Infrared and Visible Image Fusion via Interactive Compensatory Attention Adversarial Learning

Zhishe Wang, Member, IEEE, Wenyu Shao, Yanlin Chen, Jiawei Xu, and Xiaozin Zhang, Senior Member, IEEE

Abstract—The existing generative adversarial fusion methods generally concatenate source images or deep features, and extract local features through convolutional operations without considering their global characteristics, which tends to produce a limited fusion performance. Toward this end, we propose a novel interactive compensatory attention fusion network, termed ICAFusion. In particular, in the generator, we construct a multi-level encoder-decoder network with a triple path, and design infrared and visible paths to provide additional intensity and gradient information for the concatenating path. Moreover, we develop the interactive and compensatory attention modules to communicate their pathwise information, and model their long-range dependencies through a cascading channel-spatial model. The generated attention maps can more focus on infrared target perception and visible detail characterization, and are used to reconstruct the fusion image. Therefore, the generator takes full advantage of local and global features to further increase the representation ability of feature extraction and feature reconstruction. Extensive experiments illustrate that our ICAFusion obtains superior fusion performance and better generalization ability, which precedes other advanced methods in the subjective visual description and objective metric evaluation.

Index Terms—Image fusion, attention interaction, attention compensation, dual discriminators, adversarial learning.

I. INTRODUCTION

INFRARED sensors can perceive heat source target characteristics by receiving thermal radiation, and work at different times or under any weather conditions. However, the obtained images often represent high-brightness targets by pixel intensity, but lack structural textures. On the contrary, visible sensors can characterize rich scene and texture details through light reflection, but fail to identify significant targets, and are sensitive to light conditions, especially in low illumination environments. In fact, these two kinds of sensors have strong complementarity in detection conditions and imaging mechanisms. Therefore, image fusion technology can effectively fulfill their advantages to achieve a more informative image. Compared with a single modality image, the obtained fusion images can better satisfy the human visual system, and benefit other subsequent tasks, such as RGB-T tracking [1], RGB-D salient object detection [2] and multi-spectral pedestrian re-recognition [3] and so on.

Conventional fusion methods typically employ a fixed mathematical model based on prior knowledge of target characteristics to extract features, design an appropriate strategy to combine them, and then reconstruct the final fusion image through corresponding inverse operations. Representative methods are multi-scale transformation [4], [5], sparse representation [6], [7], saliency-based [8], subspace-based [9] and mimicry fusion [10] and others [11], [12]. Due to different imaging mechanisms, infrared images represent target characteristics by pixel intensity, while visible images characterize scene textures by edges and gradient. The conventional fusion methods fail to consider their inherent distinctiveness and develop a uniform mathematical model to indiscriminately extract image features. However, the designed mathematical model is only sensitive to a certain feature and may be inappropriate for other features, which inevitably leads to low fusion performance and poor visual effects in some cases. In addition, the corresponding fusion strategy is manually designed and increasingly complicated, which hinders the practical fusion application.

With the recent advances in machine learning and hardware devices, deep learning has directly facilitated the rapid development of image fusion technology [13]. The non-end-to-end convolutional neural network (CNN)-based methods [14], [15], [16], [17] utilize the pre-trained network for feature extraction and feature reconstruction, which requires the integration of features in the fusion layer. However, the fusion strategy still needs to be designed manually. The end-to-end CNN-based methods [18], [19], [20] generally concatenate input images as a source, or deep features in a dual encoder framework. Due to the lack of ground-truth, their potential fusion performance cannot be further excavated. Different from CNN-based methods, the generative adversarial network (GAN)-based methods [21],
We construct a multi-level encoder-decoder network with norm were adopted as
We propose an end-to-end fusion network for infrared and visible images. Extensive experiments indicate our ICAFusion precedes other representative state-of-the-art fusion methods in terms of subjective visual description and objective metric evaluation.

The rest of this paper is organized as follows. Section II presents the development of CNN-based and GAN-based fusion methods. Section III describes the network framework, interactive and compensatory attention modules and loss function. The related experiments and conclusions are discussed in Sections IV and V, respectively.

II. RELATED WORK

In this section, we comprehensively review the representative CNN-based and GAN-based fusion methods, and further discuss their superiority and drawbacks.

A. CNN-Based Fusion Methods

Compared with the conventional fusion methods, convolutional neural networks employ more filter banks to automatically extract features from the training dataset [24], which can reduce the imperfection of the hand-craft feature extraction model. For example, Jian et al. [14] proposed a modified residual dense network to decompose deep features, and then a visual saliency mechanism was applied to guide feature combinations. However, the proposed network is simple, and not specially trained for fusion tasks. Li et al. [15] presented DenseFuse where densely connected blocks were applied to reemploy intermediate features, and average and $L_1$ norm were adopted as fusion strategies. Luo et al. [16] exploited a multi-branch network with contrastive constraints and designed a general fusion rule based on the disentangled representation. Zhang et al. [17] introduced a general training network with a simple average rule for multi-task image fusion. These methods entirely rely on convolutional operations to extract local features, but ignore their long-range dependencies and inevitably lose some important global information.
In order to exploit local and global features to achieve better representational capacity, Jiang et al. [25] introduced SEDRFuse in which a symmetric network framework was proposed, and a spatial attention fusion strategy was designed. Li et al. [26] presented NestFuse where a decoder network based on nest connections was adopted for better feature reconstruction, and spatial-wise and channel-wise attention models were proposed as fusion strategies. Wang et al. [27] developed Res2Fusion in which two multiple receptive field aggregation blocks were proposed to generate multi-level features, and the nonlocal attention models were designed in the fusion layer. Subsequently, Wang et al. [28] introduced UNFusion where a unified multi-scale dense network was designed, and $L_p$ normalized attention models were proposed to establish long-range dependencies. Moreover, Wang et al. [29] developed SwinFuse for infrared and visible image fusion. Although these methods have achieved supernormal results, their attention-based fusion strategies are not learnable.

