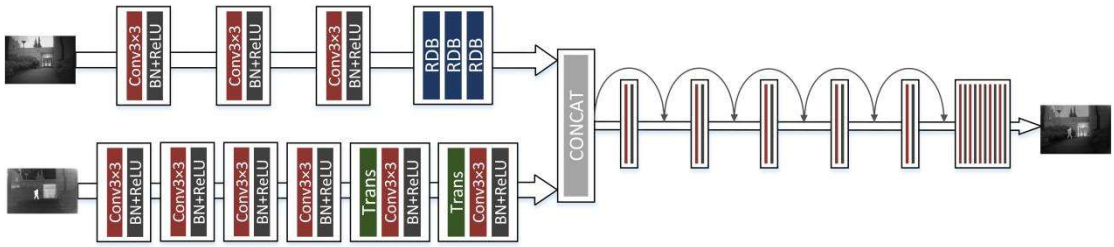


Figure 1. DSA-Net framework structure.

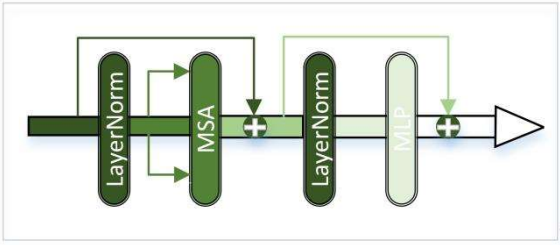


Figure 2. Transformer framework structure.

3.2. Infrared Feature Extraction Module

Infrared images have strong target information, in order to obtain infrared feature maps with local enhancement, we use a combination of CNN and transformer to extract infrared image features. The framework of transformer, adopts multi-head self-attention and has good global contextual feature exploration capability. Which is shown in Figure 2. Transformer consists of two LayerNorm, multi-head self-attention (MSA), and multilayer perceptron (MLP). LayerNorm normalizes the features, which can keep similarities between different channels' statistical properties and enhances the generalization ability of the model. After normalization, the features are linearly projected into multiple feature subspaces to obtain attention weights Q, weight indexes K, and feature vectors V. Then parallel processing is performed using multiple independent scaled dot product attention, as shown in Figure 3. Compared with single attention, MSA can effectively prevent the model from over-focusing on its own location when encoding information about the current location. The scaled dot product attention is shown in Figure 4. The similarity matrix is obtained by using Q and K for dot product operation; Scale represents the quantization operation, which can prevent the similarity matrix variance from being too large and make the training gradient update more stable; Mask is the padding operation, but unlike the ordinary padding 0, it will be padded with negative infinity, and then normalized by Softmax layer to get the attention weight matrix, and the attention of the padding part will be 0, which does not affect the subsequent operations. Finally, the attention map image is generated by multiplying the feature vector V with the corresponding attention weights. The attention process can be expressed as Equation (1). The MLP is shown in Figure 5 and consists of Full Connection, GELU activation function, and Dropout. The MLP layer can perform nonlinear transformations on the features, which can be better adapted to complex image tasks. Moreover, MLP layer can extract higher-level features from the input features, which can represent information such as objects and backgrounds in an image. And the residual connection in Transformer can effectively solve the problem of gradient disappearance and the degradation of the weight matrix. Through the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2023 doi:10.20944/preprints202306.2220.v1

above operations, the Transformer uses the self-attention mechanism to establish the relationship between image features, which can capture the global information and long-distance dependence in an image.

$$Attention(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

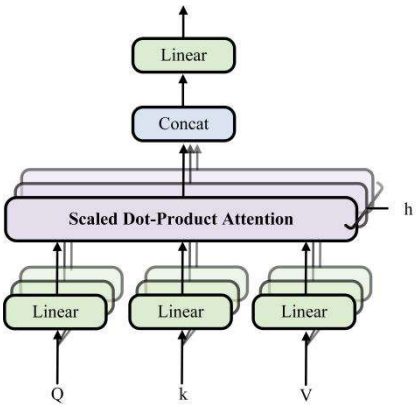


Figure 3. MSA structure details.

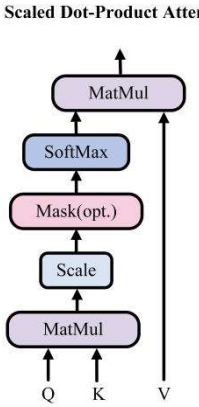


Figure 4. Scaled Dot-Product Attention structure details.

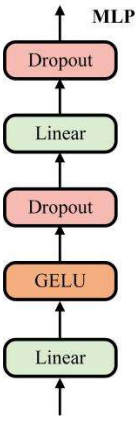


Figure 5. MLP structure details.

3.3. Visible Feature Extraction Module

Visible images have a higher spatial resolution and contain more texture details. Therefore, we design a visible feature extraction module with a residual dense block (RDB) [41] to extract the visible features. In this module, we first extract and obtain visible shallow features using three convolution

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2023 doi:10.20944/preprints202306.2220.v1

layers, followed by deep feature extraction by the RDB. The RDB is shown in Figure 6, which consists of five convolution layers. Each convolution layer can acquire the features of all previous layers through local dense connections, thus making full use of the features of each layer. The final convolution layer filters all the previous features and adaptively controls the output information. Finally, the shallow and deep feature results are combined using residual connectivity, and the residual connectivity enhances the gradient connectivity, which can effectively prevent the gradient from disappearing. The visible image can fully extract its local features through its feature extraction module, prevents the degradation of intermediate features, and obtains a feature map with rich texture detail features.

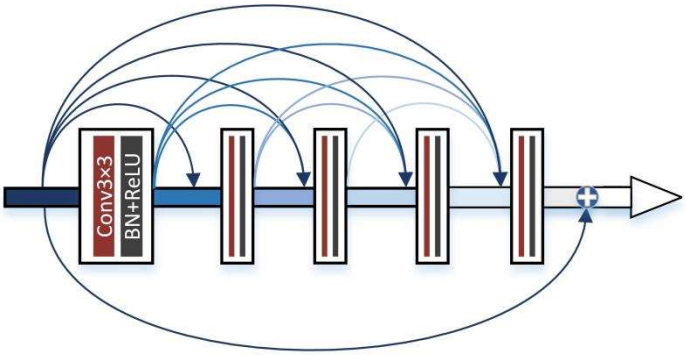


Figure 6. RDB structure details.

3.4. Merge Module

The features obtained from the infrared feature extraction module and the visible feature extraction module are concatenated as the input to the merge module. The merge module consists of ten convolution layers and skip connections. The convolution layers all consist of 3×3 convolution, BN, and ReLU activation functions. The first five convolution layers are used to extract the depth features of the infrared and visible images. As the network depth increases, the issue of feature degradation is more likely to arise, which can be addressed by incorporating skip connections. And the skip connections also use the learned features of the previous layer in this layer, which achieves feature reuse. Finally, the extracted depth features are used in the last five convolution layers, to achieve features decode, and to obtain the fused image.

3.5. Loss Function

Since our method is unsupervised learning, the loss function plays a crucial role in the fusion effect. It is an important challenge to fully retain the features of the source images, such as the infrared salient targets in infrared images and the detailed textures in visible images. Therefore, in order to fully retain the source image information, our loss function consists of three types of loss terms, structure loss L_{ssim}, intensity loss L_{int}, and gradient loss L_{grad}. The structure loss constrains the similarity between the fused image and the source images. The intensity loss constrains the fused image to maintain a similar intensity distribution as the source image while the gradient loss enforces the presence of additional texture details in the fused image. The loss function of the network can be expressed as follows:

$$L = \alpha L_{ssim} + \beta L_{int} + \gamma L_{grad}, \tag{2}$$

where α, β and γ are the weighting factors of the three loss functions, which are used to control the total loss function balance.

Ensure that the fused image has similar structural information to the source images, which can be expressed as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2023 doi:10.20944/preprints202306.2220.v1

$$L_{ssim} = \lambda_{VIS_{ssim}} \left(1 - SSIM\left(I_F, I_{VIS}\right)\right) + \lambda_{IR_{ssim}} \left(1 - SSIM\left(I_F, I_{IR}\right)\right), \tag{3}$$

I_{VIS}, I_{IR} and I_F denote the visible image, the infrared image, and the fused image of both, respectively. $\lambda_{VIS_{ssim}}$ and $\lambda_{IR_{ssim}}$ represents the SSIM loss weights between the fused image and the visible and infrared images. $SSIM(\bullet)$ denotes the structural similarity operation between the fused image and the source images, which is defined as follows:

$$SSIM\left(I_X, I_Y\right) = \frac{2\mu_X\mu_Y + C_1}{\mu_X^2 + \mu_Y^2 + C_1} \cdot \frac{2\sigma_X\sigma_Y + C_2}{\sigma_X^2 + \sigma_Y^2 + C_2} \cdot \frac{\sigma_{XY} + C_3}{\sigma_X\sigma_Y + C_3}, \tag{4}$$

where μ denotes the mean and σ denotes the standard deviation or covariance. C_1, C_2, and C_3 are constants to prevent $\mu_X^2 + \mu_Y^2$, $\sigma_X^2 + \sigma_Y^2$ and $\sigma_X\sigma_Y$ being 0 from causing formula instability. It constrains the loss and distortion of the fused image from the similarity of brightness, contrast, and structural information.

