This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2023.3273451

TGFuse: An Infrared and Visible Image Fusion Approach Based on Transformer and Generative Adversarial Network

Dongyu Rao, Tianyang Xu, and Xiao-Jun Wu*

***Abstract*—The end-to-end image fusion framework has achieved promising performance, with dedicated convolutional networks aggregating the multi-modal local appearance. However, long-range dependencies are directly neglected in existing CNN fusion approaches, impeding balancing the entire image-level perception for complex scenario fusion. In this paper, therefore, we propose an infrared and visible image fusion algorithm based on the transformer module and adversarial learning. Inspired by the global interaction power, we use the transformer technique to learn the effective global fusion relations. In particular, shallow features extracted by CNN are interacted in the proposed transformer fusion module to refine the fusion relationship within the spatial scope and across channels simultaneously. Besides, adversarial learning is designed in the training process to improve the output discrimination via imposing competitive consistency from the inputs, reflecting the specific characteristics in infrared and visible images. The experimental performance demonstrates the effectiveness of the proposed modules, with superior improvement against the state-of-the-art, generalising a novel paradigm via transformer and adversarial learning in the fusion task.**

***Index Terms*—visual object tracking, RGBT tracking, temporal information, decision-level fusion,**

I. Introduction

With the development of imaging equipment and analysis approaches, multi-modal visual data is emerging rapidly with many practical applications. In general, image fusion has played an important role in helping human vision to perceive information association between multi-modal data. Among them, the fusion of infrared and visible images has important applications in military, security, and visual tracking [1], [2], [3] etc., becoming an important part of image fusion tasks.

In order to design a natural and efficient image fusion algorithm, researchers have developed many fusion algorithms on the basis of traditional image processing. Firstly, the fusion algorithms based on multi-scale transformation are proposed [4], [5], which applied traditional image processing methods to image fusion. In this way, a multi-scale representation of the source image can be extracted. The fused image can be obtained by fusing and restoring the multi-scale representation of the source image. Subsequently, fusion algorithms based on sparse / low-rank representation were applied [6], [7]. This method decomposes and represents the source image through an overcomplete dictionary that has been learned. The representation coefficients of the source image are fused and then restored to obtain a fused image. These algorithms use specific image processing methods to obtain image representations, and obtain the output images by fusing the image representations. Some progress has been made in image fusion methods based on traditional image processing techniques. Limited by the ability of representation, these methods are not competent for the fusion of complex scenes. In addition, the complex transformation of these methods may introduce certain noise, and the low operating efficiency also affects their application. With the development of deep learning, convolutional neural network has become the mainstream of research due to its strong representation ability and flexible structure [8], [9]. However, since most image fusion tasks are unsupervised, the supervised end-to-end training framework is not suitable for training fusion tasks. Drawing on this, some fusion algorithms [10], [11] based on large-scale pre-trained networks are proposed. This is a preliminary combination of neural network and image fusion tasks. Subsequently, Li et al. [12], [13] proposed a fusion algorithm based on an auto-encoder network, using ordinary data sets for encoder-decoder training. This approach greatly expands the flexibility of the fusion framework. In order to obtain better performance for specific fusion tasks, the end-to-end image fusion methods [14], [15], [16] are proposed to learn more targeted network parameters through a specific network structure and loss function. This method is dedicated to training fusion tasks, which can usually achieve better fusion results. This puts forward higher requirements for the representative ability of the network and the effectiveness of the fusion method. At present, the end-

Fig. 1. Infrared image (a), visible image (b) and fused image generated by the proposed method (c).

D. Rao, T. Xu and X.-J. Wu (Corresponding Author) are with the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, 214122, P.R. China. (e-mail: raodongyu@163.com; {tianyang.xu; wu_xiaojun}@jiangnan.edu.cn)

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2023.3273451

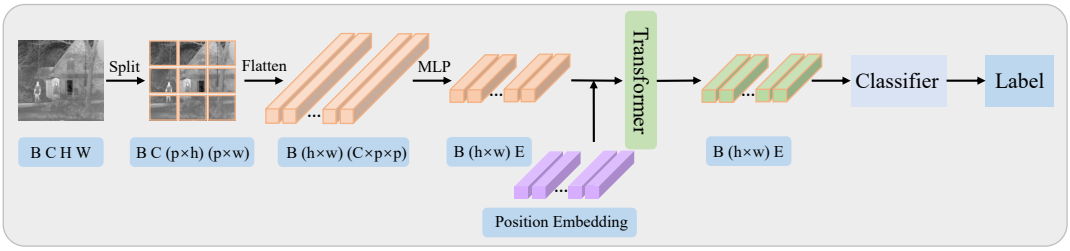


Fig. 2. The framework of ViT (Vision Transformer). "B C H W" respectively represent the batch size, channels, height and width. "p" means patch size. "h w" is the number of patches in height and width. "E" is the reduced dimension.

to-end fusion algorithm mainly uses a convolutional neural network for feature extraction and achieves the fusion effect. Subsequently, more complex networks and more ingenious fusion methods are developed. Typically, transformer-based image fusion methods are representative techniques, delivering excellent performance with their complex structure and precise design [17], [18], [19], [20]. Besides, these works not only focus on the improvement of aligned fusion effects, but also focus on the fusion of unregistered image pairs [21], [22] and the fusion of downstream tasks [23], [24]. Such multi-task setting inspires further development for image fusion in vision tasks. The fusion of unregistered images can greatly improve the application range of multi-source image fusion. It can also better help the downstream tasks (such as segmentation, tracking, etc.). However, due to the characteristics of CNN, this process usually ignores the global dependency infusion. We believe that the global dependence in image fusion should include the correlation between different modes as well as the global interactions within the image itself. Some methods using transformers can better understand the global relationship of images, but the correlation information between image modes is not obtained.

In order to solve the problem of global dependence and effective integration, we propose an infrared and visible image fusion algorithm based on the transformer and adversarial learning. Our method uses a general visual transformer for image spatial relationship learning. In particular, we propose a novel cross-channel transformer model to learn the channel relationship. The composite transformer fusion module has learned the global fusion relationship with space and channels. In addition, adversarial learning is introduced in the training process. We use two discriminators (infrared and fused image, visible and fused image) for adversarial training respectively. This allows the fused image to obtain higher-quality infrared and visible image characteristics. Unlike existing transformer-based fusion methods [25], [26], we design two kinds of transformer structures (space and channel) which can pay attention to the global information of the image and the relevance between different modes. This can provide more comprehensive guidance for image fusion. In addition, we explore feature-level discrimination, different from traditional image judgment, to provide perceptual improvement for fused images in the design of the discriminator network.

The proposed method mainly has the following three innovations

- A channel-token transformer is proposed to explore the channel relationships, which is effectively applied in the fusion method.
- A transformer module is designed to achieve global fusion relationship learning in complex scenarios.
- Adversarial learning is introduced into the training process. The discriminator of the two modalities introduces the characteristics of different modalities to the fused image to improve the fusion effect.