To overcome the limitations of hand-designed feature fusion, Long et al. [18] exploited an unsupervised aggregated residual dense network where the pixel-wise and feature-wise loss functions were designed to supervise the training. Li et al. [30] employed a two-stage training mode, namely RFN-Nest, which first trained the encoder-decoder network, and then trained the residual fusion module. Furthermore, for the multi-task image fusion, Zhao et al. [31] designed a novel universal framework to learn specific and general features and proposed a realm activation mechanism to facilitate high generalization of across-realm. Xu et al. [19] proposed a novel unified and unsupervised network to solve multiple fusion problems, which applied the information preservation degrees to constrain the loss function by measuring the importance of corresponding source images. Zhang et al. [20] presented PMGI where the gradient and intensity paths were performed to realize different image fusion tasks. These methods emphasize the design of network structure and loss function, but fail to consider attention mechanism, which causes the loss of some contextual information.

### B. GAN-Based Fusion Methods

In contrast to the aforementioned methods, some researchers translated the fusion problem into feature adversarial training. Typically, Ma et al. presented FusionGAN [21] and its extended version [22] for image fusion tasks. Since their methods only use a discriminator, the obtained fused result is similar to a sharpened infrared image, and seriously lost the texture details of the visible image. To alleviate this problem, they specifically designed two discriminators to realize fusion balance and exploited DDeGAN [32] to implement multi-resolution fusion tasks. Li et al. [33] proposed a Wasserstein GAN with dual discriminators, namely D2WGAN, which had attained better fusion performance than that of the original GAN. In addition, Zhou et al. [34] developed SDDGAN where an information quantity discrimination block was proposed to supervise semantic information of source images. Ma et al. [35] translated image fusion into multi-classification constraints, namely GANMeC, which proposed two multi-classification discriminators to generate a more balanced result. Song et al. [36] introduced a triple-discriminator GAN, which was balanced by adding a new difference image discriminator. By concatenating infrared and visible images as an input, most methods maintain a limited fusion balance.

As a way to solve these issues, Yang et al. [37] proposed a texture conditional GAN, termed TC-GAN, which applied a squeeze-and-excitation module to enhance the texture information from the channel dimension. Li et al. [38] employed a multi-grained attention network with two independent encoders, namely MgAN-Fuse, which integrated a channel attention model into multi-scale layers of the encoder, and then multi-grained attention maps were reconstructed as a fusion image by the decoder. Subsequently, they extended the attention mechanism into a generator and dual-discriminator, termed AttentionFGAN [39], which designed two multi-scale attention networks to generate the attention maps of infrared and visible images, and were directly concatenated with source images for the fusion network to produce a fused result. These methods employ the channel attention mechanism to enhance feature representation, but ignore its spatial characteristics. More importantly, the attention interaction and compensation are also not considered, which limits their performance.

### III. Method

In this section, we first present the overview of the proposed ICAFusion, and then emphatically describe the principle of interactive and compensatory attention modules. Finally, we explain the design of loss function.

#### A. Network Overview

As shown in Fig. 2, the proposed ICAFusion is based on the Wasserstein generative adversarial network, which consists of a generator and dual discriminators.

**Generator Architecture:** The generator includes the encoder part, fusion layer, and decoder part. In the encoder, a triple path, namely infrared, visible, and concatenating path, is designed as input sources. We use four convolutional layers to extract multi-level features, in which the third and fourth layers employ convolution with a stride of two. The features of infrared and visible paths are respectively concatenated with that of the concatenating path, termed $\Phi_m$ and $\Phi_n$, and then fed into an interactive attention module to produce their interactive attention maps, termed $\Phi_F$. After three-level feature interactions, the final interactive attention maps are obtained. In the fusion layer, the final interactive attention maps are directly concatenated with the compensatory attention maps of infrared and visible paths to generate the fused attention maps. Subsequently, in the decoder, we also use four convolutional layers to reconstruct features, where the first two layers are along with upsampling operation. The obtained output is concatenated with the corresponding compensatory attention maps of infrared and visible paths for subsequent reconstruction. All the layers use $3 \times 3$ convolution kernels along with PReLU activation, except for the last layer with Tanh function.

**Discriminator Architecture:** The Discriminator-IR and Discriminator-VIS share the same network framework, which
consists of four convolution layers and a fully connected layer. All the convolution layers adopt strided convolution with $3 \times 3$ kernel size and LeakyRelu activation function, except for the last layer, which is the Tanh function. The stride is set to 2, and the corresponding filter banks are set to 16, 32, 64, and 128. During the training process, we input the initial fusion image $I_f$, infrared image $I_r$, and visible image $I_v$ into the corresponding discriminator, which aim to distinguish $I_f$ from $I_r$ and $I_v$. The Discriminator-IR force $I_f$ to gradually preserve more and more infrared pixel intensity information, while the Discriminator-VIS force $I_f$ to increasingly contain more and more visible detail information. When the adversarial game of the generator and dual discriminators reaches equilibrium, this indicates that the generator has fooled dual discriminators, and the desired fused result is obtained.

### B. Interactive and Compensatory Attention Modules

Inspired by CBAM [40], we refine it to construct our interactive and compensatory attention modules. Their differences are summarized in three aspects. First, we design two convolutional layers to replace the originally shared multi-layer perceptron (MLP), which can reduce the computational complexity. Second, we propose concatenating operation rather than element-wise summation used in the CBAM, which can retain different spatial or channel contextual information by the global average and maximum pooling layers. The key issue is that we use the softmax function to interact with spatial or channel weighted coefficients of infrared and visible images, and the generated attention maps more focus on their respective representative contents.