The SSIM loss function is weakly constrained in terms of pixel intensity, while the significant targets in visible images have great pixel intensity. Therefore, we also design the intensity loss to retain the infrared targets in the source image.

$$L_{int} = \lambda_{VIS_{int}} \left(\frac{1}{HW}\left\|I_F - I_{VIS}\right\|_2^2\right) + \lambda_{IR_{int}} \left(\frac{1}{HW}\left\|I_F - I_{IR}\right\|_2^2\right), \tag{5}$$

where $\lambda_{VIS_{int}}$ and $\lambda_{IR_{int}}$ represent the intensity loss weights between the fused image and the visible and infrared images. H and W denote the height and width of the fused image. $\|\bullet\|_2$ is the l2-norm.

In addition, we use the gradient loss constraint to fuse the images to retain the detailed textures in the visible images as well as the target edges of the infrared images.

$$L_{grad} = \lambda_{VIS_{grad}} \left(\frac{1}{HW}\left\|\nabla I_F - \nabla I_{VIS}\right\|_2^2\right) + \lambda_{IR_{grad}} \left(\frac{1}{HW}\left\|\nabla I_F - \nabla I_{IR}\right\|_2^2\right), \tag{6}$$

where $\lambda_{VIS_{grad}}$ and $\lambda_{IR_{grad}}$ represents the gradient loss weight between the fused image and the visible and infrared images. ∇ denotes the gradient operator.

Due to the optimization of the above loss function, the fused image can well retain the structural information, intensity information, and gradient information of the source images. We hope that the fusion image retains more structural information of the visible image, combined gradient information, and more infrared image intensity information. Therefore, the loss weights described above should meet the following conditions:

$$\lambda_{VIS_{ssim}} > \lambda_{IR_{ssim}}, \lambda_{VIS_{int}} < \lambda_{IR_{int}}, \lambda_{VIS_{grad}} > \lambda_{IR_{grad}}, \tag{7}$$

4. Experiments

In this section, the experimental configuration and experimental details will be outlined in Section 4.1. Then, we present the comparison methods and objective evaluation metrics in subsection 4.2. The ablation experiments on the network structure are presented in subsection 4.3, demonstrating the rationality of our network structure. The comparison experiments and generalization experiments are presented in subsection 4.4 and subsection 4.5, respectively, revealing the superiority of our proposed method. Finally, we perform target detection task-driven evaluation experiments in 4.6 to evaluate different fusion methods from the perspective of advanced vision tasks.

4.1. Experimental Configuration and Experimental Details

Two mainstream datasets TNO dataset and RoadScene dataset were used in this work. We collected 51 and 83 pairs of infrared and visible image pairs from these two datasets, respectively. Then 50 pairs are randomly selected from the RoadScene datasets as the training data, while all the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2023 doi:10.20944/preprints202306.2220.v1

remaining image pairs were used as the test data. To obtain sufficient training samples, the training data is expanded using an overlapping cropping strategy. It is worth mentioning that the cropping strategy is a widely used data enhancement method in the image domain. In our experiments, the RGB images in the RoadScene dataset are converted to the YUV color model, the Y channel is used for image fusion, and finally the fused images are converted to RGB images.

Specifically, 40964 pairs of infrared and visible image patches with 120 × 120 size are generated for network training. Since the cropping strategy is only used for data expansion, the test data is not used. Therefore, by feeding the entire image into the trained model, fusion results can be generated. In our experiments, the epoch is 25 and the batch size is fixed at 29. The learning rate was set to 0.001 and the Adam optimizer was used for model optimization. The three weighting factors α, β and γ in the loss function are specified as 1.1, 10 and 10, respectively. All experiments were conducted on a computer with an Intel(R) Core(TM) i9-10920X CPU @ 3.50GHz and an NVIDIA GeForce RTX 3090 GPU. And the proposed deep model is implemented on the PyTorch framework.

4.2. Comparison Methods and Evaluation Indicators

To ensure a thorough evaluation of the proposed algorithm, we performed experiments on both the RoadScene and TNO datasets. We compared our approach with nine state-of-the-art methods, including three representative traditional methods, namely GF [42], ADF [43], and IVFusion [44], and six deep learning-based methods, namely DenseFuse [22], GAN-FM [45], DDcGAN [24], YDTR [27], IFCNN [46], and PMGI [25]. The implementations of all nine methods are publicly available and we set the optional parameters in the same way as reported in the original paper.

For quantitative evaluation, six metrics are selected to objectively assess the fusion performance, including structure similarity index measure (SSIM) [47], mean square error (MSE) [48], correlation coefficient (CC) [49], peak signal-to-noise ratio (PSNR) [50], the sum of correlations of differences (SCD) [51], and Chen-Blum Metric (Q_{CB}) [52]. SSIM evaluates the structural loss and distortion of fused images from the human visual system's perspective, and MSE calculates the error between the fused images and the source images. CC measures the degree of linear correlation between the fused images and the source images. PSNR measures the ratio of peak power to the noise power in the fused images. SCD measures the maximum information of the fused images containing each source image. Q_{CB} evaluates the image quality of the fused images based on the human visual system model. In addition, larger SSIM, CC, PSNR, SCD, and Q_{CB} indicate better fusion performances. Smaller MSE indicates better fusion performances.

4.3. Ablation Experiments

To investigate the effectiveness of our asymmetric network structure, Transformer-based infrared feature extraction module, and RDB-based visible feature extraction module, we perform ablation validation on TNO and RoadScene datasets. We divide the model structure into five groups of types, a) Transformer-based dual-stream symmetric network (D-Trans): in order to verify the effectiveness of the asymmetric network structure, we apply the infrared feature extraction module of this paper to visible image feature extraction as well, and construct a symmetric network. b) RDB-based dual-stream symmetric network (D-RDB): also to verify the effectiveness of the asymmetric network structure, we apply the visible feature extraction module of this paper to infrared image feature extraction. c) Without Transformer (O-Trans): in order to investigate the importance of the Transformer, we move the Transformer out of the infrared feature extraction module to study its function. d) Without RDB (O-RDB): to verify the necessity of RDB, we remove the RDB in the network to illustrate its validity. e) To verify the effectiveness of the Transformer and RDB for infrared and visible feature extraction, respectively, we exchange their extraction modules (E-FEM).

4.3.1. Qualitative Comparisons

Figure 7 and Figure 8 show the fusion results of the TNO and RoadScene datasets, respectively. To allow for better comparison, we zoomed in for a close-up of a local area in each fusion result. From

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2023 doi:10.20944/preprints202306.2220.v1

the fusion results of D-Trans and D-RDB, it can be seen that we changed the asymmetric network to a two-stream symmetric network, resulting in blurred edges of infrared targets and insufficient clarity of the scene, which indicates that the proposed asymmetric network has better complementary information preservation capabilities. In the absence of the Transformer module, the infrared feature extraction module cannot capture the infrared protruding target well due to the failure to build the long-distance dependency well. So the infrared character target in Figure 7 and the clouds in the sky in Figure 8 are relatively blurry. As for the case without RDB, we can see that its results fail to fully extract the visible details although the infrared target is better maintained, and the background texture of the fused image and the landmark lines of the RoadScene are not clear enough. In addition, the E-FEM fusion is not well preserved in both infrared target and texture details, which can prove the effectiveness of the Transformer and RDB for infrared and visible feature extraction, respectively.

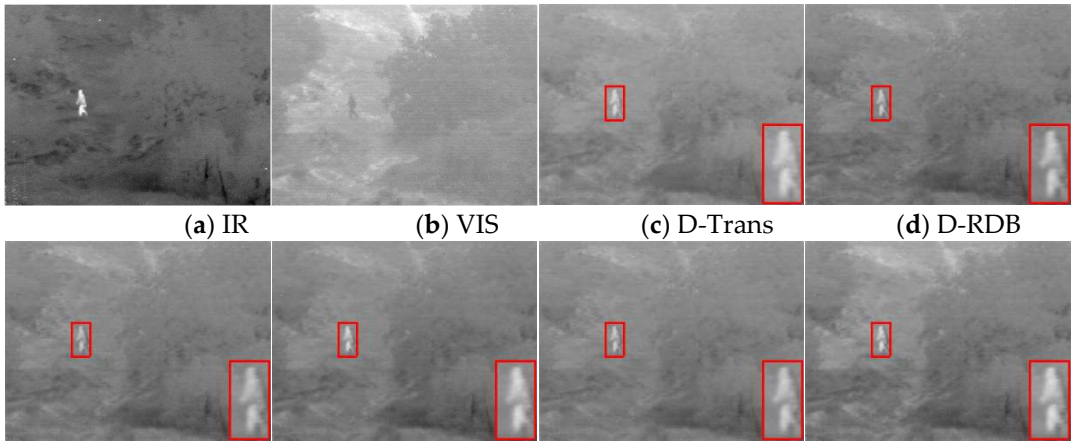


Figure 7. Subjective results of the ablation experiment TNO datasets.