II. Related Work

A. Image Fusion Method Based on Deep Learning

The fusion algorithm based on deep learning has shown excellent performance in infrared and visible image fusion, multi-focus image fusion and medical image fusion, etc. Li et al. [10], [11] used a pre-trained neural network in image fusion tasks. The features extracted by the pre-trained network are processed to assign fusion weights to participate in the fusion process. Such an attempt has achieved a good improvement in terms of fusion effect. But deep features obtained by pre-trained classification networks are not fully suitable for fusion tasks. In order to obtain the depth features suitable for reconstructing images, Li et al. [12] first proposed an algorithm based on an auto-encoder network. The method first trains an auto-encoder network that implements image reconstruction. While, there is no fusion module involved in the training process. In this way, features delivered by the encoder are more suitable for fusion tasks. At the same time, the algorithm can also achieve a good fusion effect in the absence of specific data. However, the hand-crafted designed fusion rules bring more possible interference factors to the fusion process. Furthermore, a model trained without using the specific data cannot achieve the ideal performance on the typical domain. With the advancement of visual data collection equipment, some large-scale multi-mode data sets have appeared, so end-to-end fusion algorithms [27], [28] have received more attention and applications. The end-to-end image fusion algorithm integrates the fusion process in the neural network training stage. This integrated training model allows the neural network to learn the features as well as the fusion approach in a unified manner. This end-to-end fusion algorithm based on convolutional neural networks achieves better performance on a single task. But it still has some limitations, such as the spatial limitation of the fusion method

based on a convolutional neural network. In order to explore the global information of the image, some researchers have proposed the fusion based on transformer [25], [26], [29]. The transformer architecture is derived from natural language processing tasks, presenting superior performance on vision tasks. However, how to preserve its characteristics to obtain promising fusion results is an unsolved issue. In this paper, the proposed method is an end-to-end image fusion algorithm. But compared to the CNN-based fusion network, we expand the network structure of the end-to-end algorithm and introduce the transformer that focuses on building global relationships into the fusion module. Unlike other methods, our approach provides a more comprehensive fusion relationship learning and the whole fusion process is built on the image information obtained by the two kinds of transformers. This is a new exploration of fusion methods.

B. Generative Adversarial Network

A generative adversarial network (GAN) is an algorithm that obtains high-quality generated images by training two networks against each other. Goodfellow et al. [30] first proposed the idea of a generative adversarial network. The generator generates an image, and the discriminator determines whether the input image is a real image (True) or a generated image (False). Subsequently, many improvements based on the original GAN focused on speeding up the training of the network and improving the quality of the generated images [31], [32], [33]. These improvements also help GAN gain a wider range of applications [34], [35], [36]. Methods based on GAN are also widely used in image generation tasks [37], [38]. Recently, generative adversarial networks have been introduced to perform image fusion tasks. The FusionGAN [14] method is proposed for infrared and visible image fusion. In this method, the source images are concatenated and directly input into the generator to obtain a preliminary fusion image. Then, the fused image and the infrared image are input into the discriminator for judgment. Through the adversarial training of the cyclic generation and discrimination process, the infrared information in the fused image is effectively strengthened. However, the disadvantage of this method is that it overemphasizes the importance of infrared information, which leads to the fusion results being more inclined to infrared images.

Adversarial learning is an important part of our approach. It improves the infrared and visible image characteristics in the fusion result by obtaining competitive consistency from the inputs. However, we abandon the discriminator of the classification mode and use the difference in the feature level to promote the fused image to have more infrared or visible image information.

C. Visual Transformer

The transformer is a model based on a pure attention mechanism [39]. Its success in natural language processing inspires its application in computer vision. Due to the long-range dependence of the transformer in processing input, the visual transformer also has the ability to pay attention to the global relationship in image tasks. As a pioneering work of visual transformer, Dosovitskiy et al. [40] proposed ViT (Vision Transformer) for image classification tasks (Figure.2). This is a simple and effective application of transformer in visual tasks. Subsequently, Chen et al. [41] proposed a multi-task model based on the transformer, which achieved good results on multiple low-level visual tasks. The global spatial dependence of transformers has gained many applications in the field of computer vision. This also inspires the potential of combining the model based on the transformer structure with the image fusion task. At present, an attempt has successfully proposed an image fusion method based on the transformer framework (IFT) [42]. The basic idea is to extract the features of the cascaded source images through a convolutional neural network and divide them into blocks (tokens). Then the designed Transformer module is used to obtain the global relationship in the image, so that the images can complete the fusion process during the training process. This is an interesting and straightforward attempt of the transformer structure in the fusion task.

Inspired by the characteristics of the transformer, we pay attention to the global correlation of images space and channels during the fusion process. We propose a new transformer model that focuses on channel relationships and applies it in the field of image fusion. This is a new exploration of transformer applications.

III. Proposed Method

A. The Framework of Network

As shown in Figure. 3, our model is mainly composed of two parts: one transformer-based generator and two discriminators. Typically, the fused image is obtained by the generator. Then, the output is refined during the adversarial learning between the generator and the discriminator.

Generator. The generator is used for the generation of the fused image. "IR" and "VIS" represent infrared and visible images, respectively. "Convolutional layer" is 3×3 convolution. "Res-Block" is used to extract deep features at different scales. "Relationship Map" is a feature enhancement coefficient obtained from the "Trans Module". $\otimes$ represents the operation of elementwise multiplication.

After the source images are merged in the channel dimension, the initial feature extraction is performed through the convolutional neural network. The mixed CNN features are input to the transformer fusion module to learn global fusion relations. Taking into account the consumption of computing resources and representation of features, three downsampling operators are added before the transformer fusion module. The fusion relationship learned in this process is up-sampled to different scales and multiplied by the corresponding features to achieve the preliminary result. The fusion features of different scales are up-sampled to the original image size and then superimposed to obtain the final fusion result.

Discriminator. The discriminator is used to refine the perception quality of the fused image. We set up two discriminators: fused image and infrared image ("Dis-IR"), fused image and visible image ("Dis-VIS"). These two discriminators provide high-resolution details of the visible image and

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2023.3273451

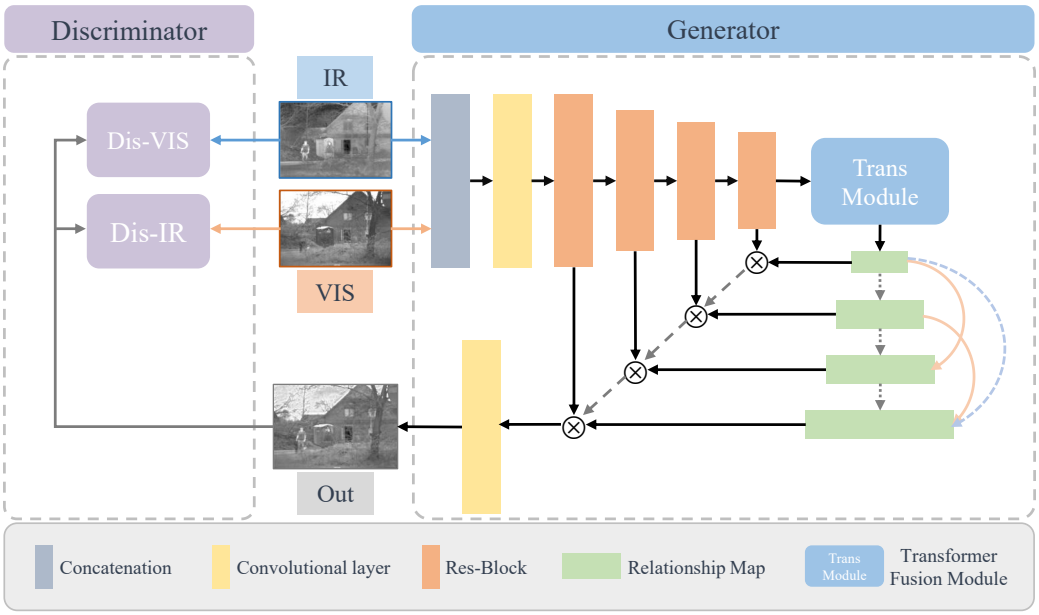


Fig. 3. The framework of our method.