The framework of the interactive attention module is shown in Fig. 3. For the intermediate features $\Phi_m, \Phi_n \in \mathbb{R}^{H \times W \times C}$, where $H, W$ and $C$ respectively represent the height, width, and the number of channels of feature maps, we firstly employ the global average and maximum pooling operations to aggregate feature maps into channel descriptions, respectively. Both descriptions pass through two convolutional layers with $1 \times 1$ kernel size and a PReLU activation layer, the output feature vectors are concatenated, and forwarded to the convolutional layer and sigmoid activation layer. In short, after the channel attention model, we obtain their respective initial channel weighted coefficients $\phi_m^{ca}$ and $\phi_n^{ca} \in \mathbb{R}^{1 \times 1 \times C}$, which are computed as (1) and (2), respectively.

$$\phi_m^{ca}(c) = \delta(Conv(Conv(Conv(\sigma(Conv(AP(\Phi_m)))))))$$
$$\phi_n^{ca}(c) = \delta(Conv(Conv(Conv(\sigma(Conv(AP(\Phi_n)))))))$$ (1)

where $Conv$ and $Conv$ represent the convolution and concatenation operations, $AP(\cdot)$ and $MP(\cdot)$ denote global average and maximum pooling operations. $\sigma$ and $\delta$ represent PReLU and sigmoid activation functions. $c = 1, 2, \ldots, C$.

And then, we apply the softmax function to produce their final channel weighted coefficients, i.e., $\beta_m^{ca}, \beta_n^{ca} \in \mathbb{R}^{1 \times 1 \times C}$, which are formulated as (3) and (4), respectively.

$$\beta_m^{ca}(c) = \frac{\exp(\phi_m^{ca}(c))}{\exp(\phi_m^{ca}(c)) + \exp(\phi_n^{ca}(c))}$$
$$\beta_n^{ca}(c) = \frac{\exp(\phi_n^{ca}(c))}{\exp(\phi_m^{ca}(c)) + \exp(\phi_n^{ca}(c))}$$ (4)

We multiply the final channel weighted coefficients with their respective input features to obtain their corresponding channel attention maps, namely $\Phi_m^{ca}$ and $\Phi_n^{ca} \in \mathbb{R}^{H \times W \times C}$, which are
expressed as (5) and (6).

\[
\Phi_{m}^{ca}(i, j) = \Phi_{m}(i, j) \times \beta_{m}^{ca}(c) \quad (5)
\]
\[
\Phi_{n}^{ca}(i, j) = \Phi_{n}(i, j) \times \beta_{n}^{ca}(c) \quad (6)
\]

Subsequently, the corresponding channel attention maps are taken as the input of the spatial attention model, and forwarded to the global average and maximum pooling layers. The output spatial feature maps are concatenated, and fed into a convolutional layer and a sigmoid activation layer, we obtain their respective initial spatial weighted coefficients \(\varphi_{m}^{sa}\) and \(\varphi_{n}^{sa}\in R^{H\times W \times 1}\), which are computed as (7) and (8).

\[
\varphi_{m}^{sa}(i, j) = \delta(Conv(Conv[AP(\Phi_{m}^{ca}), MP(\Phi_{m}^{ca})])) \quad (7)
\]
\[
\varphi_{n}^{sa}(i, j) = \delta(Conv(Conv[AP(\Phi_{n}^{ca}), MP(\Phi_{n}^{ca})])) \quad (8)
\]

And then, we again apply the softmax operation to produce their final spatial weighted coefficients, \(\beta_{m}^{sa}\) and \(\beta_{n}^{sa}\in R^{H\times W \times 1}\), which are formulated as (9) and (10), respectively.

\[
\beta_{m}^{sa}(i, j) = \frac{\exp(\varphi_{m}^{sa}(i, j))}{\exp(\varphi_{m}^{sa}(i, j)) + \exp(\varphi_{n}^{sa}(i, j))} \quad (9)
\]
\[
\beta_{n}^{sa}(i, j) = \frac{\exp(\varphi_{n}^{sa}(i, j))}{\exp(\varphi_{m}^{sa}(i, j)) + \exp(\varphi_{n}^{sa}(i, j))} \quad (10)
\]

Similarly, we multiply the final spatial weighted coefficients with their channel attention maps to produce their respective spatial attention maps, namely \(\Phi_{m}^{sa}\) and \(\Phi_{n}^{sa}\in R^{H\times W \times C}\), which are computed as (11) and (12), respectively.

\[
\Phi_{m}^{sa}(i, j) = \Phi_{m}^{ca}(i, j) \times \beta_{m}^{sa}(i, j) \quad (11)
\]
\[
\Phi_{n}^{sa}(i, j) = \Phi_{n}^{ca}(i, j) \times \beta_{n}^{sa}(i, j) \quad (12)
\]

Finally, we directly concatenate their corresponding spatial attention maps to produce the fused attention maps, which are expressed as (13).

\[
\Phi_{F}(i, j) = Con[\Phi_{m}^{sa}(i, j), \Phi_{n}^{sa}(i, j)] \quad (13)
\]

Notably, the compensatory attention module is equivalent to the upper part of the interactive module with only a feature input, and does not require the softmax and concatenation operations. In other words, the features of infrared or visible image are sequentially fed into the channel and spatial attention models to produce their attention maps, which are used to compensate information for feature reconstruction.

### C. Loss Function

In the proposed ICAFusion, we need to design the loss function of the generator and dual discriminators, respectively. In the generator, the loss function consists of adversarial loss \(L_{adv}\) and content loss \(L_{con}\), which is expressed as (14).

\[
L_{G} = L_{con} + \lambda L_{adv} \quad (14)
\]

Considering that infrared image represents target characteristics by pixel intensity, while visible image characterizes scene textures by edges and gradient. In this paper, we adopt Frobenius norm and \(L_{1}\) norm to constrain the fused result similar with source images. Therefore, the content loss function is expressed as (15).

\[
L_{con} = \frac{1}{HW}(||I_f - I_r||_{F}^{2} + \xi ||I_f - I_v||_{1}) \quad (15)
\]

where \(|| \cdot ||_{F}\) and \(|| \cdot ||_{1}\) denote frobenius norm and \(L_{1}\) norm, respectively.