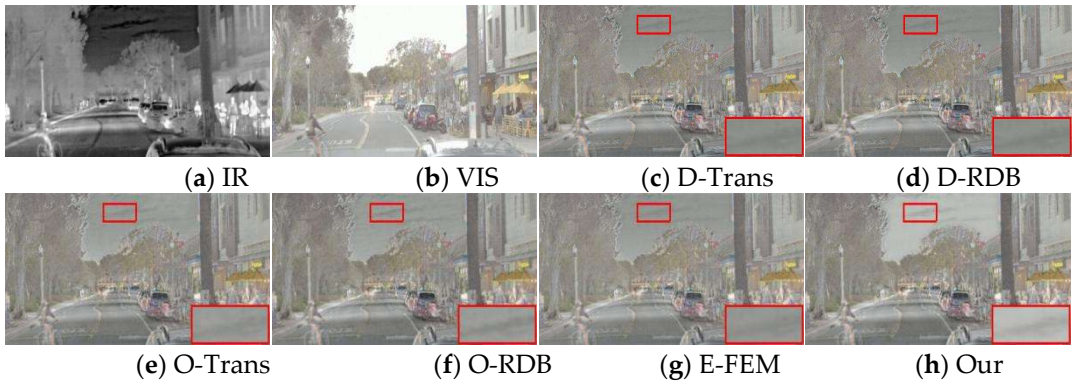


Figure 8. Subjective results of the ablation experiment RoadScene datasets.

4.3.2. Quantitative Comparisons

To evaluate the ablation experiments more objectively, we assessed the quality of their fusion results using image quality metrics. Table 1 shows the objective results in two different datasets. The table highlights the top-performing results in bold font, while the second-best results are indicated in italic font. It is easy to see that our final method has the best overall score ranking in both the TNO datasets and the RoadScene datasets. Combining this with the subjective evaluation can demonstrate the effectiveness of our network structure and the individual modules in the network.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2023 doi:10.20944/preprints202306.2220.v1

Table 1. Objective results of the ablation experiment RoadScene and TNO datasets.

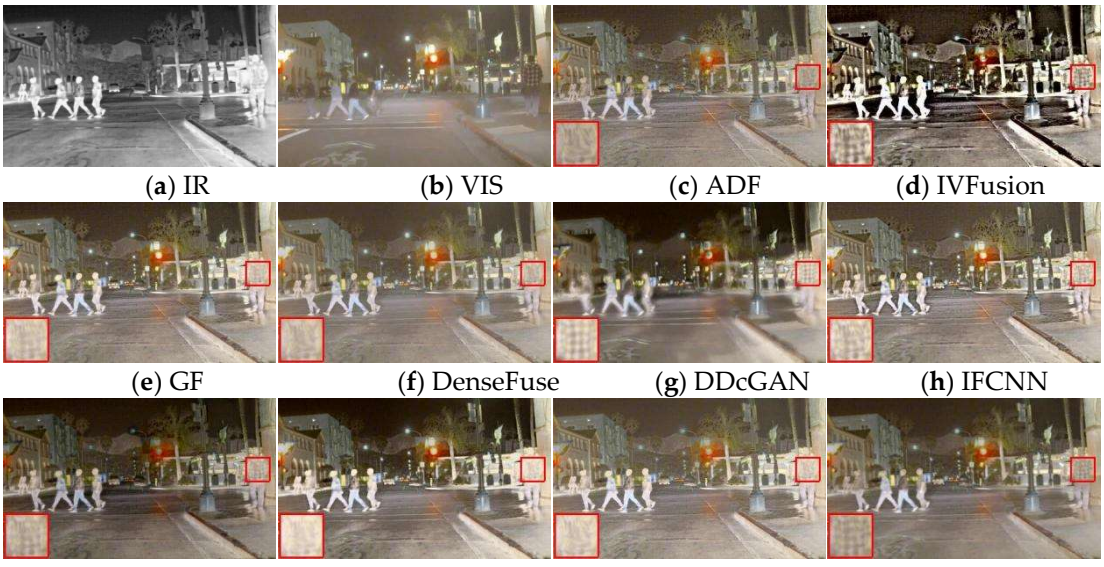
Datasets	Methods	Quality Metrics					
		SSIM	MSE	CC	PSNR	SCD	Q_{CB}
RoadScene	D-Trans	0.7141	125.1179	0.7754	27.1705	1.2882	0.4773
	D-RDB	*0.7317*	72.9455	0.7762	29.5688	1.2598	0.5086
	O-Trans	0.7306	68.2725	*0.7837*	29.8866	1.2907	0.5027
	O-RDB	0.7192	*66.7998*	0.7732	*30.0231*	*1.3115*	*0.5112*
	E-FEM	**0.7330**	82.1417	0.7815	29.0312	1.2557	0.5021
	Our	0.7277	**46.6912**	**0.7990**	**31.5830**	**1.3218**	**0.5469**
TNO	D-Trans	0.7125	117.3409	*0.5523*	27.4721	1.5780	0.4768
	D-RDB	0.7289	83.4061	0.5416	29.1615	1.4354	0.4805
	O-Trans	0.7539	*83.2827*	0.5279	*29.1897*	1.4338	0.4892
	O-RDB	*0.7605*	83.8169	**0.5553**	29.1480	1.5119	*0.4945*
	E-FEM	**0.7624**	90.1372	0.5351	28.7422	1.4452	0.4844
	Our	0.7587	**80.6446**	0.5452	**29.3554**	*1.5180*	**0.5000**

4.4. Comparative Experiments

To fully evaluate the fusion performance of our approach, we first compared the proposed method with nine other algorithms on the RoadScene datasets.

4.4.1. Qualitative Comparisons

We randomly selected 50 of the 83 infrared and visible image pairs as the training set and used the remaining 33 image pairs as the test set. As shown in Figure 9, our fusion results outperform the other methods in improving the visual quality and integrating complementary information. To show the difference more clearly, we zoom in the red boxed area and can observe that the three traditional methods ADF and GF have a loss of clothing texture details and a general prominence of significant targets. IVFusion fusion results do not match human visual effects and look unnatural. DDcGAN retains texture details while failing to poor image quality, producing significant artifacts. The methods of DenseFuse, IFCNN, and GAN-FM maintain the intensity information of infrared and have high overall contrast, but the detailed information of visible images is more severely weakened. (e.g., stripes on clothes, bicycle markings on the ground). And the PMGI and YDTR methods have too much useless information.



(**a**) IR (**b**) VIS (**c**) ADF (**d**) IVFusion

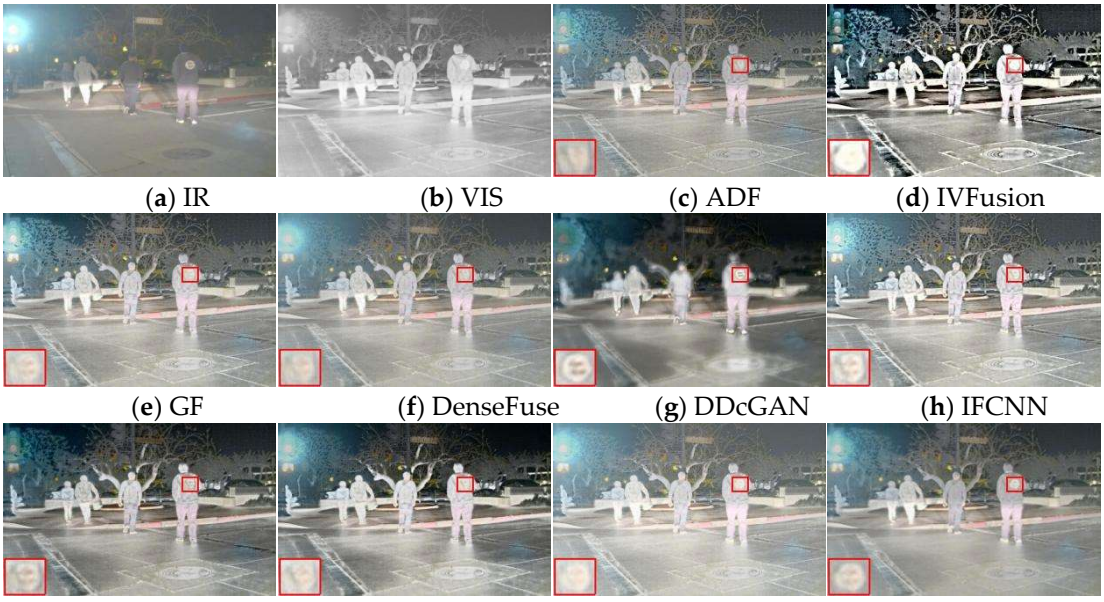
(**e**) GF (**f**) DenseFuse (**g**) DDcGAN (**h**) IFCNN

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2023 doi:10.20944/preprints202306.2220.v1

(**i**) PMGI (**j**) GAN-FM (**k**) YDTR (**l**) Our

Figure 9. Graph of fusion results of different methods.