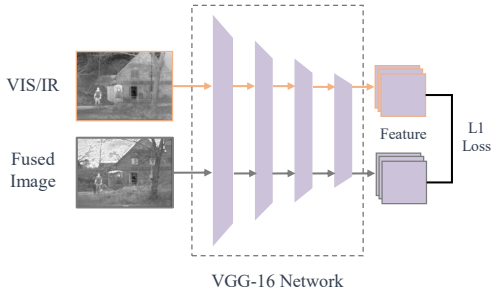


Fig. 4. The framework of discriminator.

a significant part of the infrared image for the fused image. The pre-trained VGG-16 network is used as the discriminator, which can be further fine-tuned during training. The network is shown in Figure.4. Taking the visible image discriminator ("Dis-VIS") as an example, the fused image and the visible image are separately input into the VGG-16 network to extract features. We calculate the L1 loss between the two features so that the fused image approximates the visible image from the context perspective. According to the number of down-sampling, VGG-16 is divided into 4 layers. Different layers have different feature depths and different feature shapes. Inspired by Johnson et al. [43], we use the features of different depths extracted by VGG-16 to distinguish between infrared and visible features. The infrared discriminator uses the features of the fourth layer of VGG-16 to retain more saliency information. While the visible discriminator uses the features of the first layer of VGG-16 to retain more detailed information.

In the training stage, source images are input to the generator to obtain the preliminary fused image. The preliminary fused image then passes through two discriminators with the effect of the fused image being fed back through the loss function. The above two steps are performed alternately to realize the confrontation training between the generator and the discriminator. Finally, we get a generator with an ideal generation effect to achieve the purpose of image fusion.

B. The Transformer Fusion Module

As shown in Figure. 7, the transformer fusion module consists of two parts: general transformer ("spatial transformer") and cross-channel transformer ("channel transformer"). This helps us to obtain a more comprehensive global integration relationship.

Spatial Transformer As shown in Figure. 2, the image is divided into blocks and stretched into vectors, where "p" means patch size, "w" and "h" respectively represent the number of image blocks in the width and height dimensions of the image, "E" is the reduced dimension. Then, the vector group enters the transformer model for relation learning. The number of image blocks is used to learn the global relationship of the image. Therefore, we consider that the general transformer mainly learns the global spatial relationship between image patches. Inspired by the transformer-based low-level image task, we build a spatial transformer for the fusion task. As shown in Figure. 5, the spatial transformer is basically the same as the first half of ViT (Figure. 2). The difference is that we cancelled the addition of position embedding, and subsequent experiments also proved the rationality and effectiveness of this operation. In addition, when restoring from the vector group to the image, we compress the channel dimension, so that we get a relationship map with a channel number of 1. This corresponds to the spatial relationship of the image we obtained, avoiding the interference of other dimensional relationships.

Channel Transformer For image fusion tasks, we believe that the cross-channel relationship of images also plays an

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2023.3273451

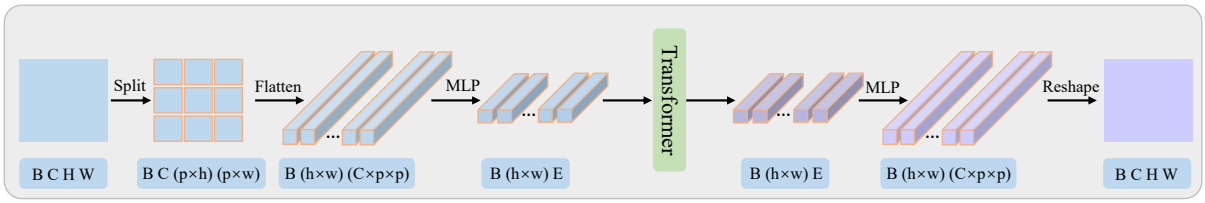


Fig. 5. The framework of spatial transformer.

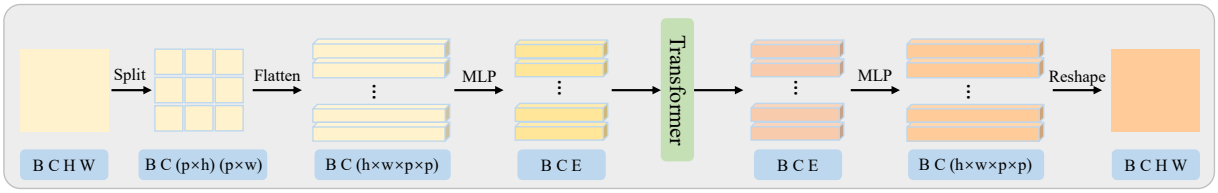


Fig. 6. The framework of channel transformer.

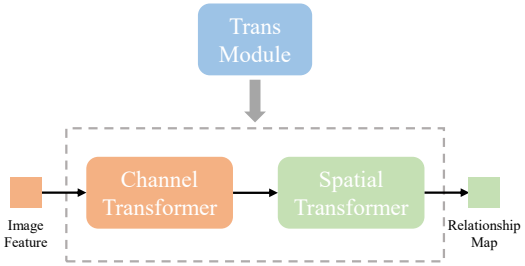


Fig. 7. The framework of transformer fusion module.

important role in fusion. Therefore, we propose a new cross-channel transformer model, which learns the correlation of information across the channel dimension. In the new transformer module, the number of tokens input to the encoder has changed from the number of image blocks to the number of image channels. Since position embedding is not required to provide category information in the image generation task, we have removed position embedding, which also makes the size of the input image more flexible. The channel transformer is also a structure similar to the spatial transformer. The main difference is that we change the object modelled by the transformer from the spatial relationship of the image block to the channel relationship. In this specific implementation, we use the number of channels as the token number, which is a simple but effective operation. Through two kinds of the transformer, we can get the relation mapping for the image fusion task.

Composite Transformer The transformer of the two modes is combined into a transformer fusion module, which enables our fusion model to simultaneously learn spatial and channel relationships with global correlation. Through experiments, we find that using a channel transformer first and then using a spatial transformer can achieve better results. This shows that the combination of these two fusion modules is used to learn the coefficients that are more suitable for the fusion of infrared and visible images.

C. Loss Function

Generator Loss Function Previous image fusion algorithms based on deep learning usually use multiple loss functions to optimize the fused image from different perspectives during training. But this causes mutual conflict among loss functions. Inspired by [44], we make improvements on the basis of the SSIM loss. A single loss function achieves a good fusion effect and avoids the problem of entanglement of multiple loss functions.