In the dual discriminators, the Discriminator-IR (\(D_{r}\)) and Discriminator-VIS (\(D_{v}\)) are designed to balance the authenticity of the fused result and source images, so that the generated result tends to the real data distribution of the source images.
The adversarial loss function is expressed as (16).

\[ L_{\text{adv}} = -\frac{1}{N} \sum_{n=1}^{N} [D_r(I^r_n)] - \frac{1}{N} \sum_{n=1}^{N} [D_v(I^v_n)] \]  

(16)

Meanwhile, the respective loss function of two discriminators are expressed as (17) and (18), respectively.

\[ L_{D_r} = \frac{1}{N} \sum_{n=1}^{N} \left[ D_r(I^r_{n,f}) + \gamma(1 - ||\nabla D_r(I^r_{n,v})||_2^2) \right] \]  

(17)

\[ L_{D_v} = \frac{1}{N} \sum_{n=1}^{N} \left[ D_v(I^v_{n,f}) + \gamma(1 - ||\nabla D_v(I^v_{n,v})||_2^2) \right] \]  

(18)

where \( \gamma \) is the regularization parameter, \( || \cdot ||_2 \) denotes \( L_2 \) norm, and \( \nabla \) indicates the gradient operator. The first term represents the Wasserstein distance between fused result and infrared or visible image, while the second term is the gradient penalty, which limits the learning ability of the discriminator.

IV. EXPERIMENTS AND DISCUSSIONS

In this section, experimental settings are firstly described, and then the ablation study on the attention mechanism is discussed. Finally, we conduct the related experiments on different datasets to demonstrate the effectiveness and superiority of our ICAFusion.

A. Training and Testing Details

In the training process, the TNO dataset [41] including 25 infrared and visible image pairs is used for the training. To expand the training dataset, we use the sliding step of 12 to divide original image pairs into smaller images of size 128 × 128, and convert the gray value range to \([-1, 1]\). Thus, we can obtain 18813 patch pairs. In addition, the Adam optimizer is applied to update model parameters, batchsize and epoch are set to 4 and 16, respectively. The learning rate of the generator and discriminator are set as 1 \times 10^{-4} and 4 \times 10^{-4}, and the corresponding iterations are set to 1 and 2, respectively. In the loss function, the parameters \( \lambda, \xi \) and \( \gamma \) are set to 1, 1 and 10, respectively. The experimental training platform is Intel I9-10850 K, 64 GB memory and NVIDIA GeForce GTX 3090. The programming environment is Python and PyTorch.

In the testing process, the TNO (25 image pairs), Nato_camp sequence (31 image pairs), Roadsence (40 image pairs) [42] and OTCBV5 (40 image pairs) [43] datasets are successively selected for the testing. We adopt nine representative methods, namely MDLatLR [7], DenseFuse [15], IFCNN [17], Res2Fusion [27], SEDRFuse [25], RPN-Nest [30], PMGI [20], FusionGAN [21] and GANMelC [35], to compare with our ICAFusion. Besides, eight metrics, such as average gradient (AG), entropy (EN) [44], standard deviation (SD) [45], mutual information (MI) [46], spatial frequency (SF) [47], nonlinear correlation information entropy (NCIE) [47], \( Q_{\text{obj}} \) [48] and visual information fidelity (VIF) [49] are employed for objective evaluation. The optimal values are bold in red, and the suboptimal values are underlined in blue.

B. Ablation Study

In ablation studies, we validate the design of triple path, attention mechanism, dual discriminators and loss function by subjective and objective experiments. The TNO database is proposed as a testing benchmark.

1) Experimental Validation of Triple Path: In the generator, we design a triple path to achieve feature interaction and compensation. To verify its effectiveness and superiority, five validation models are proposed for comparison, namely, retaining the infrared-visible concatenating path without infrared and visible paths (termed Con_Path), only retaining the infrared path (termed IR_Path), only retaining the visible path (termed Vis_Path), only retaining the interactive attention modules (termed No_Comp) and only retaining the compensatory attention modules (termed No_Inter).

Fig. 4 gives the subjective experimental validation of the triple path. Con_Path has a very serious loss of details, where the billboard is completely unclear. Due to only a single modality of compensatory information, Vis_Path has clear texture details, while IR_Path generates an opposite effect. Moreover, No_Inter has a similar result with our ICAFusion, and has a better visual effect than No_Comp. In addition, the objective experimental results are presented in Table I. Our ICAFusion achieves all the first ranks except for SD, which is the second rank. This indicates that our method is superior to other models, and its design is reasonable and effective.

2) Experimental Validation of Attention Mechanism: In the interactive and compensatory attention modules, the channel and spatial models are designed to model the long-range
Fig. 5. The subjective experimental validation of attention mechanism. The first two images are source image, and the others are generated by No_Att, Only_CA, Only_SA, No_Sigm, CBAM and our ICAFusion, respectively.

Table II

| Metrics | No_Att | Only_CA | Only_SA | No_Sigm | CBAM | Ours |
|---------|--------|---------|---------|---------|------|------|
| AG      | 3.33647| 6.16466 | 6.05384 | 6.05747 | 6.39599| 6.19600|
| EN      | 7.06631| 7.06256 | 7.06595 | 7.00938 | 7.05715| 7.08028|
| SD      | 39.560 | 39.9227 | 40.3394 | 38.5365 | 39.7510| 40.4480|
| MI      | 2.80873| 4.17490 | 4.12244 | 3.77709 | 3.76527| 4.21469|
| NCIE    | 0.80669| 0.81382 | 0.81324 | 0.81144 | 0.81156| 0.81432|
| Q_{abf} | 0.30799| 0.47225 | 0.46105 | 0.45381 | 0.44867| 0.47125|
| VIF     | 0.32995| 0.49072 | 0.49084 | 0.47350 | 0.46719| 0.49254|

The objective experimental validation of attention mechanism is shown in Fig. 5. Due to the lack of attention mechanism, No_Att achieves a limited visual effect, e.g., the street lamp is blurry. By contrast, the other four models, such as Only_CA, Only_SA, No_Sigm and CBAM, generate the similar results with our ICAFusion, and their visual differences cannot be distinguished by subjective observation. Table II shows their objective experimental results. We can find that our ICAFusion acquires EN, SD, MI, NCIE and VIF, the second rank for AG, SF and Q_{abf}. This demonstrates that the designed attention modules are crucial to improve feature extraction and reconstruction capacity.