Figure 10 shows the second set of source images of different methods and their fused image results. All nine methods have their own advantages but still have some drawbacks compared with our method. Specifically, both IVFusion and DDcGAN are inferior to all other methods from a visual sensory perspective. From the perspective of texture detail preservation, the methods of IFCNN, PMGI, GAN-FM, and YDTR inevitably suffer from the infrared thermal radiation information, blurring the background and visible features (e.g., patterns in zoomed-in regions, distant tree branches). However, it is worth mentioning that they retain sufficient infrared salient target information. In contrast, the ADF, GF, and DenseFuse methods are able to balance visible and infrared information, highlighting salient targets while retaining rich texture details. However, they are still inferior to DAS-Net, and in the enlarged area of the red box, only our method clearly shows the pattern on the clothes. In summary, only our method can effectively integrate the complementary information from the source image and simultaneously ensure the visual quality of the fused image.



(**a**) IR (**b**) VIS (**c**) ADF (**d**) IVFusion

(**e**) GF (**f**) DenseFuse (**g**) DDcGAN (**h**) IFCNN

(**i**) PMGI (**j**) GAN-FM (**k**) YDTR (**l**) Our

Figure 10. Graph of fusion results of different methods.

4.4.2. Quantitative Comparisons

We selected 33 image pairs from the RoadScene datasets for quantitative evaluation. The quantitative results for the six statistical metrics are shown in Figure 11 and Table 2. For each metric, the best and second best fusion results for all methods are marked in bold and italics, respectively. It can be observed that our method has outstanding stability and advantages on the RoadScene datasets. The best CC show a strong correlation between the fused image and the source image, which means that our method is more compatible with the human visual system. In addition, although PMGI performs well in individual metrics, it ranks low in three metrics, SSIM, CC, and SCD. In contrast, our method performs well overall, ranking higher across all metrics. A large number of qualitative and quantitative results on the RoadScene datasets show that our method can generate fused images that conform to human vision and retain the information of the source images to the maximum extent.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2023 doi:10.20944/preprints202306.2220.v1

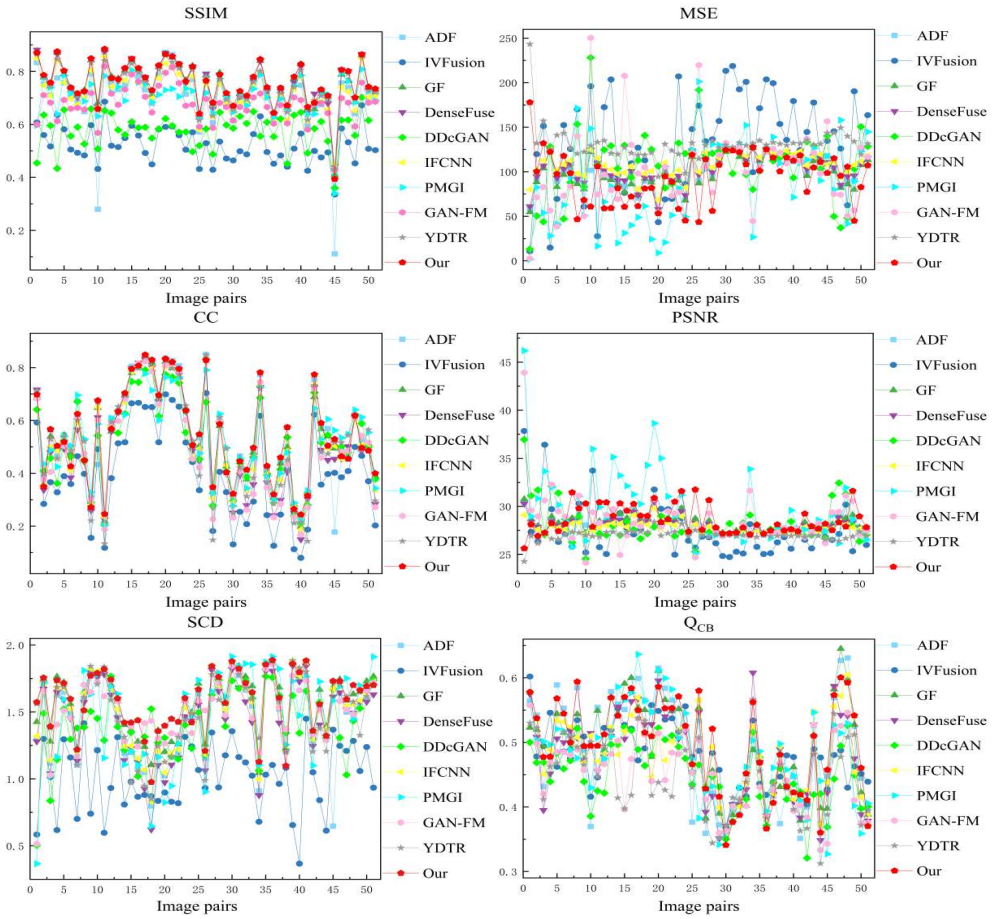


Figure 11. Graphs of performance metrics for different fusion methods.

Table 2. Objective results of the comparative experiment RoadScene datasets.

Methods	Quality Metrics					
	SSIM	**MSE**	**CC**	**PSNR**	**SCD**	**Q_{CB}**
ADF	0.6909	93.2764	0.7810	28.4483	1.0776	0.5285
IVFusion	0.4642	60.1766	0.6855	30.3363	0.9305	0.4535
GF	0.7190	90.6036	0.7769	28.5830	1.2868	0.5465
DenseFuse	**0.7453**	93.8907	*0.7851*	28.4170	1.0819	0.5440
DDcGAN	0.5589	48.4804	0.7410	31.5353	1.1833	0.4594
IFCNN	0.7045	99.6146	0.7694	28.1568	1.1557	0.4973
PMGI	0.6777	**26.9935**	0.7120	**34.2283**	0.9673	**0.5852**
GAN-FM	0.6590	52.2970	0.7680	30.9994	**1.3848**	0.5327
YDTR	0.7231	131.9041	0.7771	26.9622	1.1619	0.5236
Our	*0.7277*	*46.6912*	**0.7990**	*31.5830*	*1.3218*	*0.5469*

4.5. Generalization Experiments

Generalization performance is an important aspect of evaluating deep learning-based methods. Therefore, we provide generalization experiments on the TNO datasets to demonstrate the generalizability of the proposed approach. It is worth mentioning that our fusion model is trained on the RoadScene datasets and tested directly on the TNO datasets.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2023 doi:10.20944/preprints202306.2220.v1

4.5.1. Subjective Results

As shown in Figure 12, the fusion results obtained by the different methods in the TNO dataset introduce some meaningless information, which is reflected in the loss of texture details and the diminution of significant targets. To visualize the effect of the fused images, we zoom in on the region with rich texture details in the red box. We can observe that compared with our method, the three traditional methods ADF, IVFusion, and GF have some degree of loss of door frame details, which is because the texture details in the background region are contaminated by the thermal radiation information, especially the IVFusion method, which failing to preserve the useful information of the source image well and the overall visual effect is poor. The DDcGAN method has a limited ability to extract texture details from the visible image, and not only has a distortion problem but also cannot preserve the sharpened edges of the target. As for the method of YDTR, the intensity information of the significant targets is diminished to different degrees, and the overall contrast is low. In the PMGI results, the thermal radiation information is obviously disturbed, the sky color is not clean enough, while the texture of door frames, grassy areas and roads are not adequately preserved. It is worth mentioning that DenseFuse, IFCNN and GAN-FM interfere less with the useless information, but the texture details are still lost and the edges of the door frame are not clear enough. Overall, DAS-Net provides a good visual effect, on one hand, our method maintains clear background information, such as bright skies, layered grasses, and door frames with distinct edges, on the other hand, the major significant information from the infrared image is clearly highlighted.