SSIM [45] is a measure of structural similarity between images. As shown in Eq. (1), X, Y represent two images respectively. μ and σ stand for mean and standard deviation respectively. σ_{XY} means the covariance between X and Y. C_1 and C_2 are stability coefficients.

$$SSIM(X,Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X{}^2 + \mu_Y{}^2 + C_1)(\sigma_X{}^2 + \sigma_Y{}^2 + C_2)} \quad (1)$$

Variance reflects the contrast of the image, and an image with high contrast is more helpful for the human visual system to capture information. As shown in Eq. (2), M and N are the image size in the horizontal and vertical directions respectively. μ represents the mean of the image. We use variance as the standard and choose one as the reference image from infrared and visible images. The structural similarity between the fused image and the reference image is calculated, so that the fused image gradually approaches the reference image during the optimization process. This operation allows the fusion result to better obtain the important information from the infrared or visible image.

$$\sigma^2(X) = \frac{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}[X(i,j) - \mu]^2}{MN} \quad (2)$$

In Eq.(3), Var_SSIM calculates the structural similarity of the divided image. σ^2 is the variance of the image. I_X and I_Y represent two source images respectively. I_F means a fused image. W is the number of image blocks after division, and the size of each image block is set to 11×11. Image

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2023.3273451

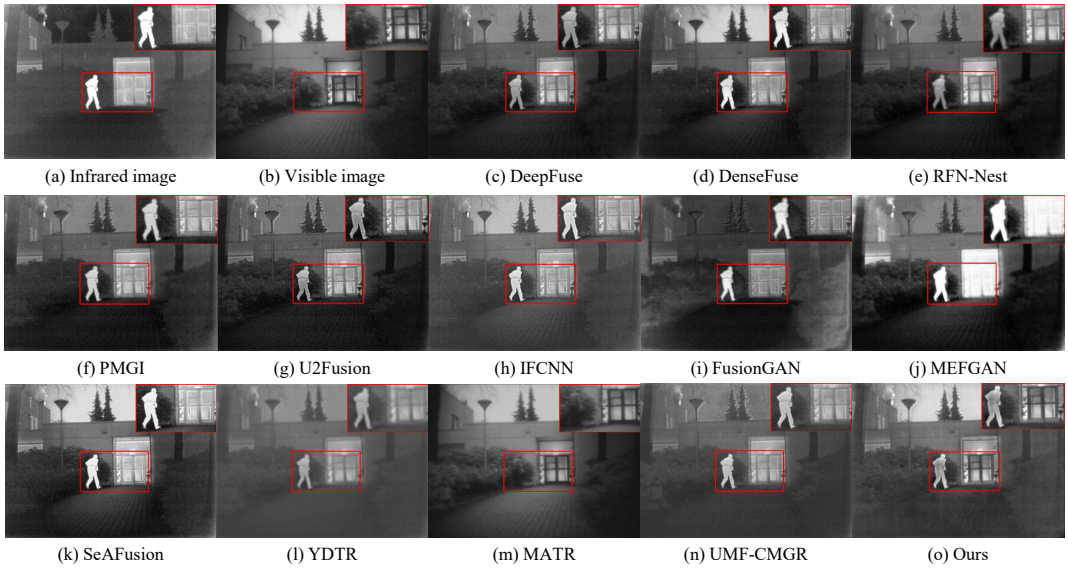


Fig. 8. Infrared and visible image fusion experiment on "human" images.

TABLE I
QUANTITATIVE EVALUATION RESULTS ON TNO DATASET. THE BEST TWO RESULTS ARE HIGHLIGHTED IN **BOLD** AND *Italic* FONTS.

Method	SF	EN	$Q^{ab/f}$	FMI_w	MS_SSIM	FMI_{pixel}	MI	SD	VIF
DeepFuse [46]	8.3500	6.6102	0.3847	0.4214	*0.9138*	0.9041	13.2205	66.8872	0.5752
DenseFuse [12]	9.3238	6.8526	0.4735	*0.4389*	0.8692	0.9061	13.7053	81.7283	0.6875
RFN-Nest [9]	5.8457	6.7274	0.3292	0.3052	0.8959	*0.9063*	13.4547	67.8765	0.5404
PMGI [47]	8.7195	6.8688	0.3787	0.4018	0.8684	0.9001	13.7376	69.2364	0.6904
U2Fusion [27]	11.0368	6.7227	0.3934	0.3594	0.9147	0.8942	13.4453	66.5035	0.7680
IFCNN [15]	*11.8590*	6.6454	*0.4962*	0.4052	0.9129	0.9007	13.2909	73.7053	0.6090
FusionGAN [14]	8.0476	6.5409	0.2682	0.4083	0.6135	0.8875	13.0817	61.6339	0.4928
MEFGAN [48]	7.8481	6.9727	0.2076	0.1826	0.6709	0.8844	13.9454	43.7332	0.7330
SeAFusion [24]	**11.9355**	**7.0331**	0.4908	0.3264	0.8981	0.9052	**14.0663**	*93.3851*	**0.8919**
YDTR [25]	3.2567	6.1933	0.1410	0.1548	0.7823	0.8959	12.3865	56.0668	0.2792
MATR [26]	5.3632	6.5353	0.2723	0.1964	0.7722	0.9047	13.0705	78.0720	0.3920
UMF-CMGR [22]	8.2388	6.3151	0.3671	0.4020	0.8688	0.9040	12.6301	60.7236	0.3934
TGFuse	11.3149	*6.9838*	**0.5863**	**0.4452**	**0.9160**	**0.9219**	*13.9676*	**94.7203**	*0.7746*

segmentation is achieved through sliding windows. Through the sliding window, the fused image can well coordinate the consistency between different image blocks. The calculation of the loss function is shown in Eq.(4).

$$Var\_SSIM(I_X, I_Y, I_F|W) = \begin{cases} SSIM(I_X, I_F), \\ if\sigma^2(X) > \sigma^2(Y) \\ SSIM(I_Y, I_F), \\ if\sigma^2(Y) >= \sigma^2(X) \end{cases} \tag{3}$$

$$L_{var\_SSIM} = 1 - \frac{1}{N}\sum_{W=1}^{N} Var\_SSIM(I_X, I_Y, I_F|W) \tag{4}$$

Discriminator Loss Function The proposed method contains two discriminators: the infrared image and fused image discriminator ("Dis-IR"), the visible image and fused image discriminator ("Dis-VIS"). Different from general classification discriminators, we employ feature-level L1 loss to allow the fused images to obtain features consistent with infrared and visible light images at different levels. Inspired by perceptual loss, we extract image features using a pre-trained VGG-16 network. As shown in Figure. 4, the features extracted by the pre-trained network include four components of different scales and depths, and the features obtained by taking different layers can represent different degrees of abstract meaning of the image.