3) Experimental Validation of Dual Discriminators and Loss Function: In our fusion model, dual discriminators and special loss function are designed for the network training. In this ablation study, we compare it with other five different models, namely, only retaining visible discriminator (termed No_Att), only retaining spatial attention mechanism (termed Only_SA), removing the Sigmoid function (termed No_Sigm) and the alternative CBAM (termed CBAM).

The subjective experimental validation of attention mechanism is shown in Fig. 5. Due to the lack of attention mechanism, No_Att achieves a limited visual effect, e.g., the street lamp is blurry. By contrast, the other four models, such as Only_CA, Only_SA, No_Sigm and CBAM, generate the similar results with our ICAFusion, and their visual differences cannot be distinguished by subjective observation. Table II shows their objective experimental results. We can find that our ICAFusion acquires EN, SD, MI, NCIE and VIF, the second rank for AG, SF and Q_{abf}. This demonstrates that the designed attention modules are crucial to improve feature extraction and reconstruction capacity.

Table III

| Metrics | Dis_Vis | Dis_IR | F_Norm | L1_Norm | Con_Loss | Ours |
|---------|---------|--------|--------|---------|----------|------|
| AG      | 3.40164| 1.51714| 4.83394| 3.60554 | 3.67882 | 6.19600|
| EN      | 6.96149| 6.58204| 6.93304| 6.67666 | 6.82531 | 7.08028|
| SD      | 48.7097| 32.8346| 35.5284| 30.0395 | 35.5237 | 40.4480|
| MI      | 2.58706| 2.71653| 2.02297| 2.00306 | 2.93859 | 4.21469|
| SF      | 6.72470| 3.52594| 9.11660| 7.01229 | 7.33913 | 11.7012|
| NCIE    | 0.80635| 0.80617| 0.80448| 0.80469 | 0.80731 | 0.81432|
| Q_{abf} | 0.16775| 0.14830| 0.43094| 0.36008 | 0.33118 | 0.47125|
| VIF     | 0.12872| 0.12537| 0.40615| 0.40747 | 0.42525 | 0.49254|

The objective experimental validation of dual discriminators and loss function is shown in Fig. 6. Due to a modality dependencies. To test their superiority, five comparable models are used for validation, which are without attention modules (termed No_Att), only retaining channel attention mechanism (termed Only_CA), retaining spatial attention mechanism (termed Only_SA), removing the Sigmoid function (termed No_Sigm) and the alternative CBAM (termed CBAM).

The objective experimental validation of attention mechanism is shown in Fig. 5. Due to the lack of attention mechanism, No_Att achieves a limited visual effect, e.g., the street lamp is blurry. By contrast, the other four models, such as Only_CA, Only_SA, No_Sigm and CBAM, generate the similar results with our ICAFusion, and their visual differences cannot be distinguished by subjective observation. Table II shows their objective experimental results. We can find that our ICAFusion acquires EN, SD, MI, NCIE and VIF, the second rank for AG, SF and Q_{abf}. This demonstrates that the designed attention modules are crucial to improve feature extraction and reconstruction capacity.

C. Results on TNO Dataset

We conduct experiments on the TNO dataset to demonstrate the effectiveness of the proposed ICAFusion. Seven typical image pairs, such as Soldiers_with_jeep, Street, Nato_camp, Kaptein_1654, Movie_01, Sandpath and soldier_in_trench_1, are chosen for the subjective validation, and the corresponding comparative results are presented in Fig. 7. From these results, the traditional method MDLatLRR proposes the learnable low-rank representation, and the obtained fused results exist the undesired artifacts. The CNN-based methods, such as DenseFuse and IFCNN, adopt an average fusion rule under the simple network framework, the obtained results have obvious detail missing and low contrast. However, SEDRFuse and Res2Fusion achieve relatively better performance, because these methods propose a fusion strategy based on the attention mechanism. Their results can retain typical infrared targets, but produce
Fig. 7. The subjectively comparative results of seven examples selected from the TNO dataset, such as Soldiers_with_jeep, Street, Nato_camp, Kaptein_1654, Movie_01, Sandpath and soldier_in_trench_1. The top two lines are source image, and the others are generated by MDLatLRR [7], DenseFuse [15], IFCNN [17], Res2Fusion [27], SEDRFuse [25], PMGI [20], RFN-Nest [30], FusionGAN [21] and GANMcC [35] and our ICAFusion, respectively.
some sharpened effects to a certain degree, and some useful texture information is lost. In addition, for the end-to-end methods, RFN-Nest is inclined to preserve abundant visible details while missing typical infrared targets. PMGI achieves satisfactory results by maintaining the proportional gradient and intensity, but its ability to perceive infrared targets and characterize visible details is still limited. FusionGAN and GANmC intend to retain prominent target information from infrared images. Due to a discriminator, FusionGAN generates unbalanced results, which sharpens the infrared target edges and lacks the important visible details. Although GANmC proposes two discriminators to realize some visual improvement, some useful texture details of visible images are still missing. Compared with the above methods, our ICAFusion achieves the optimal visual effects in maintaining typical infrared targets and unambiguous visible details.

To facilitate the visual observation, we mark some typical infrared targets in the red box and magnify the representative visible details in the green box. As shown in Fig. 7, for the first column images, i.e., the results of Soldiers_with_jeep, MD-LatLRR, DenseFuse, IFCNN, and RFN-Nest can preserve the texture details of the housetop but lost the brightness of pedestrians. On the contrary, FusionGAN and GANmC can retain the targets of infrared images, while the edges of pedestrians are blurred, and the details of the housetop are missing. SEDRFuse and Res2Fusion achieve better results, but their visual effects are also limited. Especially, Res2Fusion lacks some useful scene information, such as trees and clouds. For the results of Street, compared with other methods, our ICAFusion can preserve higher brightness of pedestrians and clearer details of billboards, and our result has higher image contrast. The other five image pairs can draw a similar conclusion. In general, the subjective experiments demonstrate that our method can obtain better fusion performance and is more appropriate to the human visual system (HVS).