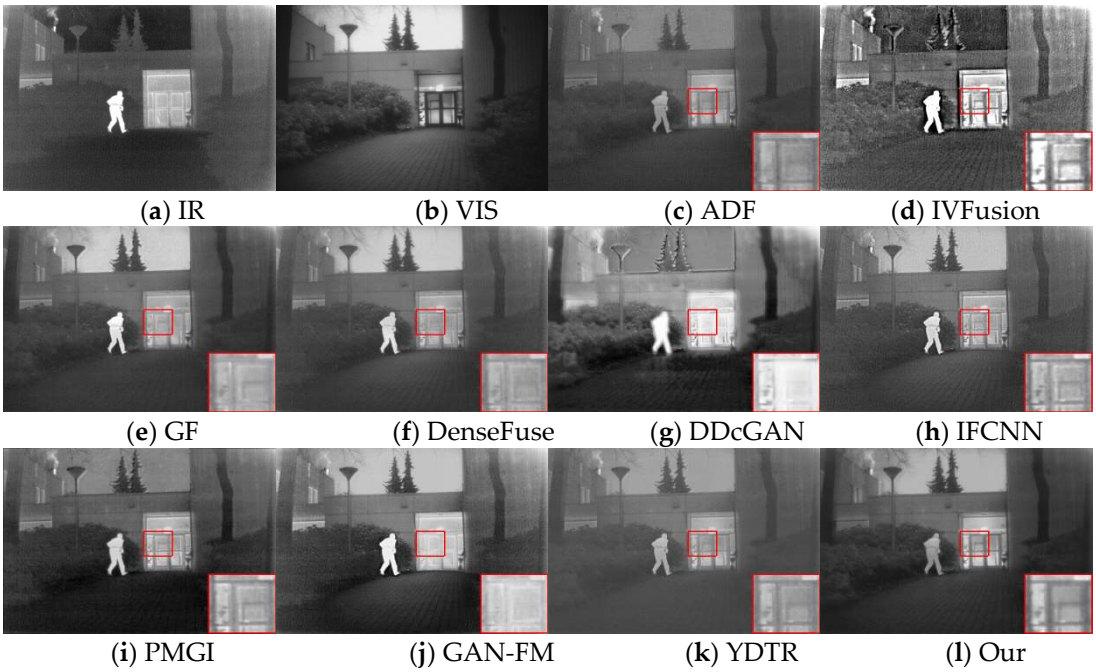


Figure 12. Fusion images of "person walking" for different methods, along with their respective source images.

The second set of source images and their fused image results for different methods are shown in Figure 13. It is obvious that our fusion results are better than the other nine methods from the viewpoint of visual effect, preservation of texture details, and significant target. The methods of IFCNN, PMGI, GAN-FM, and YDTR do not retain enough texture details, the background information of tree branches and railings is blurred, and the intensity information of infrared is too low, resulting in low contrast between light and dark in the fused image. Although the ADF and GF methods retain relatively clear background information, it can be observed from the areas with rich texture details in the red box that the fusion effect is still our method.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2023 doi:10.20944/preprints202306.2220.v1

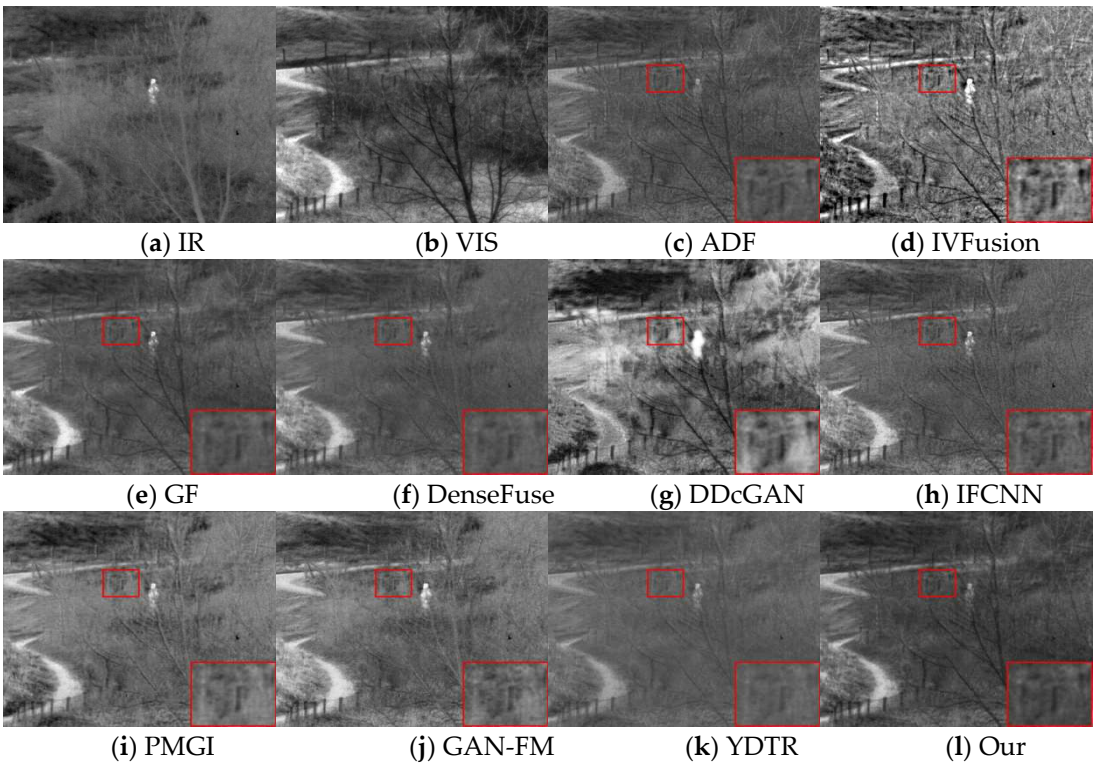
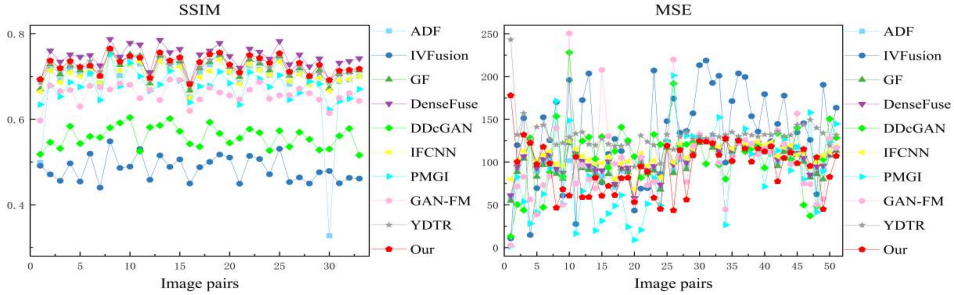


Figure 13. Fusion images of "forest trail" for different methods, along with their respective source images.

4.5.2. Objective Results

We selected 51 image pairs from the TNO dataset for quantitative evaluation. Figure 14 displays the performance metrics for each fusion result, while Table 3 shows the average performance metric values for these fusion methods. For each metric, the best and second best fusion results for all methods are marked in bold and italics, respectively. As can be seen in several figures, our method has significant advantages on SSIM, MSE, CC, and Q_{CB} on the TNO dataset. This phenomenon implies that our fused images have the best visual effect and contain rich texture information and infrared salient target information. In addition, our method ranks second in PSNR and SCD, which indicates that our method transfers enough source image information to the fused images. In summary, a large number of qualitative and quantitative results on the TNO dataset show that our method has outstanding generalization and stability, and is able to retain sufficient texture details and intensity information. We boil this advantage down to several aspects. On the one hand, we design an asymmetric network for the different modal characteristics of the infrared and visible images, preserving thermal radiation information of infrared images and texture details of visible images, respectively. On the other hand, we embed the transformer into the CNN network, which allows the network to preserve the global and local features to the maximum extent.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2023 doi:10.20944/preprints202306.2220.v1

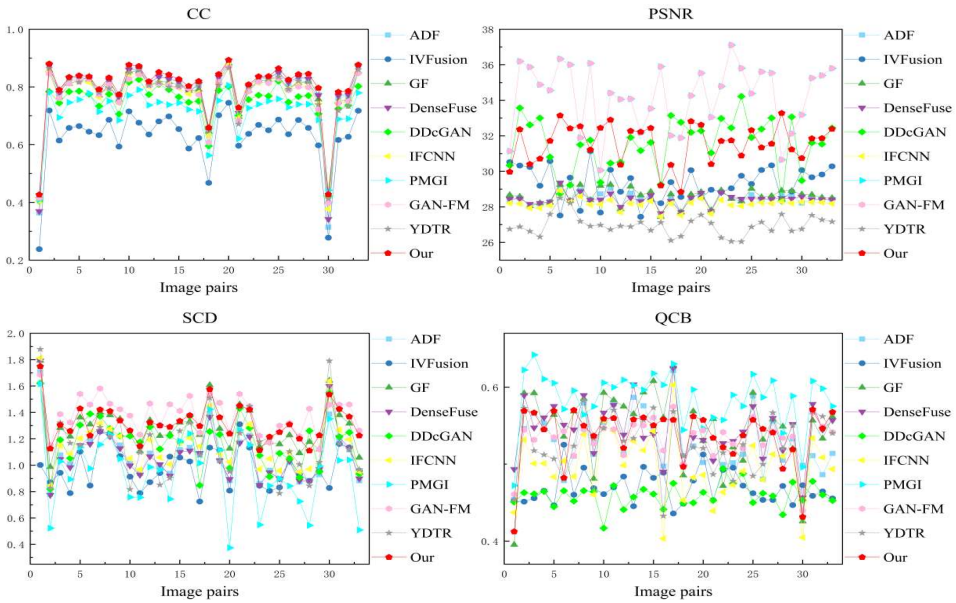


Figure 14. Graphs of performance metrics for different fusion methods.