For infrared image and fused image discriminator ("Dis-IR"), we use the deepest feature to represent the saliency information of the image after deep abstraction, and the loss function is as follows:

$$L_{Dis-IR} = ||\varphi_j(IR) - \varphi_j(Fused)||_1 \tag{5}$$

φ_j represents the j-th layer feature extracted by the pre-trained network, where j=4. For the visible image and fused image discriminator ("Dis-VIS"), we use shallow features to represent the texture information of the image, and the loss function is as follows:

$$L_{Dis-VIS} = ||\varphi_j(VIS) - \varphi_j(Fused)||_1 \tag{6}$$

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2023.3273451

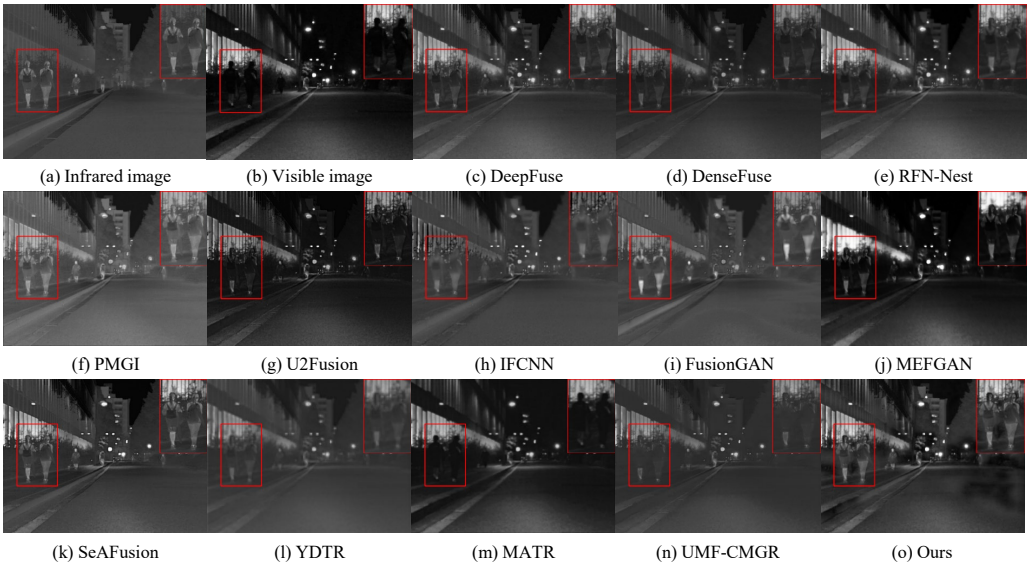


Fig. 9. Infrared and visible image fusion experiment on "street" images.

Fig. 10. Visualization of global information.

and j=1. The loss function for the entire generation and discrimination process is as follows:

$$\min_{G}\min_{D_V,D_I}(G(VIS,IR,Fused)+D_V(Fused,VIS) +D_I(Fused,IR)) \tag{7}$$

Different from classical GAN, the adversarial nature of the proposed loss function is manifested in semantics. First, we train the fusion network (generator) in the training mode of a normal image fusion task. Then, on the basis of the fixed feature extraction network (VGG-16, which is also the discriminator in our design), the generator network is trained again using two loss functions that are semantically opposed to each other. The generator and discriminator are optimized in the same direction, but exhibit opposite semantics. Thus, the performance of the fusion network is improved in semantic adversarial learning.

IV. Experiments

A. Setup

Datasets. In the training phase, 40,000 pairs of corresponding infrared and visible images are selected as the training data from the KAIST [49] data set. KAIST data set is a pedestrian data set containing various general scenes of campus, street and countryside. Each picture contains a visible image and a corresponding infrared image. At present, some end-to-end image fusion algorithms [9] use it as training data. The training image size is set to 256×256 pixels. During the test phase, we use two different test sets for infrared and visible image tasks. One is 10 pairs of images in [12], the other is 188 pairs of images from the test set of [50]. The former is a partial image selected in the TNO dataset. The latter is part of the images selected in the road scene dataset.

Hyper-Parameters. In the training phase, we choose Adam as the optimizer and the learning rate is set to a constant of 0.0001. Training data includes 40,000 pairs of images and batch size is set to 16. Complete training requires 20 epochs. Inspired by [40], [41], we chose fixed values for some parameters in the transformer fusion module. The patch size of the spatial transformer and channel transformer is set to 4 and 16 respectively. Taking into account the different dimensions of the data processed by a spatial transformer and channel transformer, the embedding dimensions are set to 2048 and 128 respectively. Our model is implemented with NVIDIA TITAN Xp and Pytorch.

Compared Methods.

The proposed method is compared with 13 methods in subjective and objective evaluation in three datasets. These are: DenseFuse [12], DeepFuse [46], a general end-to-end fusion network(IFCNN) [15], FusionGAN [14], NestFuse [13], PMGI [47], U2Fusion [27], RFN-Nest [9], and MEF-GAN [48], SeAFusion [24], UMF-CMGR [22], YDTR [25] and MATR [26], respectively. The three data sets include the

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2023.3273451

TABLE II
QUANTITATIVE EVALUATION RESULTS ON ROAD SCENE DATASET. THE BEST TWO RESULTS ARE HIGHLIGHTED IN **BOLD** AND *Italic* FONTS.

Method	SF	EN	$Q^{ab/f}$	FMI_w	MS_SSIM	FMI_{pixel}	MI	SD	VIF
DeepFuse [46]	5.6405	6.0044	*0.5743*	**0.3753**	*0.9711*	*0.9355*	12.0087	54.8582	1.5240
DenseFuse [12]	4.0095	5.5711	0.3804	0.3480	0.9325	0.9354	11.1421	38.7495	0.8090
RFN-Nest [9]	5.2540	6.1904	0.5257	0.2438	**0.9719**	0.9324	12.3809	59.9035	1.6212
PMGI [47]	6.7328	*6.2473*	0.4481	0.3466	0.9007	0.9154	*12.4947*	51.4928	1.5766
U2Fusion [27]	7.0909	5.5092	0.4912	0.2867	0.9572	0.9283	11.0184	49.6143	1.5171
IFCNN [15]	7.4354	5.6613	0.5558	0.3468	0.9403	0.9312	11.3226	47.6592	1.4451
FusionGAN [14]	4.2803	5.6386	0.2835	0.2868	0.8401	0.9264	11.2771	39.7419	1.2244
MEFGAN [48]	6.0142	5.7924	0.2690	0.1954	0.9188	0.9272	11.5847	**74.9715**	*2.0240*
SeAFusion [24]	*7.5154*	5.7905	0.5667	0.2909	0.9180	0.9344	11.5810	54.9380	.7323
YDTR [25]	2.8696	5.2291	0.1509	0.1627	0.8819	0.9248	10.4582	35.0020	0.7249
MATR [26]	4.9477	5.2217	0.3363	0.1865	0.9162	0.9271	10.4435	59.3100	1.3307
UMF-CMGR [22]	4.6683	4.5300	0.2641	0.2610	0.8530	0.9290	9.0601	32.4017	0.7142
TGFuse	**7.6542**	**6.3403**	**0.6012**	*0.3557*	0.9227	**0.9386**	**12.6807**	*62.1760*	**2.1134**

TABLE III
QUANTITATIVE EVALUATION RESULTS ON LLVIP DATASET. THE BEST TWO RESULTS ARE HIGHLIGHTED IN **BOLD** AND *Italic* FONTS.