We continue to verify our ICAFusion from the perspective of objective evaluation. Fig. 8 gives the comparative results of different methods for the TNO dataset. Notably, the abscissa represents the number of testing images, and the ordinate denotes the average value of the evaluation metric for the corresponding fusion images. The metric curves of our method are described by a red dotted line. We can find that our ICAFusion achieves the highest average values of most metrics. Specifically speaking, our ICAFusion acquires the first rank for AG, MI, SF, NCIE, and VIF, and the second rank for EN, SD, and Qabf, which correspondingly follow behind SEDRFuse and Res2Fusion. In addition, the objective comparative results of the Nato_camp sequence are shown in Fig. 9. Our ICAFusion acquires the first rank for EN, SD, MI, NCIE, Qabf, and VIF, and the third rank for AG and SF, which are lower than IFCNN and Res2Fusion. In conclusion, our ICAFusion implements higher fusion performance and surpasses other representative methods in the subjective visual description and objective metric evaluation.

D. Results on Roadscene Dataset

To further illustrate the superiority of the proposed method, 40 infrared and visible image pairs are selected from the Roadscene dataset for experimental verification. Figs. 10 and 11 give the subjectively comparative results of different fusion methods for FLIR_07210 and FLIR_07081. Our ICAFusion owns three distinct advantages. Firstly, our method can retain the high-brightness target information from the infrared image. For the typical infrared targets, e.g., street lamp and car, our results have higher brightness than those of other methods. Secondly, our method can preserve abundant and unambiguous texture details from the visible image. For example, the representational details, e.g., signboard and decorative lights, obtained by our method are clearer than that of other methods. Thirdly, our method can achieve higher contrast and better visual perception.
Compared with source images and other fused results, due to the application of interactive and compensatory attention modules, our ICAFusion can well preserve prominent target characteristics and unambiguous scene details.

Fig. 12 shows the objective results of different methods for the Roadscene dataset. The proposed method obtains the first rank for metrics SD, MI, NCIE and VIF, and the second rank for metrics AG, EN and SF. The objective experiments also demonstrate that the fusion performance of our ICAFusion surpasses other methods. In addition, the competitive values EN and SF indicate that our results can maintain abundant useful information from source images. This is because our method proposes a triple path where infrared and visible paths can provide additional intensity and gradient information for the fused image. The largest MI and NCIE demonstrate that our results have a strong correlation and similarity with source images. The reason is that our method adopts two discriminators to supervise and optimize the generator with a specific loss function, and can produce a more balanced fusion result. The largest SD and VIF explain that our results can achieve better image contrast and visual effect. This is because our interactive and compensatory attention modules model long-range dependencies and improve their capabilities, allowing them to focus more on infrared target recognition and visual detail characterization.
Fig. 12. The objectively comparative results of different fusion methods for the Roadscene dataset. The corresponding average values of different fusion methods are also presented. Note that our ICAFusion is indicated by a red dotted line.

Fig. 13. The subjectively comparative results of different fusion methods for video_1007 selected from the OTCBVS dataset. The left two images are source image, and the others are generated by MDLatLRR [7], DenseFuse [15], IFCNN [17], Res2Fusion [27], SEDRFuse [25], PMGI [20], RFN-Nest [30], FusionGAN [21] and GANMcC [35] and our ICAFusion, respectively.

E. Results on OTCBVS Dataset

We further conduct the experiments on the OTCBVS dataset to clarify the generalization ability of the proposed ICAFusion. We select 40 image pairs of the pedestrian change sequence, and the comparative results are shown in Fig. 13. By contrast, our ICAFusion presents a richer background scene and involves unambiguous details of the ash-bin. The typical target region, e.g., the pedestrians, can also be contained. As a whole, our method generates a more balanced result and produces better visual perception. The objectively comparative results are shown in Fig. 14. Our method acquires the 1st ranking for EN, SD, MI, NCIE, and VIF, and the 2nd ranking for AG, SF, and Qabf, which only follows behind IFCNN.

In addition, we also verify the computational efficiency of different fusion methods. The traditional method MDLatLRR is tested on the CPU, while the others are implemented on the GPU. Table IV shows the comparative results of different fusion methods. The experiments show that our ICAFusion achieves the competitive fusion efficiency, which is slightly lower than that of DenseFuse and IFCNN. The main reason is that both methods propose a simple network framework with a weighted average fusion rule. In conclusion, the above subjective and objective experiments demonstrate that our ICAFusion achieves remarkable results, and is superior to other methods on three different datasets, indicating that it has better fusion performance and stronger generalization ability.

| Method       | EN      | SD      | MI      | NCIE    | VIF     | AG      | SF      | Qabf   | TNO | Roadscene | OTCBVS |
|--------------|---------|---------|---------|---------|---------|---------|---------|--------|-----|----------|--------|
| MDLatLRR     | 7.941\times10^1 | 3.839\times10^1 | 1.956\times10^1 | 
| DenseFuse    | 8.509\times10^{-2} | 4.001\times10^{-2} | 2.315\times10^{-2} | 
| IFCNN        | 4.554\times10^{-2} | 1.149\times10^{-2} | 2.246\times10^{-2} | 
| Res2Fusion   | 1.886\times10^1  | 4.267    | 1.337   | 
| SEDRFuse     | 2.676   | 1.445    | 8.031\times10^{-1} | 
| PMGI         | 5.445\times10^{-1} | 2.928\times10^{-1} | 1.262\times10^{-1} | 
| RFN-Nest     | 1.777\times10^{-1} | 8.609\times10^{-2} | 5.181\times10^{-2} | 
| FusionGAN    | 2.015   | 1.093    | 4.903\times10^{-1} | 
| GANMcC       | 4.210   | 2.195    | 1.017   | 
| Ours         | 1.309\times10^{-1} | 7.610\times10^{-2} | 3.245\times10^{-2} | 

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Fig. 14. The objectively comparative results of different fusion methods for the OTCBVS dataset. The corresponding average values of different fusion methods are also presented. It is noted that our ICAFusion is indicated by a red dotted line.