Table 3. Objective results of the generalization experiment TNO datasets.

Methods	Quality Metrics					
	SSIM	**MSE**	**CC**	**PSNR**	**SCD**	**Q_{CB}**
ADF	0.7085	101.6540	*0.5407*	28.1317	1.4522	0.4888
IVFusion	0.5224	133.1557	0.4016	27.4738	1.0104	0.4795
GF	0.7529	99.9707	0.5368	28.2123	**1.5735**	*0.4891*
DenseFuse	*0.7547*	104.4317	0.5068	28.0037	1.4427	0.4774
DDcGAN	0.5785	107.4098	0.5120	28.2019	1.4089	0.4497
IFCNN	0.7230	108.8827	0.5246	27.7912	1.5154	0.4804
PMGI	0.7067	*87.2343*	0.5363	**29.7962**	1.4668	0.4823
GAN-FM	0.6763	100.7944	0.5049	28.3592	1.5058	0.4540
YDTR	0.7443	131.2808	0.5213	26.9846	1.4979	0.4478
Our	**0.7587**	**80.6446**	**0.5452**	*29.3554*	*1.5180*	**0.5000**

4.6. Detecting Performance

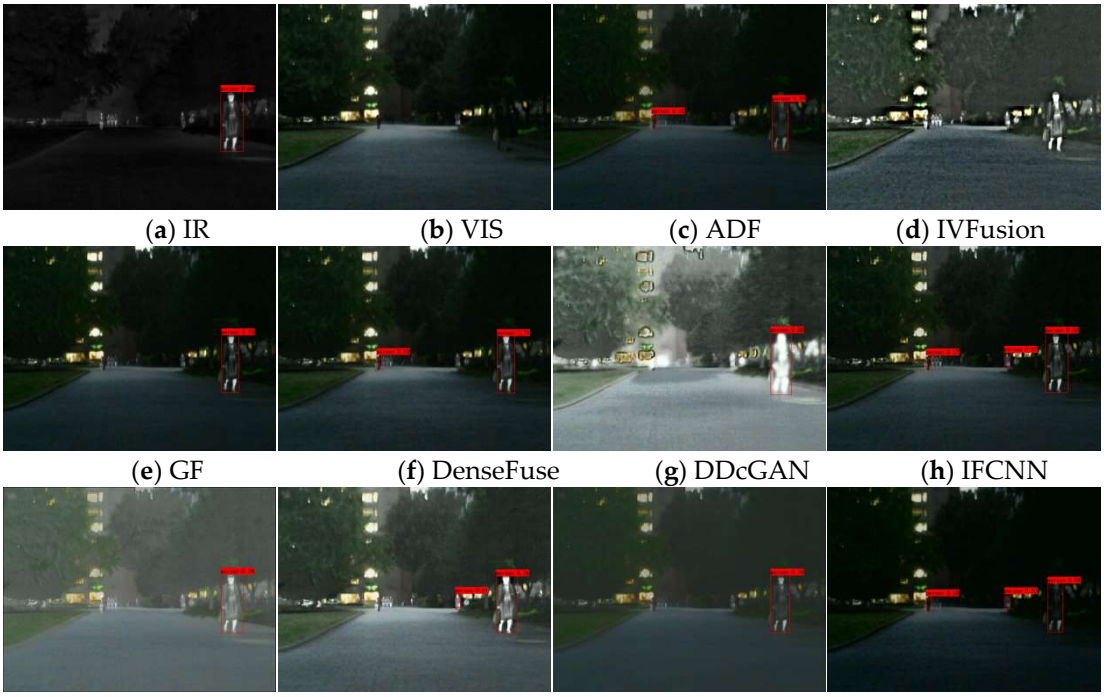
Target detection is an important research direction in the field of computer vision, and its performance reflects well the semantic information integrated into the fused images. To be able to better evaluate the target detection performance of fused images, we use the YOLOX detector [53] for detection. We conducted experiments on 50 randomly selected pairs of images from the MFNet dataset, including 25 pairs of nighttime images and 25 pairs of daytime images, which almost describe the urban scenes.

4.6.1. Subjective Results

Figure 15, Figure 16 and Figure 17 show some typical source images and the detection results of different methods. From the visualization results, we can find that visible images contain rich background information but are difficult to detect salient targets, while infrared images can provide sufficient semantic information about salient targets (e.g., people) and the target has high contrast with the background, which is more helpful for detectors to detect salient targets. Different fusion algorithms can integrate the complementary information of these two images, however, the

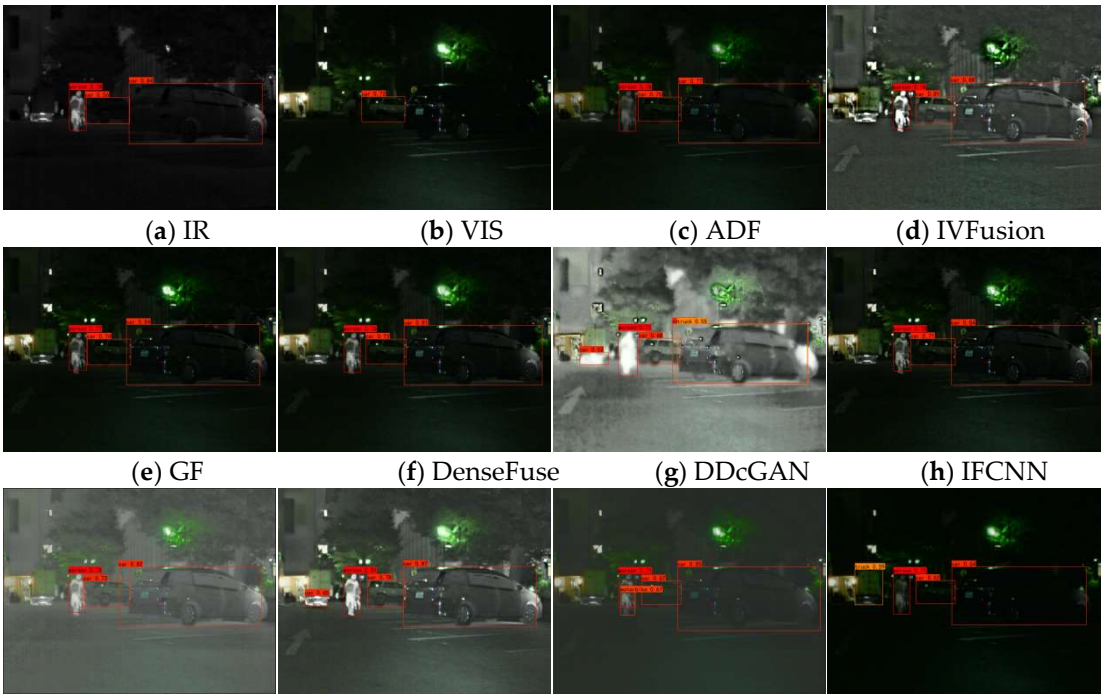
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2023 doi:10.20944/preprints202306.2220.v1

performance of fusion and detection differs due to the difference in methods. For example, in the 00004N scenario, the methods of IVFusion, GF, DDcGAN, PMGI, and YDTR detect only one person, and ADF, DenseFuse, and GAN-FM detect two people, while our method and IFCNN detect three people with higher confidence level than IFCNN. A similar scenario occurs in the 00726N scene, where our method accurately detects people, cars, and also trucks, but trucks cannot be detected by the source image as well as other methods. This shows that our method fully integrates the intensity information of infrared images and the texture information of visible, and is suitable for subsequent image applications.



(**a**) IR (**b**) VIS (**c**) ADF (**d**) IVFusion
(**e**) GF (**f**) DenseFuse (**g**) DDcGAN (**h**) IFCNN
(**i**) PMGI (**j**) GAN-FM (**k**) YDTR (**l**) Our

Figure 15. Subjective results of target detection on scene 00004N.



(**a**) IR (**b**) VIS (**c**) ADF (**d**) IVFusion
(**e**) GF (**f**) DenseFuse (**g**) DDcGAN (**h**) IFCNN
(**i**) PMGI (**j**) GAN-FM (**k**) YDTR (**l**) Our

Figure 16. Subjective results of target detection on scene 00726N.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2023 doi:10.20944/preprints202306.2220.v1

In daytime images, the confidence level of visible images for detecting pedestrians is lower than that of infrared images due to daytime illumination factors, and even some pedestrian targets may not be detected. In the 00420D scene, DDcGAN cannot keep the sharpened edges of pedestrians and other objects, resulting in low confidence of both targets. Due to the interference of useless information, ADF, IVFusion, IFCNN, and YDTR have some targets undetected. The poor fusion of GF, GAN-FM, and PMGI leads to lower confidence in detecting pedestrians and other objects than the source image. In contrast, our method and DenseFuse fully integrate the semantic information in the source images, preserving the source image targets and details. Compared with others, our fused images can detect all targets with confidence levels closer to the source image for all detected targets, which demonstrates the advantage of our method in facilitating advanced vision tasks.