Method	SF	EN	$Q^{ab/f}$	FMI_w	MS_SSIM	FMI_{pixel}	MI	SD	VIF
DeepFuse [46]	12.4175	7.0222	0.4620	0.3978	0.9008	0.8923	14.0444	0.4586	38.3328
DenseFuse [12]	12.5900	7.0361	0.4700	0.3956	0.9030	0.8923	14.0723	0.4669	38.7011
RFN-Nest [9]	10.6825	7.0642	0.3844	0.3670	0.8939	0.8860	14.1284	0.4658	39.7194
PMGI [47]	12.0997	7.0368	0.3951	0.3460	0.8153	0.8868	14.0737	0.4487	37.9572
U2Fusion [27]	17.2889	6.7070	0.4985	0.3859	0.8746	0.8817	13.4141	0.4917	37.4284
IFCNN [15]	*21.7698*	*7.2417*	0.6092	0.3954	**0.9227**	0.8905	*14.4835*	0.6762	44.0938
FusionGAN [14]	9.2062	6.4490	0.0600	0.1387	0.3607	0.8068	12.8981	0.1141	26.9133
MEFGAN [48]	15.1905	6.9787	0.3644	0.2697	0.8597	0.8774	13.9575	*0.8720*	**59.7947**
SeAFusion [24]	20.9194	**7.4508**	*0.6181*	*0.3982*	*0.9157*	*0.8952*	**14.9016**	0.8392	*51.8096*
YDTR [25]	7.0755	6.6929	0.1961	0.2076	0.8146	0.8809	13.3858	0.3365	33.1625
MATR [26]	13.5066	5.9994	0.4282	0.3288	0.7007	0.8914	11.9989	0.4575	34.1515
UMF-CMGR [22]	13.4481	6.7018	0.3707	0.3482	0.8258	0.8866	13.4037	0.3841	35.1731
TGFuse	**21.9161**	7.0003	**0.6535**	**0.4051**	0.8388	**0.8966**	14.0007	**0.8737**	43.6293

commonly used TNO data set, the road scene data set and the high-resolution high-quality image data set (LLVIP).

B. Results Analysis

We use three test sets of different scenarios and scales for testing. Then the visualized results and objective indicator evaluations are used to measure the performance of the fusion algorithm. Subjective evaluation judges whether the fusion result conforms to human visual perception, such as clarity, salient information, etc. Therefore, the subjective evaluation method puts the fused images obtained by different algorithms together for intuitive visual comparison.

In Figure. 8, pedestrians in infrared images are salient information. While the red box in the visible image is the background information. We show the fusion results of thirteen different methods, highlighting the major different parts with red boxes. Other involved fusion methods can achieve fusion in a certain degree, but cannot highlight the salient information of infrared images and the low-noise background of visible images simultaneously. Compared with other methods, our method not only highlights the infrared information of the person in the red frame but also maintains the visible details of the door. The sky as the background also retains the high-resolution visible scene. Such a fused image is friendly and easy to accept for human vision. The results generated by some fusion algorithms based on traditional methods introduce more noise, which affects the quality of the fusion image. In contrast, the fusion result produced by the deep learning method is more in line with human vision. Most methods based on deep learning can maintain the basic environmental information of the visible image and the salient human of the infrared image at the same time.

Therefore, we mainly compare the latest fusion algorithms based on deep learning in the Figure. 9. Figure. 9 also includes pedestrian and scene information. Our method exhibits an exquisite balance between the two modalities. To further demonstrate the advantage of the algorithm, we compare the metrics with state-of-the-art methods on a high-resolution large-scale test set. Meanwhile, Figure. 10 shows the fused results of the proposed method. (a) is the input infrared and visible image, (b) is the fused results, (c) is the difference map for visible images and infrared images, respectively. The red

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2023.3273451

TABLE IV
THE OBJECTIVE EVALUATION ON WHETHER TO USE GAN. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD** FONTS.

	SF	EN	$Q^{ab/f}$	FMI_w	MS_SSIM	FMI_{pixel}	MI	SD	VIF
w/o GAN	11.2253	6.9547	0.5794	0.4425	**0.9240**	0.9212	13.9094	92.4749	**0.7870**
GAN	**11.3149**	**6.9838**	**0.5863**	**0.4452**	0.9160	**0.9219**	**13.9676**	**94.7203**	0.7746

TABLE V
THE OBJECTIVE EVALUATION ON DIFFERENT TRANSFORMER FUSION METHOD. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD** FONTS.

	SF	EN	$Q^{ab/f}$	FMI_w	MS_SSIM	FMI_{pixel}	MI	SD	VIF
Spatial	10.8364	6.8665	0.5491	0.4281	**0.9337**	0.9173	13.7330	86.2626	0.7247
Channel	11.1283	6.9520	0.5622	0.4328	0.9107	0.9169	13.9040	91.2356	0.7417
Spatial+Channel	10.8808	6.9161	0.5304	0.4139	0.9172	0.9089	13.8323	**94.6343**	0.7565
Channel+Spatial	**11.2253**	**6.9547**	**0.5794**	**0.4425**	0.9240	**0.9212**	**13.9094**	92.4749	**0.7870**

TABLE VI
THE OBJECTIVE EVALUATION ON WHETHER TO USE POSITION EMBEDDING. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD** FONTS.

	SF	EN	$Q^{ab/f}$	FMI_w	MS_SSIM	FMI_{pixel}	MI	SD	VIF
w/o PE	**11.2253**	**6.9547**	**0.5794**	**0.4425**	0.9240	**0.9212**	**13.9094**	**92.4749**	**0.7870**
PE	10.8748	6.9332	0.5522	0.4186	**0.9340**	0.9174	13.8664	90.5422	0.7654

boxes in (a) and (b) show the salient objects in the infrared image, while the red boxes in (c) show the difference map between the fused image and the visible image. It can be seen that the fused image is significantly different from the visible image in the salient object region, while the difference in other parts is very marginal. Similarly, the blue boxes in (a) and (b) show the detail information in the visible image, and the blue boxes in (c) show the difference map between the fused image and the infrared image. The difference map shows a large difference in the detail information. From the results, we can find that the global relationship focused by our method can accurately obtain the saliency information in the infrared image as well as the detailed background of the visible image, thus providing a good fusion effect.

There are many different evaluation indicators for objective evaluation. We have selected nine common evaluation indicators for the quality of fused images. These are: Spatial Frequency (SF) [51], Entropy (EN) [52], quality of images ($Q^{ab/f}$) [53], feature mutual information with wavelet transform(FMI_w) [54], multiscale SSIM (MS_SSIM) [55], feature mutual information with pixel(FMI_pixel) [54] Standard Deviation of Image (SD) [56], Visual Information Fidelity (VIF) [57], and mutual information (MI) [58], respectively. In Table.I, We compared the performance of all methods on 9 evaluation indicators. The best three results are highlighted in **bold** and *italic* fonts. our method performed best on 5 indicators and also achieved second place on the three indicators on 10 pairs of images from TNO dataset. Besides, our method also performed best on 6 indicators on the 188 pairs of images from road scene dataset in Table. II. In Table.III, Our method is compared with the latest methods on a large high-resolution dataset. Our method achieves 5 optimal metrics and still shows excellent performance. Through subjective and objective evaluation, our method is proved to have obvious advantages in performance.

C. Ablation Study

GAN.

Adversarial learning during training is very effective in image generation tasks, but how to combine it with fusion tasks is a problem in its application. Our original method only has the generation part of the fused image and does not include two discriminators. In this case, our method has surpassed the previous method in most objective evaluation indicators. In order to enhance the characteristics of the fused image: the high resolution of the visible image and the highlighted part of the infrared image, we introduce adversarial learning into the training process. We use the pre-trained VGG-16 network as a discriminator to enhance the characteristics of different modalities at the feature level. The objective evaluation results are shown in the Table. IV. Compared with the method that does not use adversarial training, the new method with GAN has improved on seven indicators. This also proves the effectiveness of introducing generative confrontation methods.