V. CONCLUSION

In this paper, an interactive compensatory attention adversarial learning network, termed as ICAFusion, is developed. We construct a multi-level encoder-decoder network with a triple path, where infrared and visible paths provide additional intensity and gradient information for the subsequent processing. The interactive and compensatory attention modules are developed to communicate their pathwise information and model the long-range dependencies. The obtained attention maps can more focus on infrared target perception and visible detailed characterization, and further increase the representation power of feature extraction and feature reconstruction. In addition, dual discriminators are designed to identify the similar distribution between fused results and source images. Moreover, the specific loss function is adopted and the generator is optimized to produce a more balanced result.

We carry out extensive experiments on the TNO, Roadscene, and OTCBVS datasets, and the related results demonstrate that our ICAFusion achieves satisfactory fusion performance along with high computational efficiency and strong generalization ability, preceding other nine state-of-the-art fusion methods in the subjective visual description and objective metric evaluation. In future work, we will continue to optimize the network architecture and introduce attention mechanisms into discriminators to further improve the equilibrium and effectiveness of the adversarial training. In addition, we further extend this network for other tasks, such as multi-band, multi-exposure, multi-focus image fusion, etc.

REFERENCES

[1] Q. Xu, Y. Mei, J. Liu, and C. Li, “Multimodal cross-layer bilinear pooling for RGBT tracking,” IEEE Trans. Multimedia, vol. 24, pp. 567–580, 2022.
[2] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, “CCAFNet: Crossflow and cross-scale adaptive fusion network for detecting salient objects in RGB-D images,” IEEE Trans. Multimedia, vol. 24, pp. 2192–2204, 2022.
[3] Z. Feng, J. Lai, and X. Xie, “Learning modality-specific representations for visible-infrared person re-identification,” IEEE Trans. Image Process., vol. 29, pp. 579–590, 2020.
[4] Z. Wang, F. Yang, Z. Peng, L. Chen, and L. Ji, “Multi-sensor image enhanced fusion algorithm based on NSST and top-hat transformation,” Optik, vol. 126, no. 23, pp. 4184–4190, 2015.
[5] Z. Wang, J. Xu, X. Jiang, and X. Yan, “Infrared and visible image fusion via hybrid decomposition of NSCT and morphological sequential toggle operator,” Optik, vol. 201, pp. 1–11, 2020.
[6] Q. Zhang, Y. Liu, R. S. Blum, J. Han, and D. Tao, “Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review,” Inf. Fusion, vol. 40, pp. 57–75, 2018.
[7] H. Li, X. Wu, and J. Kittler, “MDLatLRR: A novel decomposition method for infrared and visible image fusion,” IEEE Trans. Image Process., vol. 29, pp. 4733–4746, 2020.
[8] J. Ma, Z. Zhou, B. Wang, and H. Zong, “Infrared and visible image fusion based on visual saliency map and weighted least square optimization,” Infrared Phys. Technol., vol. 82, pp. 8–17, 2017.
[9] W. Kong, L. Yang, X. Feng, and H. Zhao, “Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization,” Inf. Fusion, vol. 67, pp. 161–172, 2014.
[10] P. Hu, F. Yang, H. Wei, L. Ji, and D. Liu, “A multi-algorithm block fusion method based on set-valued mapping for dual-modal infrared images,” Infrared Phys. Technol., vol. 102, 2019, Art. no. 102977.
[11] Z. Li, H. Hu, W. Zhang, S. Pu, and B. Li, “Spectrum characteristics preserved visible and near-infrared image fusion algorithm,” IEEE Trans. Multimedia, vol. 23, pp. 306–319, 2021.
[12] R. Nie, C. Ma, J. Cao, H. Ding, and D. Zhou, “A total variation with joint norms for infrared and visible image fusion,” IEEE Trans. Multimedia, vol. 24, pp. 1460–1472, 2022.
[13] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, “Image fusion meets deep learning: A survey and perspective,” Inf. Fusion, vol. 76, pp. 323–336, 2021.
[14] L. Jian et al., “Infrared and visible image fusion based on deep decomposition network and saliency analysis,” IEEE Trans. Multimedia, vol. 24, pp. 3314–3326, 2022.
[15] H. Li and X.-J. Wu, “DenseFuse: A fusion approach to infrared and visible images,” IEEE Trans. Image Process., vol. 28, no. 5, pp. 2614–2623, May 2019.
[16] X. Luo, Y. Gao, A. Wang, Z. Zhang, and X. Wu, “IFSepR: A general framework for image fusion based on separate representation learning,” IEEE Trans. Multimedia, early access, Nov. 19, 2021, doi: 10.1109/TMM.2021.3129354.
[17] Y. Zhang et al., “IFCNN: A general image fusion framework based on convolutional neural network,” Inf. Fusion, vol. 54, pp. 99–118, 2020.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
[18] Y. Long, H. Jia, Y. Zhong, Y. Jiang, and Y. Jia, “RXDNFuse: A aggregated residual dense network for infrared and visible image fusion,” *Inf. Fusion*, vol. 69, pp. 128–141, 2021.

[19] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, “U2fusion: A unified unsupervised image fusion network,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.

[20] 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,” *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 12797–12804, 2020.

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

[22] J. Ma et al., “Infrared and visible image fusion via detail preserving adversarial learning,” *Inf. Fusion*, vol. 54, pp. 85–98, 2020.

[23] J. Fu, X. Wu, and T. Durrani, “Image fusion based on generative adversarial network consistent with perception,” *Inf. Fusion*, vol. 72, pp. 110–125, 2021.