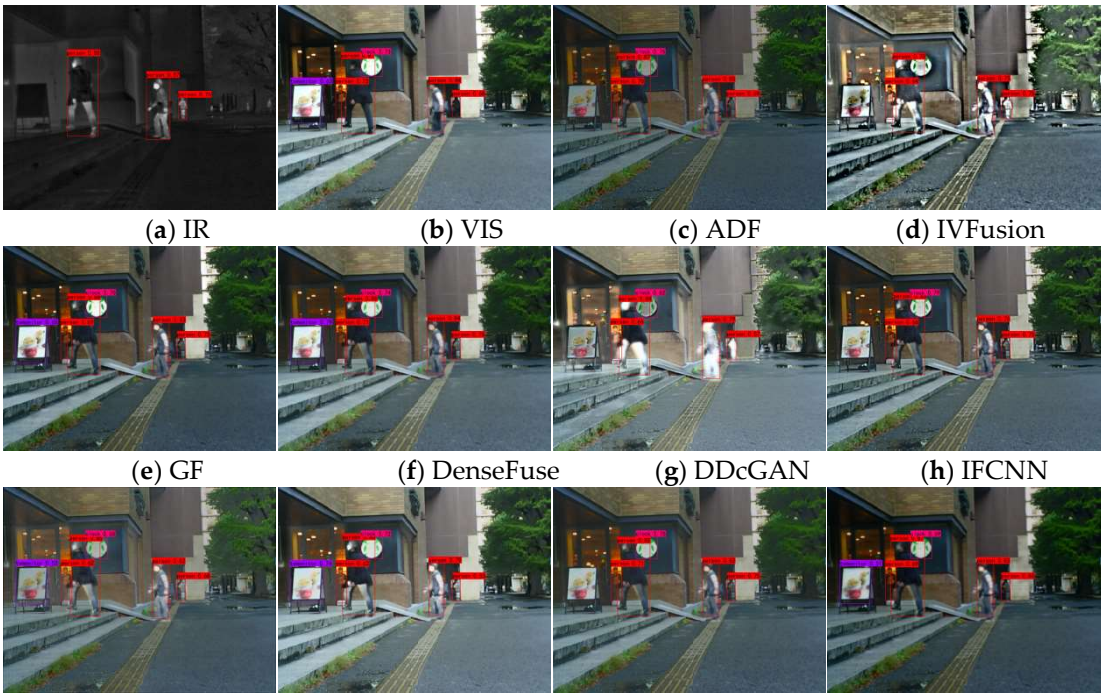


Figure 17. Subjective results of target detection on scene 00420D.

4.6.2. Objective Results

To further evaluate the performance of different methods for the detection task, we use the mean evaluation precision (mAP) for quantitative evaluation. The mAP has a value between 0 and 1, the closer to 1, the better in the model. mAP@0.5 and mAP@0.9 denotes the mAP values at confidence thresholds of 0.5 and 0.9, respectively. The results are shown in Table few and it can be seen that our method performs better under both thresholds. Especially in terms of mAP@0.9, the fused images of our method have a clear advantage and rank first in terms of average accuracy. In terms of mAP@0.5, GAN-FM performs the best and our method is second, while GAN-FM performs poorly in the other threshold. This further indicates the excellent stability of our method. Based on the above subjective and objective analysis, we conclude that the fused images of the proposed method can perform well in the image fusion task and also help improve the performance of target detection.

Table 4. Detection results of source images and different fusion methods.

Methods	mAP@0.5			mAP@0.9		
	Person	Car	Average	Person	Car	Average
IR	0.6307	0.3023	0.4665	**0.2562**	0.3013	0.2788
VIS	0.4953	0.7240	0.6096	0.1901	0.4358	0.3129
ADF	0.6935	0.7208	0.7072	0.2456	0.4505	0.3480

 doi:10.20944/preprints202306.2220.v1

IVFusion	*0.7288*	0.7040	0.7164	0.1768	0.3733	0.2750
GF	0.6562	0.7300	0.6931	0.2415	0.4603	0.3509
DenseFuse	0.6915	0.7353	0.7134	0.2413	0.4425	0.3419
DDcGAN	0.4010	0.6968	0.5489	0.1072	0.3550	0.2316
IFCNN	0.7038	0.7305	0.7172	*0.2541*	0.4108	0.3324
PMGI	0.6990	0.6788	0.6889	0.2238	0.3448	0.2843
GAN-FM	**0.7450**	**0.7548**	**0.7499**	0.2409	0.4178	0.3293
YDTR	0.7149	0.5708	0.6428	0.2348	**0.4972**	*0.3660*
Our	0.7241	*0.7388*	*0.7223*	0.2458	*0.4865*	**0.3661**

5. Conclusions

In this paper, we propose a new end-to-end network to solve the problem of infrared and visible image fusion. For the characteristics of two different modal images, we design dual-stream asymmetric branched paths to extract infrared and visible image features, use Transformer to capture global information and long-distance dependencies in infrared images, and use residual dense blocks to fully extract texture details in visible images. Finally, the captured features are fully merged by the main path to further retain important information. This approach enables the preservation of both texture details from visible images and thermal radiation targets from infrared images in a superior manner. We conducted a large number of comparison experiments and generalization experiments testing using RoadScene datasets and TNO datasets. The experimental results reveal that our approach outperforms existing techniques in both subjective and objective evaluations, demonstrating its outstanding performance and generalization ability. And target detection experiments were carried out on the MFNet datasets to showcase the prowess of our approach in elevating high-level visual tasks.

Author Contributions: Conceptualization, R.Y. and B.Y.; methodology, R.Y.; software, Z.H.; validation, R.Y. and B.Y.; writing—original draft preparation, R.Y.; writing—review and editing, B.Y. and X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Nos.61871210) and Chuanshan Talent Project of the University of South China.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Source of data set in experimental analysis: https://fifigshare.com/articles/TNOImageFusionDataset/1008029, https://github.com/hanna-xu/RoadScene, and https://github.com/haqishen/MFNet-pytorch.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yang, Z.; Yu, W.; Liang, P.; Guo, H.; Xia, L.; Zhang, F.; Ma, Y.; Ma, J. Deep Transfer Learning for Military Object Recognition under Small Training Set Condition. *Neural Comput & Applic* **2019**, *31*, 6469–6478, doi:10.1007/s00521-018-3468-3.
2. Ma, J.; Ma, Y.; Li, C. Infrared and Visible Image Fusion Methods and Applications: A Survey. *Inf. Fusion* **2019**, *45*, 153–178, doi:10.1016/j.inffus.2018.02.004.
3. Schnelle, S.R.; Chan, A.L. Enhanced Target Tracking through Infrared-Visible Image Fusion. In Proceedings of the 14th International Conference on Information Fusion; 2011; pp. 1–8.
4. Ma, W.; Wang, K.; Li, J.; Yang, S.X.; Li, J.; Song, L.; Li, Q. Infrared and Visible Image Fusion Technology and Application: A Review. *Sensors* **2023**, *23*, 599, doi:10.3390/s23020599.
5. Zhang, L.; Yang, X.; Wan, Z.; Cao, D.; Lin, Y. A Real-Time FPGA Implementation of Infrared and Visible Image Fusion Using Guided Filter and Saliency Detection. *Sensors* **2022**, *22*, 8487, doi:10.3390/s22218487.
6. Jia, W.; Song, Z.; Li, Z. Multi-Scale Fusion of Stretched Infrared and Visible Images. *Sensors* **2022**, *22*, 6660, doi:10.3390/s22176660.