Transformer Fusion Module. We propose two transformer fusion methods: spatial transformer and channel transformer. They can work alone or in combination with each other. Spatial or Channel Transformer modules are used alone will focus the fusion relationship on one of them. In order to pay attention to the space and channel relationship at the same time, it is necessary to make an appropriate combination of the two modules. In Table. V, we separately verify the results of using the two transformer fusion modules alone and in combination. The effect of passing through the channel transformer first and then passing through the space transformer will be better. We believe that it is more beneficial for fusion to first pay attention to the channel relationship between corresponding blocks in the process of modelling.

Position Embedding. In our transformer fusion method, position embedding is removed because the category information provided by position embedding is not needed in the

TABLE VII
THE OBJECTIVE EVALUATION ON DIFFERENT ENCODER LAYERS OF TRANSFORMER. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD** FONTS.("/" MEANS TRAINING FAILURE)

	SF	EN	$Q^{ab/f}$	FMI_w	MS_SSIM	FMI_{pixel}	MI	SD	VIF
3-layers					/				
4-layers	**11.2253**	**6.9547**	**0.5794**	**0.4425**	0.9240	**0.9212**	**13.9094**	**92.4749**	**0.7870**
5-layers	11.1740	6.8722	0.5623	0.4209	**0.9404**	0.9198	13.7443	86.7715	0.7539

TABLE VIII
THE OBJECTIVE EVALUATION ON DIFFERENT LAYERS OF CNN. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD** FONTS.("/" MEANS TRAINING FAILURE)

	SF	EN	$Q^{ab/f}$	FMI_w	MS_SSIM	FMI_{pixel}	MI	SD	VIF
2-layers	10.3438	6.7281	0.5560	0.4314	0.9006	0.9097	13.4562	**94.2280**	0.6862
3-layers	11.0769	6.8959	0.5497	0.4272	**0.9298**	0.9157	13.7919	92.5518	0.7517
4-layers	**11.2253**	**6.9547**	**0.5794**	**0.4425**	0.9240	**0.9212**	**13.9094**	92.4749	**0.7870**
5-layers					/				

TABLE IX
THE OBJECTIVE EVALUATION ON DIFFERENT CHANNELS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD** FONTS.

	SF	EN	$Q^{ab/f}$	FMI_w	MS_SSIM	FMI_{pixel}	MI	SD	VIF
32-channels	10.6360	6.9228	0.5715	0.4370	0.9276	0.9206	13.8456	90.1796	0.7061
64-channels	**11.2253**	**6.9547**	**0.5794**	**0.4425**	0.9240	**0.9212**	**13.9094**	**92.4749**	0.7870
128-channels	11.1181	6.9388	0.5545	0.4142	**0.9368**	0.9163	13.8776	88.5524	**0.8069**

TABLE X
THE OBJECTIVE EVALUATION ON DIFFERENT LOSS FUNCTION. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD** FONTS.

	SF	EN	$Q^{ab/f}$	FMI_w	MS_SSIM	FMI_{pixel}	MI	SD	VIF
TGFuse(RMI-Loss)	12.2629	6.3697	0.3013	0.3741	0.7664	0.9045	12.7395	93.6815	0.4915
TGFuse(SF-Loss)	**16.1052**	6.4580	0.3070	0.3569	0.7174	0.8918	12.9161	**120.2446**	0.7279
TGFuse	11.3149	**6.9838**	**0.5863**	**0.4452**	**0.9160**	**0.9219**	**13.9676**	94.7203	**0.7746**

fusion task. However, whether the direct removal of position embedding has an effect on the training of the transformer has not been verified. Therefore, we train the TGFuse model with and without position embedding respectively. Comparing the indicators of the fusion results in Table. VI, we find that removing position embedding has a positive effect on the results.

Transformer Module Layers. The transformer model we use is a multi-layer encoder model based on ViT. The number of encoder layers also has a great impact on performance. Unlike classification tasks, fusion tasks are less complex and require fewer layers. But too few layers may also lead to failure of fusion relationship learning. Therefore, we set different values for experiments to find the number of layers most suitable for the fusion task. The comparative results of the experiment are shown in the Table. VII. When the number of layers is three, the test result is a meaningless black image. It may be that too few layers cause the transformer fusion module can not learn the available fusion relationship. When the number of layers is five, the test result becomes worse. This may be because the fusion relationship learned by the deep transformer fusion module is redundant. We select the most suitable number of layers (4 layers) based on the experimental results.

CNN Layers. Firstly, multi-layer CNN is used to extract features from the input image, which can help the transformer module to converge faster. The number of layers of CNN (that is, the number of "Res-Block") affects the granularity and depth of the extracted features. We set different values to experiment to find the most suitable number of CNN layers. The more layers, the more times the image is downsampled. When the image block is too small, the model cannot learn an effective fusion relationship. As shown in Table. VIII, when the depth is 4 layers, the model learns the best fusion relationship. When the layer is deeper, the resulting image is meaningless black blocks. This means that if the feature block is too small, the fusion module cannot fuse information effectively.

CNN Channels. As an important dimension of image features, the number of feature channels is also an important factor influencing algorithm performance. In the process of feature extraction, we get four image features with the same dimensions but different scales. The difference in the number of channels means that the distribution of channel dimension information is different. In the ablation experiment, we choose a few typical values as the number of channels. After comparison in Table. IX, we select the number of channels (64 channels) with the best performance.

Loss Function. Loss function is a key factor affecting the performance of image fusion algorithms. Therefore, how to

design an appropriate loss function is the main part of the algorithm design. In the base model, we use the improved structural similarity loss. To verify the role of the loss function, we introduce SF-Loss [25] and RMI-Loss [26] for comparison while keeping the base model unchanged. In Table. X, we can find that using other loss functions in the basic model is helpful for some indicators. However, more optimal indicators are obtained by using our loss function method. This shows that our proposed loss function is beneficial for the model to learn the fusion relationship.

V. Conclusion

In this paper, we proposed an infrared and visible image fusion method based on the transformer module and generative adversarial learning. The proposed transformer is deeply involved in the fusion task as a fusion relation learning module. Adversarial learning provides generators with different modal characteristics during the training process at the feature level. This is the first attempt of deep combination and application of transformer and adversarial learning in the image fusion task. Our method has also achieved outstanding performance in subjective and objective evaluation, which proves the effectiveness and advancement of our method. In future, we will further explore the possibility of combining transformers with fusion tasks and try them in downstream tasks.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (62020106012, U1836218, 62106089), and the 111 Project of Ministry of Education of China (B12018).

References

[1] H. Li, X.-J. Wu, and J. Kittler, "Mdlatlrr: A novel decomposition method for infrared and visible image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4733–4746, 2020.

[2] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "Rgb-t object tracking: Benchmark and baseline," *Pattern Recognition*, vol. 96, p. 106977, 2019.

[3] T. Xu, Z. Feng, X.-J. Wu, and J. Kittler, "Adaptive channel selection for robust visual object tracking with discriminative correlation filters," *International Journal of Computer Vision*, vol. 129, no. 5, pp. 1359–1375, 2021.

[4] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion," in *15th Pacific Conference on Computer Graphics and Applications (PG'07)*. IEEE, 2007, pp. 382–390.

[5] Z. Zhang and R. S. Blum, "A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application," *Proceedings of the IEEE*, vol. 87, no. 8, pp. 1315–1326, 1999.