[24] X. Zhang, J. Wang, T. Wang, and R. Jiang, “Hierarchical feature fusion with mixed convolution attention for single image dehazing,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 510–522, Feb. 2022.

[25] L. Jiang et al., “SedRFuse: A symmetric encoder-decoder with residual block network for infrared and visible image fusion,” *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–15, 2021.

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

[27] Z. Wang, Y. Wu, J. Wang, J. Xu, and W. Shao, “Res2Fusion: Infrared and visible image fusion based on dense Res2Net and double non-local attention models,” *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[28] Z. Wang, J. Wang, Y. Wu, J. Xu, and X. Zhang, “UNFusion: A unified multi-scale densely connected network for infrared and visible image fusion,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3560–3574, Jun. 2022.

[29] Z. Wang, Y. Chen, W. Shao, H. Li, and L. Zhang, “SwinFuse: A residual swin transformer fusion network for infrared and visible images,” *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[30] H. Li, X. Wu, and J. Kittler, “RFN-Nest: An end-to-end residual fusion network for infrared and visible images,” *Inf. Fusion*, vol. 73, pp. 72–86, 2021.

[31] F. Zhao and W. Zhao, “Learning specific and general realm feature representations for image fusion,” *IEEE Trans. Multimedia*, vol. 23, pp. 2745–2756, 2021.

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

[33] J. Li, H. Huo, K. Liu, and C. Li, “Infrared and visible image fusion using dual discriminators generative adversarial networks with Wasserstein distance,” *Inform. Sci.*, vol. 529, pp. 28–41, 2020.

[34] H. Zhou, W. Wu, Y. Zhang, J. Ma, and L. Ling, “Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network,” *IEEE Trans. Multimedia*, early access, Nov. 22, 2021, doi: 10.1109/TMM.2021.3129609.

[35] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, “GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion,” *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.

[36] A. Song, J. Liu, H. Duan, H. Pei, and L. Ding, “Triple-discriminator generative adversarial network for infrared and visible image fusion,” *Neurocomputing*, vol. 483, pp. 183–194, 2022.

[37] Y. Yang et al., “Infrared and visible image fusion via texture conditional generative adversarial network,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 12, pp. 4771–4783, Dec. 2021.

[38] J. Li et al., “Multigrained attention network for infrared and visible image fusion,” *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[39] J. Li, H. Huo, C. Li, R. Wang, and Q. Feng, “AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks,” *IEEE Trans. Multimedia*, vol. 23, pp. 1383–1396, 2021.

[40] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 3–19.

[41] A. Toet, “TN0 image fusion dataset,” Figshare data, 2014. [Online]. Available: https://figshare.com/articles/TN_Image_Fusion_Dataset/1008029

[42] H. Xu, “Roadscene database,” 2020. [Online]. Available: https://github.com/hanana-xu/RoadScene

[43] S. Artifin, “OTCBVS database,” 2016. [Online]. Available: http://vcipl-okstate.org/pbvs/bench/

[44] V. Aslantas and E. Bendes, “Assessment of image fusion procedures using entropy, image quality, and multispectral classification,” *J. Appl. Remote Sens.*, vol. 2, no. 1, pp. 1–28, 2008.

[45] Y. Rao, “In-fibre Bragg grating sensors,” *Meas. Sci. Technol.*, vol. 8, no. 4, pp. 355–375, 1997.

[46] A. Eskicioglu and P. Fisher, “Image quality measures and their performance,” *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.

[47] Z. Liu et al., “Objective assessment of multisolution image fusion algorithms for context enhancement in night vision: A comparative study,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 94–109, Jan. 2012.

[48] G. Piella and H. Heijmans, “A new quality metric for image fusion,” in *Proc. IEEE Int. Conf. Image Process.*, 2003, pp. 173–176.

[49] Y. Han, Y. Cai, Y. Cao, and X. Xu, “A new image fusion performance metric based on visual information fidelity,” *Inf. Fusion*, vol. 14, no. 2, pp. 127–135, 2013.

Zhishe Wang (Membre, IEEE) received the B.S. degree in automation from the North China Institute of Technology, Taiyuan, China, in 2002, the M.S. and Ph.D. degrees in signal and information processing from the North University of China, Taiyuan, in 2007 and 2015, respectively. He is currently an Associate Professor with the Taiyuan University of Science and Technology, Taiyuan. His research interests include computer vision, pattern recognition, and machine learning.

Wenyu Shao received the B.S. degree in engineering mechanics in 2020 from the Taiyuan University of Science and Technology, Taiyuan, China, where he is currently working toward the M.S. degree in electronic information. His research interests include image fusion and deep learning.

Yanlin Chen received the B.S. degree in communication engineering from the Hunan University of Technology, Zhuzhou, China, in 2019. He is currently working toward the M.S. degree in optical engineering with the Taiyuan University of Science and Technology, Taiyuan, China. His research interests include image fusion and deep learning.
Jiawei Xu received the M. S. degree in image processing from Hallym University, Chuncheon, South Korea, in 2011, and the Ph.D. degree with the research topic of human factors in driving from the University of Lincoln, Lincoln, U.K., in 2015. He is currently with the Key Laboratory of Intelligent Informatics for Safety & Emergency of Zhejiang Province, Wenzhou University, Wenzhou, China. He is also an Adjunct Consultant with Huawei Technologies specialized in human factors in driving from 2019. He was with Newcastle University, Newcastle upon Tyne, U.K. from 2015 to 2019 with the research topic of human factors in aviation. His research interests include human factors in intelligent driving (L2, L3, L4), human factors in aviation (Cessna, A320).

Xiaoqin Zhang (Senior Member, IEEE) received the B.S. degree in electronic information science and technology from Central South University, Changsha, China, in 2005, and the Ph.D. degree in pattern recognition and intelligent system from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2010. He is currently a Professor with Wenzhou University, Wenzhou, China. His research interests include pattern recognition, computer vision, and machine learning.