7. Liu, Y.; Wu, Z.; Han, X.; Sun, Q.; Zhao, J.; Liu, J. Infrared and Visible Image Fusion Based on Visual Saliency Map and Image Contrast Enhancement. *Sensors* **2022**, *22*, 6390, doi:10.3390/s22176390.
8. Huang, Z.; Yang, B.; Liu, C. RDCa-Net: Residual Dense Channel Attention Symmetric Network for Infrared and Visible Image Fusion. *Infrared Phys. Technol.* 2023, 130, 104589, doi:10.1016/j.infrared.2023.104589.
9. Wang, H.; Wang, J.; Xu, H.; Sun, Y.; Yu, Z. DRSNFuse: Deep Residual Shrinkage Network for Infrared and Visible Image Fusion. *Sensors* **2022**, *22*, 5149, doi:10.3390/s22145149.
10. Zheng, X.; Yang, Q.; Si, P.; Wu, Q. A Multi-Stage Visible and Infrared Image Fusion Network Based on Attention Mechanism. *Sensors* **2022**, *22*, 3651, doi:10.3390/s22103651.
11. Liu, Y.; Wang, Z. Simultaneous Image Fusion and Denoising with Adaptive Sparse Representation. *IET Image Processing* **2015**, *9*, 347–357, doi:10.1049/iet-ipr.2014.0311.
12. Yang, B.; Li, S. Visual Attention Guided Image Fusion with Sparse Representation. *Optik* **2014**, *125*, 4881–4888, doi:10.1016/j.ijleo.2014.04.036.
13. Bulanon, D.M.; Burks, T.F.; Alchanatis, V. Image Fusion of Visible and Thermal Images for Fruit Detection. *Biosyst. Eng.* **2009**, *103*, 12–22, doi:10.1016/j.biosystemseng.2009.02.009.
14. Yu, X.; Ren, J.; Chen, Q.; Sui, X. A False Color Image Fusion Method Based on Multi-Resolution Color Transfer in Normalization YCbCr Space. *Optik* **2014**, *125*, 6010–6016, doi:10.1016/j.ijleo.2014.07.059.
15. Cvejic, N.; Bull, D.; Canagarajah, N. Region-Based Multimodal Image Fusion Using ICA Bases. *IEEE Sens. J.* **2007**, *7*, 743–751, doi:10.1109/JSEN.2007.894926.
16. Mitianoudis, N.; Stathaki, T. Pixel-Based and Region-Based Image Fusion Schemes Using ICA Bases. *Inf. Fusion* **2007**, *8*, 131–142, doi:10.1016/j.inffus.2005.09.001.
17. Yin, M.; Duan, P.; Liu, W.; Liang, X. A Novel Infrared and Visible Image Fusion Algorithm Based on Shift-Invariant Dual-Tree Complex Shearlet Transform and Sparse Representation. *Neurocomputing* **2017**, *226*, 182–191, doi:10.1016/j.neucom.2016.11.051.
18. Deng, J.; Xuan, X.; Wang, W.; Li, Z.; Yao, H.; Wang, Z. A Review of Research on Object Detection Based on Deep Learning. *J. Phys.: Conf. Ser.* **2020**, *1684*, 012028, doi:10.1088/1742-6596/1684/1/012028.
19. Tian, C.; Fei, L.; Zheng, W.; Xu, Y.; Zuo, W.; Lin, C.-W. Deep Learning on Image Denoising: An Overview. *Neural Netw.* **2020**, *131*, 251–275, doi:10.1016/j.neunet.2020.07.025.
20. Liu, Y.; Chen, X.; Peng, H.; Wang, Z. Multi-Focus Image Fusion with a Deep Convolutional Neural Network. *Inf. Fusion* **2017**, *36*, 191–207, doi:10.1016/j.inffus.2016.12.001.
21. Liu, Y.; Chen, X.; Cheng, J.; Peng, H.; Wang, Z. Infrared and Visible Image Fusion with Convolutional Neural Networks. *Int. J. Wavelets Multiresolut Inf. Process.* **2018**, *16*, 1850018, doi:10.1142/S0219691318500182.
22. Li, H.; Wu, X.-J. DenseFuse: A Fusion Approach to Infrared and Visible Images. *IEEE Trans. on Image Process.* **2019**, *28*, 2614–2623, doi:10.1109/TIP.2018.2887342.
23. Ma, J.; Yu, W.; Liang, P.; Li, C.; Jiang, J. FusionGAN: A Generative Adversarial Network for Infrared and Visible Image Fusion. *Inf. Fusion* **2019**, *48*, 11–26, doi:10.1016/j.inffus.2018.09.004.
24. Ma, J.; Xu, H.; Jiang, J.; Mei, X.; Zhang, X.-P. DDcGAN: A Dual-Discriminator Conditional Generative Adversarial Network for Multi-Resolution Image Fusion. *IEEE Trans. on Image Process.* **2020**, *29*, 4980–4995, doi:10.1109/TIP.2020.2977573.
25. Zhang, H.; Xu, H.; Xiao, Y.; Guo, X.; Ma, J. Rethinking the Image Fusion: A Fast Unified Image Fusion Network Based on Proportional Maintenance of Gradient and Intensity. In Proceedings of the AAAI Conference on Artificial Intelligence, 34, 12797–12804, doi:10.1609/aaai.v34i07.6975.
26. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A Unified Unsupervised Image Fusion Network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 502–518, doi:10.1109/TPAMI.2020.3012548.
27. Tang, W.; He, F.; Liu, Y. YDTR: Infrared and Visible Image Fusion via Y-Shape Dynamic Transformer. *IEEE Trans. Multimedia* **2022**, 1–16, doi:10.1109/TMM.2022.3192661.
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; Vol. 30.
29. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding 2019.
30. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Proceedings of the Advances in Neural Information Processing Systems; Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d', Fox, E., Garnett, R., Eds.; Curran Associates, Inc., 2019; Vol. 32.
31. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale 2021.

32. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation 2021.
33. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions 2021.
34. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S.; et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Nashville, TN, USA, June 2021; pp. 6877–6886.
35. Yan, B.; Peng, H.; Fu, J.; Wang, D.; Lu, H. Learning Spatio-Temporal Transformer for Visual Tracking. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); IEEE: Montreal, QC, Canada, October 2021; pp. 10428–10437.
36. Ren, P.; Li, C.; Wang, G.; Xiao, Y.; Du, Q.; Liang, X.; Chang, X. Beyond Fixation: Dynamic Window Visual Transformer. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: New Orleans, LA, USA, June 2022; pp. 11977–11987.
37. Vs, V.; Jose Valanarasu, J.M.; Oza, P.; Patel, V.M. Image Fusion Transformer. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP); IEEE: Bordeaux, France, October 16 2022; pp. 3566–3570.
38. Zhao, H.; Nie, R. DNDT: Infrared and Visible Image Fusion Via DenseNet and Dual-Transformer. In Proceedings of the 2021 International Conference on Information Technology and Biomedical Engineering (ICITBE); IEEE: Nanchang, China, December 2021; pp. 71–75.
39. Fu, Y.; Xu, T.; Wu, X.; Kittler, J. PPT Fusion: Pyramid Patch Transformerfor a Case Study in Image Fusion 2022.
40. Rao, D.; Xu, T.; Wu, X.-J. TGFuse: An Infrared and Visible Image Fusion Approach Based on Transformer and Generative Adversarial Network. *IEEE Trans. on Image Process.* **2023**, 1–1, doi:10.1109/TIP.2023.3273451.
41. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual Dense Network for Image Super-Resolution 2018.
42. Ma, J.; Zhou, Y. Infrared and Visible Image Fusion via Gradientlet Filter. *Computer Vision and Image Understanding* **2020**, 197–198, 103016, doi:10.1016/j.cviu.2020.103016.
43. Bavirisetti, D.P.; Dhuli, R. Fusion of Infrared and Visible Sensor Images Based on Anisotropic Diffusion and Karhunen-Loeve Transform. *IEEE Sens. J.* **2016**, *16*, 203–209, doi:10.1109/JSEN.2015.2478655.
44. Li, G.; Lin, Y.; Qu, X. An Infrared and Visible Image Fusion Method Based on Multi-Scale Transformation and Norm Optimization. *Inf. Fusion* **2021**, *71*, 109–129, doi:10.1016/j.inffus.2021.02.008.
45. Zhang, H.; Yuan, J.; Tian, X.; Ma, J. GAN-FM: Infrared and Visible Image Fusion Using GAN With Full-Scale Skip Connection and Dual Markovian Discriminators. *IEEE Trans. Comput. Imaging* **2021**, *7*, 1134–1147, doi:10.1109/TCI.2021.3119954.
46. Zhang, Y.; Liu, Y.; Sun, P.; Yan, H.; Zhao, X.; Zhang, L. IFCNN: A General Image Fusion Framework Based on Convolutional Neural Network. *Inf. Fusion* **2020**, *54*, 99–118, doi:10.1016/j.inffus.2019.07.011.
47. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. on Image Process.* **2004**, *13*, 600–612, doi:10.1109/TIP.2003.819861.
48. Aslantas, V.; Kurban, R. A Comparison of Criterion Functions for Fusion of Multi-Focus Noisy Images. *Opt. Commun.* **2009**, *282*, 3231–3242, doi:10.1016/j.optcom.2009.05.021.
49. Mukaka, M.M. A Guide to Appropriate Use of Correlation Coefficient in Medical Research. *Malawi Med. J.* **2012**, *24*, 69–71.
50. Rajkumar, S.; Mouli, P.V.S.S.R.C. Infrared and Visible Image Fusion Using Entropy and Neuro-Fuzzy Concepts. In ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India- Vol I; Satapathy, S.C., Avadhani, P.S., Udgata, S.K., Lakshminarayana, S., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, 2014; Vol. 248, pp. 93–100 ISBN 978-3-319-03106-4.
51. Aslantas, V.; Bendes, E. A New Image Quality Metric for Image Fusion: The Sum of the Correlations of Differences. *AEU - International Journal of Electronics and Communications* **2015**, *69*, 1890–1896, doi:10.1016/j.aeue.2015.09.004.
52. Chen, Y.; Blum, R.S. A New Automated Quality Assessment Algorithm for Image Fusion. *Image and Vis. Comput.* **2009**, *27*, 1421–1432, doi:10.1016/j.imavis.2007.12.002.
53. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021 2021.