[6] C. Chen, Y. Li, W. Liu, and J. Huang, "Image fusion with local spectral consistency and dynamic gradient sparsity," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2760–2765.

[7] M. Nejati, S. Samavi, and S. Shirani, "Multi-focus image fusion using dictionary-based sparse representation," *Information Fusion*, vol. 25, pp. 72–84, 2015.

[8] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Information Fusion*, vol. 36, pp. 191–207, 2017.

[9] H. Li, X.-J. Wu, and J. Kittler, "Rfn-nest: An end-to-end residual fusion network for infrared and visible images," *Information Fusion*, 2021.

[10] ——, "Infrared and visible image fusion using a deep learning framework," in *2018 24th international conference on pattern recognition (ICPR)*. IEEE, 2018, pp. 2705–2710.

[11] H. Li, X.-j. Wu, and T. S. Durrani, "Infrared and visible image fusion with resnet and zero-phase component analysis," *Infrared Physics & Technology*, vol. 102, p. 103039, 2019.

[12] H. Li and X.-J. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2018.

[13] 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 Transactions on Instrumentation and Measurement*, vol. 69, no. 12, pp. 9645–9656, 2020.

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

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

[16] Y. Fu, X.-J. Wu, and T. Durrani, "Image fusion based on generative adversarial network consistent with perception," *Information Fusion*, 2021.

[17] L. Qu, S. Liu, M. Wang, S. Li, S. Yin, Q. Qiao, and Z. Song, "Transfuse: A unified transformer-based image fusion framework using self-supervised learning," *arXiv preprint arXiv:2201.07451*, 2022.

[18] Y. Yuan, J. Wu, Z. Jing, H. Leung, and H. Pan, "Multimodal image fusion based on hybrid cnn-transformer and non-local cross-modal attention," *arXiv preprint arXiv:2210.09847*, 2022.

[19] J. Zhang, A. Liu, D. Wang, Y. Liu, Z. J. Wang, and X. Chen, "Transformer-based end-to-end anatomical and functional image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.

[20] T. Xu, Z. Feng, X.-J. Wu, and J. Kittler, "Towards robust visual object tracking with independent target-agnostic detection and effective siamese cross-task interaction," *IEEE Transactions on Image Processing*, vol. 32, pp. 1541–1554, 2023.

[21] H. Xu, J. Ma, J. Yuan, Z. Le, and W. Liu, "Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 679–19 688.

[22] D. Wang, J. Liu, X. Fan, and R. Liu, "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," *arXiv preprint arXiv:2205.11876*, 2022.

[23] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5802–5811.

[24] 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," *Information Fusion*, vol. 82, pp. 28–42, 2022.

[25] W. Tang, F. He, and Y. Liu, "Ydtr: infrared and visible image fusion via y-shape dynamic transformer," *IEEE Transactions on Multimedia*, 2022.

[26] W. Tang, F. He, Y. Liu, and Y. Duan, "Matr: multimodal medical image fusion via multiscale adaptive transformer," *IEEE Transactions on Image Processing*, vol. 31, pp. 5134–5149, 2022.

[27] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[28] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, and J. Jiang, "Infrared and visible image fusion via detail preserving adversarial learning," *Information Fusion*, vol. 54, pp. 85–98, 2020.

[29] 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 Journal of Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, 2022.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[31] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[32] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial networks," in *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[33] D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717*, 2017.

[34] J. Liang, H. Zeng, and L. Zhang, "High-resolution photorealistic image translation in real-time: A laplacian pyramid translation network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9392–9400.

[35] H. Liu, Z. Wan, W. Huang, Y. Song, X. Han, and J. Liao, "Pd-gan: Probabilistic diverse gan for image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9371–9381.

[36] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "Tedigan: Text-guided diverse face image generation and manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2256–2265.

[37] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[38] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[41] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 299–12 310.

[42] V. VS, J. M. J. Valanarasu, P. Oza, and V. M. Patel, "Image fusion transformer," *arXiv preprint arXiv:2107.09011*, 2021.

[43] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[44] R. Hou, D. Zhou, R. Nie, D. Liu, L. Xiong, Y. Guo, and C. Yu, "Vif-net: an unsupervised framework for infrared and visible image fusion," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 640–651, 2020.

[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[46] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs." in *ICCV*, vol. 1, no. 2, 2017, p. 3.

[47] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "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*, vol. 34, no. 07, 2020, pp. 12 797–12 804.

[48] H. Xu, J. Ma, and X.-P. Zhang, "Mef-gan: Multi-exposure image fusion via generative adversarial networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 7203–7216, 2020.

[49] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1037–1045.

[50] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5108–5115.

[51] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Transactions on communications*, vol. 43, no. 12, pp. 2959–2965, 1995.

[52] J. W. Roberts, J. A. Van Aardt, and F. B. Ahmed, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *Journal of Applied Remote Sensing*, vol. 2, no. 1, p. 023522, 2008.

[53] C. Xydeas, , and V. Petrovic, "Objective image fusion performance measure," *Electronics letters*, vol. 36, no. 4, pp. 308–309, 2000.

[54] M. Haghighat and M. A. Razian, "Fast-fmi: Non-reference image fusion metric," in *2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT)*. IEEE, 2014, pp. 1–3.

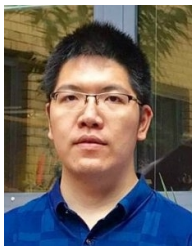
[55] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3345–3356, 2015.

[56] Y.-J. Rao, "In-fibre bragg grating sensors," *Measurement science and technology*, vol. 8, no. 4, p. 355, 1997.

[57] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on image processing*, vol. 15, no. 2, pp. 430–444, 2006.

[58] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electronics letters*, vol. 38, no. 7, pp. 313–315, 2002.

Dong-Yu Rao received the B.E. degree in School of Metallurgy and Environment from Central South University, China, in 2019. He is a master student at the Jiangsu Provincial Engineerinig Laboratory of Pattern Recognition and Computational Intelligence, School of Artificial Intelligence and Computer Science, Jiangnan University. His research interests include image fusion, style transfer and deep learning.



Tianyang Xu received the B.Sc. degree in electronic science and engineering from Nanjing University, Nanjing, China, in 2011. He received his PhD degree at the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China, in 2019. He is currently an Associate Professor at the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China. His research interests include visual tracking and deep learning. He has published several scientific papers, including IJCV, ICCV, TIP, TIFS, TKDE, TMM, TCSVT etc. He achieved top 1 performance in academic competitions, including the VOT2018 public dataset (ECCV18), VOT2020 RGBT challenge (ECCV20), Anti-UAV challenge (CVPR20), Multi-Modal Video Reasoning and Analysing Competition (ICCV21).

Xiao-Jun Wu received the B.Sc. degree in mathematics from Nanjing Normal University, Nanjing, China, in 1991. He received the M.S. degree and the Ph.D. degree in pattern recognition and intelligent systems from Nanjing University of Science and Technology, Nanjing, China, in 1996 and 2002, respectively. He is a Professor in artificial intelligence and pattern recognition at the Jiangnan University, Wuxi, China. His research interests include pattern recognition, computer vision, fuzzy systems, neural networks and intelligent systems. He has won several domestic and international awards because of his research achievements. He served as associate editor for several international journals. He is currently a Fellow of IAPR and AAIA.